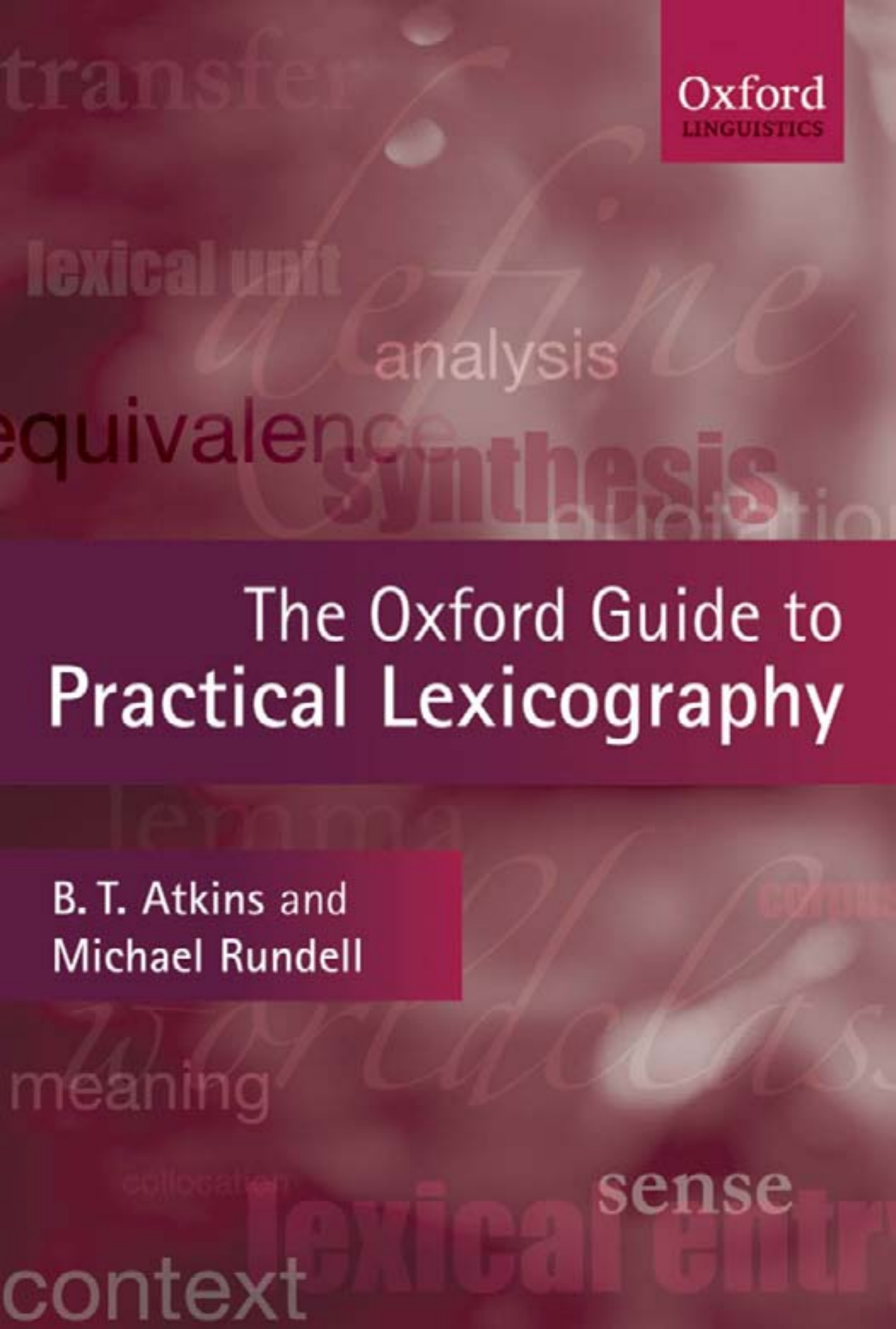


The Oxford Guide to Practical Lexicography

B. T. Atkins and
Michael Rundell

The Oxford Guide to Practical Lexicography

This page intentionally left blank

The Oxford Guide to Practical Lexicography

B. T. Sue Atkins
and Michael Rundell

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trademark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2008

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Data available
Typeset by SPI Publisher Services, Pondicherry, India
Printed in Italy by Legoprint S.p.A.
ISBN 978–0–19–927770–4 (Hbk.)
ISBN 978–0–19–927771–1 (Pbk.)
1 3 5 7 9 10 8 6 4 2

Contents

Acknowledgements

Our thanks go to the many friends and colleagues who have helped us in the making of this book: Valerie Grundy, for her painstaking reading of and gentle comments on some of the chapters in manuscript; Adam Kilgarriff, Penny Silva, and Tony Cowie, for encouraging us to start this project and see it through to the end; Thierry Fontenelle (editor of the companion volume to this, *Practical Lexicography: A Reader*) for his contagious enthusiasm which spurred us on when we were flagging; Charles Fillmore for his kindly and erudite explanations in response to showers of email queries; Philippe Climent for his generous help in producing screenshots good enough for print; Patrick Hanks, Rosamund Moon, Faye Carney, Ramesh Krishnamurthy, Krista Varantola, and Daan Prinsloo for advice and enlightenment at various points along the way; and for answering our many questions about their publications, Catherine Love at HarperCollins, and Vivian Marr and Nicholas Rollin at OUP.

We are grateful, too, to OUP's editorial team – John Davey, Chloe Plummer, Karen Morgan, and Malcolm Todd – who guided this book from manuscript to publication, spotting our mistakes and inconsistencies, and responding to our innumerable queries, with patience and good humour.

We are happy to thank the publishers whose dictionaries we have cited to illustrate points we are making: Houghton Mifflin Inc. for the *American Heritage Dictionary*; HarperCollins Publishers for *Collins English Dictionary*, the *Cobuild English Dictionary*, and the *Collins-Robert French Dictionary*; Macmillan Publishers Ltd for the *Macmillan English Dictionary for Advanced Learners*; Merriam Webster Inc. for the *Merriam-Webster Collegiate Dictionary* and *Webster's Third New International Dictionary*; Cambridge University Press for the *Cambridge Advanced Learner's Dictionary*; Pearson Education Ltd for the *Longman Dictionary of Contemporary English* and the *Longman Language Activator*; and Oxford University Press for the *Oxford Advanced Learner's Dictionary*, the *Oxford Dictionary of English*, the *Oxford English Dictionary*, and the *Oxford-Hachette French Dictionary*.

We remember with gratitude the students of our MSc course at the University of Brighton, and everyone who has attended our Lexicom Workshops over the years: we learned at least as much from them as they did from us, and the materials we developed in the process have now morphed into this book.

Finally, a special thank you to Peter and Maggy, who have helped us through the book's gestation with just the right mix of encouragement, support, and wry amusement that anyone should embark on such a project: this book is dedicated to them both.

Abbreviations and symbols

(a) Dictionary titles

We include extracts from many different dictionaries at various points in the book. For the ones we refer to most frequently, we use the following abbreviations:

AHD	*American Heritage Dictionary* Houghton Mifflin Company, Boston MA, USA
CALD	*Cambridge Advanced Learner's Dictionary* Cambridge University Press, Cambridge, UK
CED	*Collins English Dictionary* HarperCollins Publishers, Glasgow, UK
COBUILD	*Cobuild English Dictionary* HarperCollins Publishers, Glasgow, UK
CRFD	*Collins-Robert French Dictionary* HarperCollins Publishers, Glasgow, UK
LDOCE	*Longman Dictionary of Contemporary English* Pearson Education Ltd, Harlow, UK
MED	*Macmillan English Dictionary for Advanced Learners* Macmillan Publishers Ltd, Oxford, UK
MWC	*Merriam-Webster Collegiate Dictionary* Merriam Webster Inc, Springfield MA, USA
MW-3	*Merriam-Webster Third International Dictionary* Merriam Webster Inc, Springfield MA, USA
OALD	*Oxford Advanced Learner's Dictionary* Oxford University Press, Oxford, UK
ODE	*Oxford Dictionary of English* Oxford University Press, Oxford, UK
OED	*Oxford English Dictionary* Oxford University Press, Oxford, UK
OHFD	*Oxford-Hachette French Dictionary* Oxford University Press, Oxford, UK

References to these dictionaries indicate the edition referred to and its publication date. Thus *AHD-4* (2000) refers to the 4th Edition of the *American Heritage Dictionary*, published in 2000.

(b) Use of the ➔ symbol

In the *Oxford Guide to Practical Lexicography* we provide practical suggestions at many points. These are introduced by the ➔ symbol.

(c) Other Abbreviations

BNC	British National Corpus
CQL	corpus query language
CQS	corpus query software
DTD	document type definition
DV	defining vocabulary
DWS	dictionary writing system
ECD	Explanatory Combinatorial Dictionary
FE	frame element
FSD	full-sentence definition
HTML	Hypertext Markup Language
IPA	International Phonetic Alphabet
KWIC	keyword in context
LOB	Lancaster-Oslo-Bergen corpus
LU	lexical unit
MLD	monolingual learners' dictionary
MWE	multiword expression
NC	noun countable
NLP	natural language processing
NP	noun phrase
NU	noun uncountable
OEC	Oxford English Corpus
OED	Oxford English Dictionary
PDF	Portable Document Format
POS	part of speech
PP	prepositional phrase
RTF	Rich Text Format
SL	source language

TL	target language
V + O	verb + object
VP	verb phrase
WSD	word sense disambiguation
WYSIWYG	what you see is what you get
XCES	XML Corpus Encoding Standard

1

Introduction

1.1 What this book is about

The *Oxford Guide to Practical Lexicography* (*OGPL*) is a complete introduction to the job of creating a dictionary. It provides a step-by-step guide to all the tasks involved in the planning, resourcing, and compilation of reference materials for human users. The clue is in the title. It is a book about how to write dictionaries. Or, more accurately, about how we write dictionaries – something we have both been doing for the better part of our working lives.

For those who are interested, there are plenty of books 'about dictionaries' – about their macrostructure and microstructure, their strengths and their weaknesses. This is the province of the metalexicographers, for whom the dictionary itself is the object of study. There is a thriving metalexicographic community, represented by scholars such as H.-E. Wiegand, F.-J. Hausmann, Gabriele Stein, Reinhard Hartmann, and Henri Béjoint. Others – one thinks for example of Ladislav Zgusta, Bernard Quemada, Alain Rey, Josette Rey-Debove, Carla Marello, Dirk Geeraerts, and Laurence Urdang – have written eloquently about dictionaries while also being actively involved in the business of dictionary-making. Our focus in *OGPL* is on practical methodologies for transforming raw language data into dictionaries, though finding out about these will give you plenty of insights into the general nature of dictionaries.

All dictionaries are incomplete, and come under the heading 'work in progress'. And just as there is no such thing as a perfect dictionary, there is, equally, no 'right' way to produce a dictionary. So we make no special claims for the methodology we outline in this book, because there are many different ways of reaching the same goal. What we describe here is what has worked well for us over a number of years. And although the *OGPL* is written in English and most of the examples we give are from English dictionaries (or from English-French dictionaries when we exemplify bilingual issues), the lexicographic techniques we describe are for the most part language-independent.[1]

1.2 What lexicographers do

Dictionaries are often perceived as authoritative records of how people 'ought to' use language, and they are regularly invoked for guidance on 'correct' usage. They are seen, in other words, as *prescriptive* texts. Lexicographers have for long been uncomfortable with this idea – at least from the time of James Murray, the founding editor of the *Oxford English Dictionary* – and we see ourselves as working firmly within the tradition of *descriptive* lexicography. For us, a dictionary is a description of the vocabulary used by members of a speech community (for example, by 'speakers of English'). And the starting point for this description is evidence of what members of the speech community do when they communicate with one another. But between the raw linguistic data and the finished dictionary, a number of other factors come into play, as Figure 1.1 shows. Each box in the diagram represents an 'input' to the lexicographic process, and we deal with all these issues later in the book. Lexicographers need language technology to gain access to linguistic data; we need linguistic theory to help us analyse the data effectively and draw useful conclusions from it; and we have to understand the needs of our target audience if we are going to produce a language description that is accessible and relevant to the people who will use it.

[1] Inevitably there are exceptions: for example, deciding on what should be a headword – not an especially big problem for those of us working in European languages – is fraught with difficulty for lexicographers describing the languages of southern Africa.

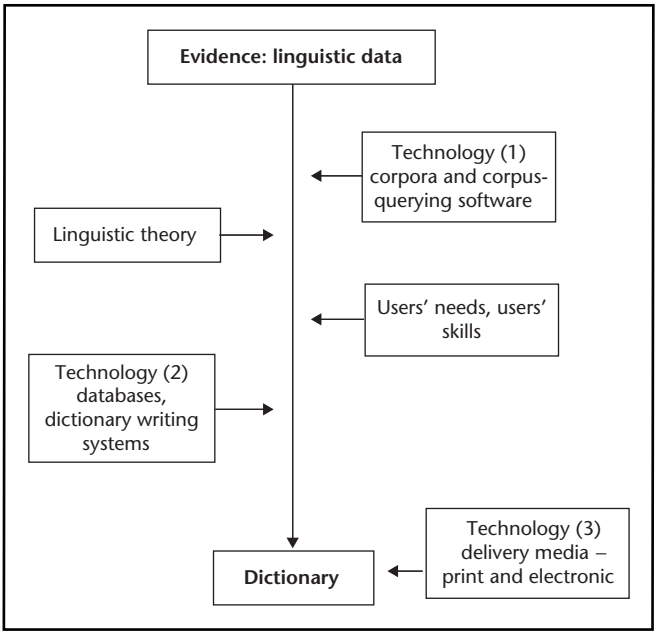


Fig 1.1 From data to dictionary

1.2.1 Lexicography and technology

Computers were first employed in the dictionary-making process in the 1960s, and in the intervening half-century the role of technology has become ever more central. In the twenty-first century, all good dictionaries take corpus data as their starting point, and the contemporary lexicographer (typically querying the corpus online and recording dictionary data in a structured database) depends on a number of technologies – most of them of recent origin. These include:

- personal computers with vast storage capacity, powerful processors, and fast internet links
- corpus data, processed using software tools developed in the Natural Language Processing community and accessed through dedicated querying programs
- software for inputting dictionary text, and databases that store and manage the text as it develops.

And once the dictionary has been compiled, technology offers a number of ways, and a number of media, for making it available to the end-user.

Improvements in hardware and infrastructure have been critical here, but the lexicographic community also owes a big debt to those computational linguists who have made our lives easier (and made the dictionaries we produce better) by applying their expertise to lexicographic tasks. In this context, Adam Kilgarriff, Pavel Rychlý, Antonio Zampolli, Roy Byrd, Ken Church, Ulrich Heid, Greg Grefenstette, and Thierry Fontenelle deserve special thanks.

1.2.2 Lexicography and theory

This is not a book about 'theoretical lexicography' – for the very good reason that we do not believe that such a thing exists. But that is not to say that we pay no attention to theoretical issues. Far from it. There is an enormous body of linguistic theory which has the potential to help lexicographers to do their jobs more effectively and with greater confidence. In the *OGPL* we refer to theoretical discussions whenever they illuminate the task in hand and help us to inject more 'system' into our work. People whose day job is writing dictionaries can't hope to remain fully abreast in every area, but fields of particular relevance to our work include lexical semantics, cognitive theory, pragmatics, and corpus linguistics. There is no question that lexicography has benefited hugely from the insights of scholars such as Charles Fillmore, Igor Mel'čuk, John Sinclair, Juri Apresjan, Alan Cruse, Eleanor Rosch, Beth Levin, Annie Zaenen, George Lakoff, and Douglas Biber (to name just a few). It's important to stress that these linguists don't (in general) address lexicographic issues directly. Their focus is language, not dictionaries, and they don't 'tell lexicographers what to do, or how to solve problems'. Rather, 'they show us different ways of looking at language, which we can take and adapt to our needs' (Atkins 1993: 29). Lexicographers have a great deal to learn from linguistic theory, and many of the recent improvements in dictionaries can be attributed to the intelligent application of theoretical ideas.

1.2.3 Lexicography and dictionary users

But making dictionaries 'is not a theoretical exercise to increase the sum of human knowledge but practical work to put together text that people can understand'. So says Sidney Landau (2001: 153), himself a distinguished lexicographer, whose classic volume, *Dictionaries: the Art and*

Craft of Lexicography, is warmly recommended for anyone who wants to know what goes on in the production of a published dictionary. 'The value of a work', as Johnson says, 'must be estimated by its use', and the most important single piece of advice we can give to anyone embarking on a dictionary project is: know your user. The *OGPL* invokes this mantra in every chapter, and we make no apology for this. This doesn't imply a superficial concern with 'user-friendliness', but arises from our conviction that the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for. Samuel Johnson (as is increasingly recognized) identified and grappled with almost all the problems that preoccupy lexicographers today.[2] But what is most impressive of all is his insistence that users' needs are paramount, and users' skills (or lack of them) must be taken into account. In a famous reflection on this theme, he says:

> It is not enough that a dictionary delights the critick, unless, at the same time, it instructs the learner; as it is to little purpose that an engine amuses the philosopher by the subtilty of its mechanism, if it requires so much knowledge in its application as to be of no advantage to the common workman.
>
> (*The Plan of an English Dictionary*, 1747)

Crudely paraphrased, this tells us that no amount of theoretical rigour is worth a hill of beans if the average user of your dictionary can't understand the message you are trying to convey.

1.3 How this book works

The *OGPL* is in three parts:

- Part 1 (Chapters 2–7): 'Pre-lexicography'
- Part 2 (Chapters 8–9): 'Analysing the data'
- Part 3 (Chapters 10–12): 'Compiling the entry'.

Part 1 deals with the things you need to know, the tasks you have to perform, and the resources you need to assemble before you can embark on writing your dictionary. In Part 2, we take you through the two principal stages of the analysis process: discovering the senses of the headword (the 'lexical units'), and recording the lexicographically relevant facts about each of these units. In Part 3 we demonstrate in detail how we compile entries for

[2] See for example Hanks (2005): 243–244.

monolingual and for bilingual dictionaries, including a discussion of the translation process that is a necessary part of bilingual entry writing. If you are using *OGPL* as a textbook, Part 1 can be seen as a reference section providing background information, while Parts 2 and 3 form a complete set of teaching modules.

At the end of each chapter we provide a reading list in two parts:

- recommended reading on the topics covered in the chapter
- further reading on these and related topics.

All the books and articles we refer to are listed in a full bibliography at the end of the book, and many of the most relevant papers also appear in a companion volume to this one: *Practical Lexicography: A Reader* edited by Thierry Fontenelle (2008). Finally, all the chapters that deal directly with the creation of dictionary text are accompanied by practical exercises. Figure 1.2 gives an outline of the contents of *OGPL*.

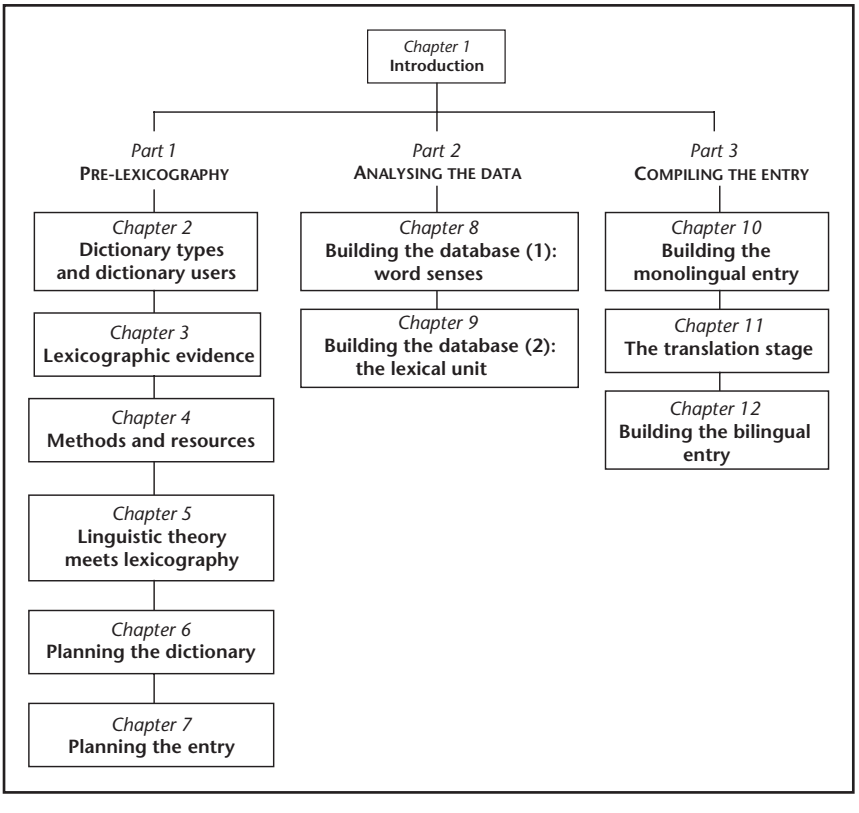


Fig 1.2 Contents of the book

In the *Pre-lexicography* section, **Chapter 2** looks at the earliest stages in the planning of a dictionary project: first, at the decisions that have to be made about the type of dictionary you are writing (monolingual or bilingual, for native speakers or learners of the language, for adults or children, and so on); and second, at the creation of a 'user profile', a description of the typical user of the dictionary with an assessment of their needs and linguistic skills. In **Chapter 3** we discuss sources of lexicographic evidence, in particular the design, collection, and processing of a text corpus for dictionary-building. **Chapter 4** starts by outlining the main stages in the editing process, from corpus to finished entries; describes software for corpus-querying and entry-writing; and introduces the Style Guide, the document that sets out, in fine detail, the way in which dictionary entries should be written. Some of the most useful ideas drawn from theoretical work in linguistics are introduced in **Chapter 5**, including sense relationships (hyponymy, synonymy, etc.), Fillmore's frame semantics, the concept of lexicographic relevance, and Mel'čuk's lexical functions. (The relevance of other theoretical areas, such as prototype theory and pragmatics, is explained in other chapters.) Dictionary 'macrostructure' is discussed in **Chapter 6**, where we look at the principal types of entry found in most dictionaries, and at the various kinds of lexical item about which inclusion decisions have to be made. But the greater part of this chapter focuses on the issues involved in building the dictionary's headword list, from simple words through proper names to multiword expressions. Part 1 of the book concludes with **Chapter 7**, where we deal with the dictionary's 'microstructure', and look at each of numerous possible entry components. We describe their function and illustrate their use in real dictionaries. The chapter ends with a brief look at the electronic dictionary, and at the microstructure decisions to be made over the internal organization of its entries.

The two chapters in the section entitled *Analysing the data* give an account of the work involved in extracting from the corpus all the information that is relevant for the dictionary. The first step in the process is to identify and record the senses of a polysemous word, and this is the central theme of **Chapter 8**. Here we describe a methodology for dividing words (or 'lemmas') into senses (or 'lexical units') and show how linguistic theory can contribute to successful word sense disambiguation. The next step in the process is the discovery and recording of facts about each lexical unit, and **Chapter 9** describes the kinds of material to be recorded, and how it

may be entered in a database, illustrating this with corpus data and extracts from sample entries. All the relevant properties of the lexical unit are covered here: its meaning, grammar, significant contexts, and combinatorial features (including the various types of multiword expression that have to be specified and recorded).

The section entitled *Compiling the entry*, as its name implies, gives a complete account of the way in which monolingual and bilingual entries are built on the basis of the facts systematically recorded during the analysis process. In **Chapter 10** we move on from preliminary database to finished entries in a monolingual dictionary. Here we discuss the options for presenting and ordering the various categories of information that make up an entry. This chapter deals with topics such as grammar, labelling, and illustrative examples, but its primary focus is on issues relating to the key function of writing definitions. Bilingual dictionaries are considered separately, and **Chapter 11** goes through the process of finding equivalences in a target language for source language items of every type. The discussion covers all the factors involved in inserting useful translations into the database, for later use by the editors writing bilingual entries; using source language and target language corpus data to find and check translations; and lastly, how these may be recorded in the database entry. Finally **Chapter 12** provides an account of how a bilingual entry is assembled from the materials created in the previous stages. We look at the ways information can be distributed in the entry and at the tasks involved in putting the entry together – starting with decisions on the presentation of senses; working through the various options for showing translations, and selecting examples; and finally proposing strategies for helping users to choose the most appropriate target language expression for their purpose.

1.4 And finally . . .

Anna Wierzbicka, who has written prolifically and insightfully about semantics and cognition (while taking the occasional sideswipe at the hapless lexicographer) famously observed that 'lexicography has no theoretical foundations, and even the best lexicographers, when pressed, can never explain what they are doing and why' (1985: 5). Her observation has a good deal of truth in it (though perhaps a little less than when she made it). It is

framed as a sort of exasperated reproof – but is this absence of theory such a bad thing? It may make more sense to think in terms of the *principles* that guide lexicographers in their work. We have already hinted at what these are in our case, but let's now attempt a summary.

Our objective in producing dictionaries is to create a description of language which is faithful to the available linguistic evidence, and optimized to take account of the specific needs and skills of those who will use the dictionary. To a significant degree, this process entails the exercise of subjective judgment – consider, for example, the way that we all (as lexicographers or ordinary language-users) go about the task of finding meaning in texts. But we recognize (and welcome) the fact that this subjective element can at many points be made more objective, either through the contribution of intelligent software or through the application of linguistic theory. This interaction between lexicography, linguistics, and language engineering has helped to make dictionaries more systematic, more internally consistent, more complete, and simply better as representations of how people use language in real communicative situations. And we have no doubt that these collaborations have more to offer as we go forward. In the end, though, we share Johnson's view that 'in lexicography, as in other arts, naked science is too delicate for the purposes of life'. Natural languages are dynamic systems, which tolerate a good deal of inventiveness, idiosyncrasy, and deviation from 'normal' behaviour. Consequently, efforts to make them conform to one particular way of looking at language, efforts – in short – to describe language 'scientifically', have usually foundered when they come up against what Landau (1993: 113) refers to as 'the stubborn diversity of actual usage'. If we have a theoretical position at all, it is a belief that most (if not all) of the things that people do with language are *motivated*. So, for example, if a phrasal verb depends on one particle rather than another, or an originally monosemous word acquires new meanings and uses, these things tend to happen in ways that are systematic rather than arbitrary. The underlying systems aren't always easy to retrieve and describe, but they are, ultimately, accessible to anyone with enough data, enough perseverance, and enough analytical nous. This is one of the challenges that make lexicography so exciting.

You learn about lexicography by doing it, by training other people to do it (which we have been doing for over two decades), and by talking about it with colleagues. We have learned a lot from the dictionary-lovers and dictionary-practitioners who belong to the major lexicographic

associations: EURALEX, the Dictionary Society of North America, Afrilex, Australex, and Asialex. But it is no accident that the most relevant things written about lexicography have been written by lexicographers – starting with Samuel Johnson, and continuing with our friends and colleagues in the profession, especially Patrick Hanks, Rosamund Moon, Tony Cowie, and the late Penny Stock. And as someone who has expanded our horizons and improved the quality of our lexicographic life, Adam Kilgarriff deserves a special mention. Though not a lexicographer himself, he is the only world-class computational linguist who genuinely understands what lexicographers do (and what they need in order to do it better). It has been our good fortune to work with some of the best people in the lexicographic world, and we have learned an enormous amount from all of them.

And finally, we leave the last word to the Great Cham of Literature (and Lexicography) . . .

When I survey the Plan which I have laid before you, I cannot, my Lord, but confess, that I am frighted at its extent, and, like the soldiers of Cæsar, look on Britain as a new world, which it is almost madness to invade. But I hope, that though I should not complete the conquest, I shall, at least, discover the coast, civilize part of the inhabitants, and make it easy for some other adventurer to proceed further, to reduce them wholly to subjection, and settle them under laws.

Samuel Johnson, *Plan of a Dictionary* (1747)

~~~
~~~

Reading

Recommended reading

Landau 2001: 6–42; Aitchison 2003; Murray 1979; Bolinger 1980; Johnson 1747, 1755.

Further reading on related topics

Zgusta 1971; McArthur 1986; Littré 1880; Hanks 2005.

Websites

Lexicography associations: http://www.euralex.org; http://polyglot.lss.wisc.edu/dsna/index.html; http://afrilex.africanlanguages.com/; http://www.asialex.org/; http://www.australex.org/; http://crcl.th.net/index.html?main=http%3A//crcl.th.net/sealex/

International Journal of Lexicography papers online: http://www3.oup.co.uk/lexico/

This page intentionally left blank

PART I

Pre-lexicography

This page intentionally left blank

Introduction to Part I

Part 1 of the book covers what we call 'Pre-lexicography' – the planning stages of a dictionary project. It explains the things you need to know, the tasks you have to perform, and the resources you need to assemble before your project can get properly under way. We deal first (in Chapter 2) with the business of specifying a dictionary – making decisions about the kind of reference book it will be, the type of information it will contain, and the kinds of people who will use it. Next, Chapter 3 takes you through the process of acquiring a corpus, a body of evidence providing the raw language data on which your dictionary will be based. In Chapter 4 we look at the other resources you will need, including software (for querying the corpus and building the dictionary database), a Style Guide, and a set of template entries. Chapter 5 provides an introduction to a number of concepts from theoretical linguistics which have particular relevance to the work we do as lexicographers. Chapters 6 and 7 deal, respectively, with the dictionary's macrostructure and microstructure, describing first the process of building a headword list and selecting the main types of entry the dictionary will include, and then the structure and components of individual dictionary entries. By the end of Part 1 you will have everything in place to begin the next stage – the lexicography itself.

This page intentionally left blank

2

Dictionary types and dictionary users

This chapter sets dictionary writing in its context. It looks at how the dictionary comes about in the first place and how dictionaries may be classified. The dictionary user is shown to play a central role in the planning process, and we illustrate the ways in which editorial decisions are influenced by our understanding of the needs and skills of our dictionary's typical user. Figure 2.1 sets out the plan of the chapter.

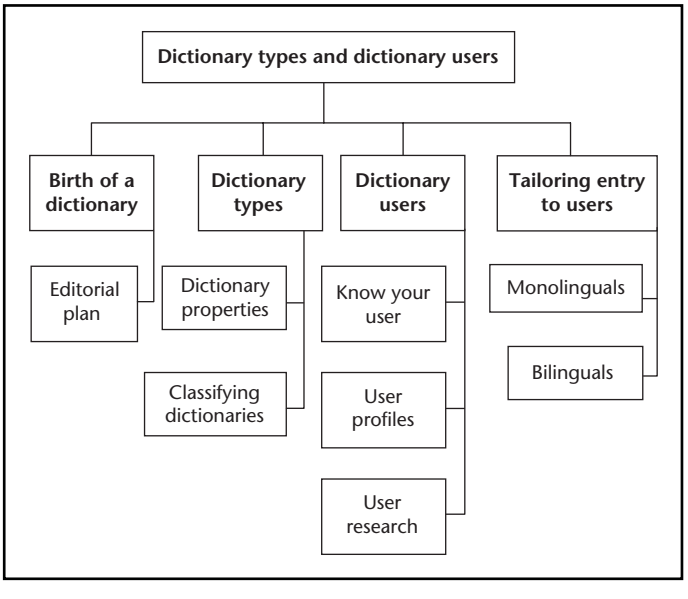


Fig 2.1 Contents of this chapter

2.1 The birth of a dictionary

Dictionaries are not born every day. They are hugely expensive to produce from scratch, and most 'new' dictionaries still owe much to some earlier incarnation. Sometimes however you get the chance to do it all from the first twinkle in the eye of the publisher. In such a case there is a 'pre-lexicography' stage, when the most fundamental decisions are taken, affecting every aspect of the lexicography. Figure 2.2 outlines this process through to dictionary publication. During this pre-lexicography stage, the decision-making process typically involves a dialogue between publisher and senior editor. The editor may have plenty of ideas about what s/he would like to do, but the final say rests with the publisher, who holds the purse strings.

The sequence of events typically goes like this:

- The marketing department spots a 'gap' on the booksellers' shelves, and commissions from the editorial department a dictionary to fill that gap. (For all but scholarly or historical dictionaries, market forces come into play here: the new work will have to sell against existing dictionaries produced by competitor publishers.)
- The marketing department specifies the type of dictionary needed, describes the market it will sell to and thus the type of user it is destined for, and paints a broad-brush picture of what its contents should be.
- The eventual selling price of the proposed dictionary is to a large extent dictated by the price of competing dictionaries, and this in turn constrains the overall budget of the project.
- The budget dictates the schedule (timeline, personnel, resources, etc.).
- The budget and schedule are passed to the editorial department where the dictionary is designed and developed.
- For the dictionary planners who will work within this budget to create a dictionary for a specific market, the needs of the end-user determine the extent of the book and its content (the number of headwords, the depth of their treatment, the type of material to be included in the front and back matter, etc.).[1] The styling of entries is specified, and sample entries are produced and circulated for comment. The Style Guide is drafted. The dictionary planners work with

[1] The editors' detailed planning of the dictionary is described in Chapters 6 and 7.

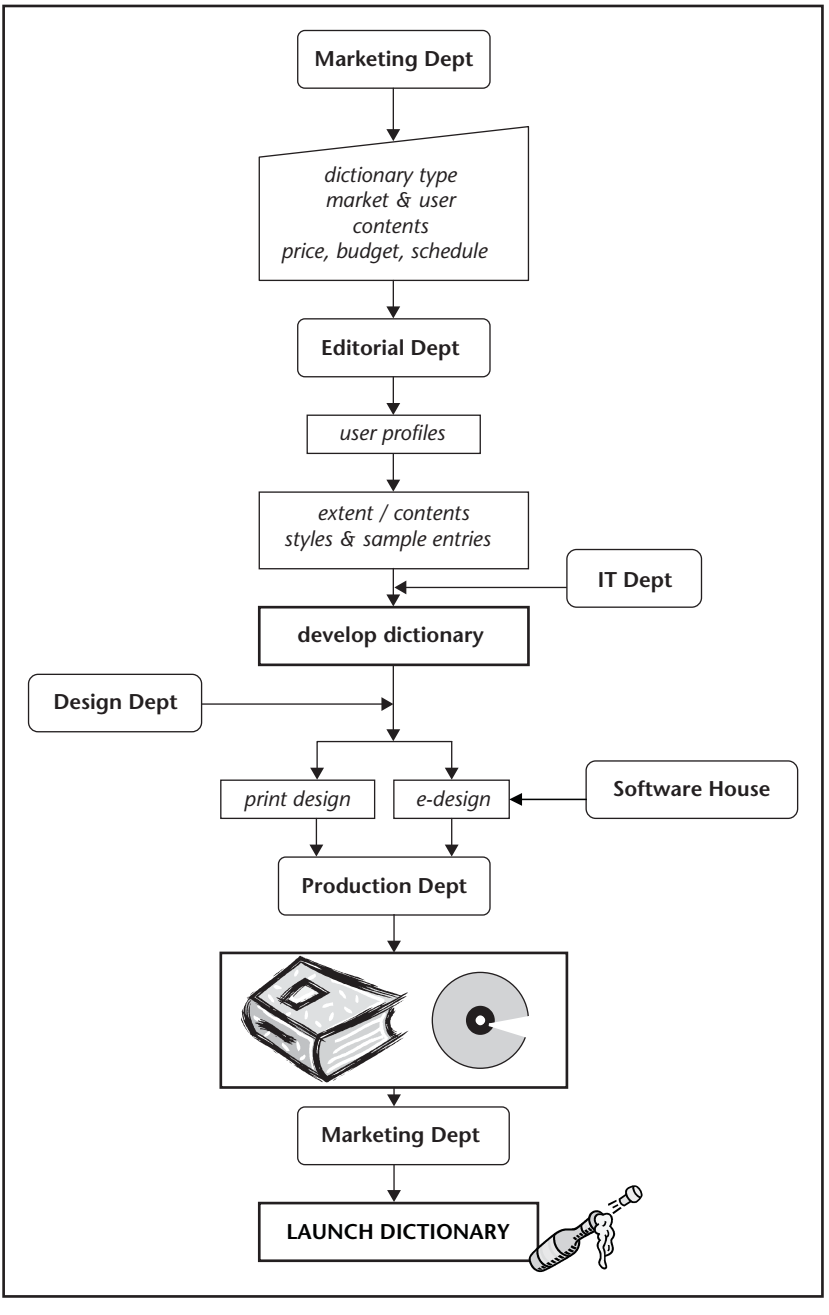


Fig 2.2 The birth of a dictionary

the IT department to customize the dictionary writing software (DWS), and – if the publishers have no corpus – to design and build a lexicographic corpus together with its corpus query software (CQS); also to provide the hardware for the project.
- The editorial planners set up a system of text flow, text back-up etc., often one which allows the dictionary editors to work online from home.
- The editors also have in mind the type of presentation needed for the dictionary to be effective and attractive, and usually there are early discussions with the design department (see e.g. Luna 2004).
- The e-dictionary (the electronic version of the dictionary), if there is to be one, is usually commissioned from an outside software firm, who develop the user interface in collaboration with the dictionary planners, allowing both print and electronic versions of the dictionary to be compiled simultaneously.
- When the dictionary text is ready, it is passed to the production department, who take it through to book and electronic form.
- The marketing department, in consultation with the editorial department, handles the launch of the new dictionary.

2.1.1 Developing the editorial plan

For lexicographers the birth of a new dictionary offers exciting opportunities. The potential for improvement and innovation is almost infinite, but two general principles have to be kept in mind:

- Space is finite and has to be used intelligently.
- A dictionary is like an eco-system: decisions about content, presentation, and design can't be made in isolation, because a change to one part of the system impacts on all the other parts of it.

2.1.1.1 *The intelligent use of space* This is a zero-sum game. Space is finite, so if you use a certain amount of it for one purpose, that amount is not available for any other purpose. Even the 20-volume *OED* makes no claim to include all the vocabulary of English.[2] Inevitably, then, the average

[2] 'There are a number of myths about the *Oxford English Dictionary*, one of the most prevalent of which is that it includes every word, and every meaning of every word, which

one-volume dictionary can cover only a small proportion of the vocabulary of a language.

It may seem obvious that the more information a dictionary contains, the more helpful it is likely to be. In fact, though, there is always a trade-off between *coverage* (how much information a dictionary includes) and *accessibility* (how easy it is for users to find the information they need and successfully process it). Over the centuries, dictionaries have evolved strategies to maximize the use of limited space, for example by the use of codes, abbreviations, and a 'telegraphic' defining style. But all this comes at a cost. Until its ninth edition (1995), the *Concise Oxford Dictionary* was a miracle of compression, packing an astonishing amount of information into a small-format one-volume dictionary. But as the extract in Figure 2.3 shows, not all of this information is readily retrievable by an unskilled user.

băg[1] *n.* **1.** receptacle of flexible material with closable opening at top (esp. w. prefixed word showing contents or purpose; DIPLOMATIC *bag*, GAME[1] *bag*, HAND[1] *bag*, KIT[1] *bag*, *mailbag, travelling-bag*, VANITY *bag*); (w. such prefix understood) particular kind of this; hence ~FUL. 2 *n.* **2.** contents of bag; MIXED *bag*; amount of game a sportsman has shot or caught (also fig.) **3. ~and baggage**, with all belongings; **~of bones** lean creature**; (whole) ~of tricks** every… [etc]

Fig 2.3 Extract from *Concise Oxford Dictionary* (1982)

Since the 1970s, a countervailing tendency has stressed user-friendliness, and this has led to a re-evaluation of the value of packing large amounts of information into a small space. But almost anything you do to make dictionary text easier to process will take up more space. Writing out 'noun', instead of the abbreviated '*n*', may seem a trivial change, but if your dictionary has 25,000 noun headwords, its effects are multiplied by 25,000. Another option is to begin the description of each new sense on a new line. This is appealing: the page looks less cluttered and users find it easier to locate the meaning they want. But it all takes up space, and that means a reduction in the amount of information that can be included.

has ever formed part of the English language' (John Simpson, Editor of *OED*: quoted on *OED* website).

2.1.2 The dictionary as eco-system

It's sometimes useful to think of the dictionary in 'database' terms, as a set of components (such as definitions, etymologies, and pronunciations) that can be dealt with discretely. But when planning a dictionary, we have to think of it as a complete system, in which all these components are inextricably related to one another. To give a concrete example: one of the decisions the editor of an English dictionary has to make at the planning stage is how to handle inflections. The options include:

- showing full inflections for every word (thus *sail, sails, sailing, sailed*);
- showing inflections only when they are irregular (and you then have to define what is meant by 'irregular');
- not showing them at all.

Each choice has its consequences – especially the first, which is very space-intensive. The benefits of any approach have to be weighed against its impact on available space: if we include more information about this particular feature, what others will have to be sacrificed? The *COBUILD* dictionaries, for example, generally avoid abbreviations, provide full inflections, and use a 'full-sentence' defining style in favour of conventional definitions. All of this, it may be argued, contributes to making the dictionaries more user-friendly. But there is an inescapable downside: these policies use up a lot of space, and consequently *COBUILD*'s dictionaries always have significantly fewer headwords than other books in the same category.[3]

Two questions arise:

- If we want our dictionary to include more information, why not just make it bigger?
- Doesn't the arrival of electronic media make this discussion irrelevant?

Dictionaries have a tendency to get bigger with each new edition. For example, the third edition of the *Oxford Advanced Learner's Dictionary* published in 1976 had 1,002 pages of A-Z text. Subsequent editions got steadily larger, and when the seventh edition came out in 2005 it weighed in at 1,780 pages – an increase of almost 80 per cent. Some of the extra space is used to include more information (more headwords, for example) and some to present existing information in more accessible forms (as by replacing coded

[3] Analysis of a fairly large sample suggests *COBUILD*'s headword count is around 23 per cent lower than that of comparable dictionaries (cf. Rundell 2006: 327).

verb patterns – like 'VP19C' – with more explicit grammatical guidance). But there are limits to how far this process can continue: more pages means higher costs to the publisher (and possibly also to the user), while the larger the book, the less portable it becomes and the more likely it is to fall apart. For all these reasons, it is likely that the dictionaries in this category are getting close to their maximum size.

In electronic media of all types (from PCs to iPods to mobile phones) data-storage capacity has become so cheap that it has ceased to be an issue. In their non-print versions, therefore, dictionaries no longer need to grapple with space constraints, and publishers are beginning to take advantage of this novel situation (see §7.2.11 for a brief discussion of electronic dictionaries). But if the careful rationing of space has ceased to be a concern for electronic dictionary planners, the opportunities offered by 'infinite capacity' bring their own challenges. The idea that we can simply include all of the lexical data available to us is fanciful; at the very least, the process calls for smart information management and sensitive design, if users are not going to suffer from a debilitating case of information overload. We need to be clear about the difference between doing things just because we *can*, and doing them because they will be of real value to the user.

Developing an editorial plan involves juggling a large number of interrelated variables. Every linguistic feature of your target language presents you with a range of options: should it be covered in the dictionary, and if so, in how much depth, and what is the most effective way to convey this information and display it on the page?

For each policy decision of this type, it is essential to be clear about:

- how much space it requires
- how this impacts the system as a whole
- whether it is in the best interest of users to devote so much space to it
- what has to be jettisoned to make that possible.

The best way of tackling these complex and challenging issues is to think first and always of the dictionary user. If you have a clear idea of who your user is and what they want from their dictionary, you stand a good chance of achieving the right fit between dictionary type and user need. The next two sections address these two aspects of dictionary planning. We look first at types of dictionary, then at types of user and ways of identifying their needs.

➔ Think first about the user when you're deciding what is to go in your dictionary, and how much prominence to give the various facts.

2.2 Types of dictionary

2.2.1 Properties of dictionaries

There are many different aspects of a dictionary to be taken into account when you are looking to classify dictionaries. If you are writing, or planning, a trade dictionary (not a scholarly and/or historical work but one that has to make its way in the hard commercial world) you need to be able to think clearly about the following:

1. **the dictionary's language(s):** is it . . .
   - **a.** monolingual
   - **b.** bilingual: if so, is it . . .
     (1) unidirectional[4] or
     (2) bidirectional[5]
   - **c.** multilingual (but we don't want to go there in this book)
2. **the dictionary's coverage:** is it . . .
   - **a.** general language
   - **b.** encyclopedic and cultural material
   - **c.** terminology or sublanguages (e.g. a dictionary of legal terms, cricket, nursing)
   - **d.** specific area of language (e.g. a dictionary of collocations, phrasal verbs, or idioms)
3. **the dictionary's size:** is it a . . .
   - **a.** standard (or 'collegiate') edition
   - **b.** concise edition
   - **c.** pocket edition
4. **the dictionary's medium:** is it . . .
   - **a.** print
   - **b.** electronic (e.g. DVD or handheld)
   - **c.** web-based

[4] A *unidirectional* bilingual dictionary, as the name implies, goes 'one way': a bilingual English-French dictionary contains a single text in which the source language (SL) is English and the target language (TL) is French, cf. §2.4.2.

[5] A *bidirectional* bilingual dictionary contains two texts and works 'both ways': in a bilingual English-French dictionary there is one text in which the SL is English and the TL is French, and a second text where the SL is French and the TL is English, cf. §2.4.2.

5. **the dictionary's organization:** is it . . .
   - **a.** word to meaning (the most common)
   - **b.** word to meaning to word (where looking up one word leads to other semantically related words)
6. **the users' language(s):** is the dictionary meant for . . .
   - **a.** a group of users who all speak the same language
   - **b.** two specific groups of language-speakers
   - **c.** learners worldwide of the dictionary's language
7. **the users' skills:** are they . . .
   - **a.** linguists and other language professionals
   - **b.** literate adults
   - **c.** school students
   - **d.** young children
   - **e.** language learners
8. **what they use the dictionary for**: is it for one or both of the following . . .
   - **a.** decoding, which is . . .
     - – understanding the meaning of a word
     - – translating from a foreign language text into their own language
   - **b.** encoding, which is . . .
     - – using a word correctly
     - – translating a text in their own language into a foreign language
     - – language teaching

2.2.2 Classifying dictionaries

You can use these properties to categorize most kinds of dictionary fairly exactly. Take three of the main types with which we are concerned in this book.

(1) You could describe a big one-volume collegiate dictionary for home, study, and office use such as the *AHD-4* (2000), the *ODE-2* (2003), or the *CED-8* (2006) as:

1a	(monolingual)
2ab	(general language, with some encyclopedic and cultural material)
3a	(standard edition)
4a	(print)
5a	(word-to-meaning)

6a (native English speakers)
7b (literate adults)
8a(b) (decoding with some encoding).

(2) A pocket-sized dictionary for school students, such as the *Collins School Dictionary* (Collins 1990) could be described as:

1a (monolingual)
2a (general language)
3c (pocket edition)
4a (print)
5a (word-to-meaning)
6a (native English speakers)
7c (school students)
8a(b) (decoding with some encoding).

(3) A collegiate one-volume English-French and French-English dictionary such as the *CRFD-2006* or the *OHFD-2001* would be categorized as:

1b(2) (bilingual: bidirectional)
2a (general language)
3a (standard edition)
4a (print)
5a (word-to-meaning)
6b (English speakers and French speakers)
7abce (linguists, adults, school students, language learners)
8ab (decoding and encoding).

(4) A dictionary such as the *Longman Language Activator* or the *Oxford Wordfinder* could be described as:

1a (monolingual)
2a (general language)
3a (standard edition)
4a (print)
5b (meaning-to-word)
6c (non-native English speakers)
7e (language learners)
8b (encoding).

As the examples above show, you can't use these categories to sort dictionaries into distinct classes, simply to describe them. The categories should be thought of as sets of *properties*. Every dictionary must have

at least one property from each category, but they can have more than one.

➔ When you're planning a new dictionary, consider the implications of each category carefully in the light of what you know about your market and typical users. This will help you to make your dictionary maximally useful.

2.3 Types of dictionary user

Creating a dictionary involves making decisions: big decisions at the planning stage and – as the project goes forward – smaller ones on a day-to-day basis. Many of these decisions entail some form of *selection*, because every dictionary contains a subset of all the available information about the target language and its vocabulary. For example, at any given point in the editorial process, you may have to decide whether to include a particular headword and, if so, how much information to give about it.

To some extent, the commercial factors outlined in the previous two sections will limit your room for manoeuvre. To give an obvious example, the length of the dictionary (usually agreed at the outset) restricts the number of headwords it can include. But within the parameters imposed by the publishing plan, there is still plenty of scope for variation, as Figure 2.4 illustrates.

clam•ber / …/ v [I always + adv/prep] to climb or move slowly somewhere, using your hands and feet because it is difficult or steep: [**+over/across etc**] *They clambered over the slippery rocks.* | *We all clambered aboard and the boat pulled out.*

LDOCE-4 (2003)

clamber / …/ **(clambers, clambering, clambered)** If you **clamber** somewhere, you climb there with difficulty, usually using your hands as well as your feet. ❑ *They clambered up the stone walls of a steeply terraced olive grove …*

VERB
= scramble

V prep/adv

COBUILD-5 (2006)

Fig 2.4 Entries for *clamber* in two dictionaries of the same type

These two dictionaries are designed for the same user-group: advanced learners of English. Both use a simple defining language, both provide illustrative examples, and both indicate (using codes) that *clamber* is typically followed by an adverbial or prepositional complement. But there remain significant differences. To give a few examples:

- *LDOCE* takes the view that its users don't need to be told about this verb's (regular) morphology, but *COBUILD* provides a full set of inflections.
- *LDOCE* lists two particles that typically follow *clamber* ('over' and 'across'). *COBUILD* doesn't, but on the other hand it gives a near-synonym ('scramble').
- *LDOCE* uses a conventional defining style, while COBUILD opts for a full-sentence definition.

This gives some idea of significant differences that can be found even among dictionaries occupying the same well-defined market slot. All these variations reflect editorial decisions made during the planning stage. But on what basis are such decisions made? And what can we do to ensure that we reliably make the 'right' decisions? There are two ways of finding out about the user: user profiling and user research. The process is never scientific, but the only possible starting point is the targeted user group. You need a clear understanding of who will use the dictionary, what they will use it for, and what kinds of skill they will bring to the task. If you have answers to all these questions, you have a firm basis for making well-informed decisions about both content and presentation.

➔ Know your users: that way, the dictionary will give them what they need.

2.3.1 User profiles and how to create them

A user profile seeks to characterize the typical user of the dictionary, and the uses to which the dictionary is likely to be put. It's true that some dictionaries have such a wide range of potential users and uses that it may be difficult to identify information specific enough to be useful. But even in such cases, the exercise is still worthwhile. To build a user profile, you need to think carefully about who your typical users will be, and what they will be using the dictionary for. The principal questions to ask yourself are given below.

2.3.1.1 *Types of user* Which of these groups do you expect them to belong to?

- adults, young children, or older children
- native speakers (of the language of the dictionary) or language-learners
  - if learners, are they beginners, intermediate, or advanced?

- general users or specialists
  - if specialists, what field are they working in?
- using the dictionary in an educational, domestic, or professional setting

2.3.1.2 *Types of use* Which of these tasks do you expect them to use the dictionary for?

- general reference purposes, such as
  - understanding unfamiliar words
  - checking spellings or pronunciations
  - doing crosswords
- studying a particular subject
- learning a language
- translating text from one language to another
- writing essays or reports
  - in their first language
  - in a language they are learning
- preparing for a written or oral exam

2.3.1.3 *Users' pre-existing skills* What skills and knowledge will they have? In particular, can you rely on . . .

- their linguistic knowledge:
  - How proficient are they in the language(s) used in the dictionary?
  - Do they know (or *need* to know) what is meant by terms like 'noun', 'present participle', and 'transitive'?
  - Can you assume they know regular morphology, or should you give information on all inflections?
- their familiarity with 'standard' dictionary conventions:
  - Do they understand abbreviations like *adj*?
  - Do they understand linguistic labels such as *informal* or *derog.*?
  - Do they understand grammatical codes, or cross-references to other entries?
  - Do they know how words are pronounced, or will you need to provide pronunciations? If so, will they know the International Phonetic Alphabet (IPA), or will you need to show pronunciation in some other way?

The more information of this type you can gather, the better-placed you will be to make informed decisions on a range of editorial and design issues.

2.3.2 User research and its relevance

'User research' refers to any method used for finding out what people do when they consult their dictionaries, what they like and dislike about them, and what kinds of problem they look to the dictionary to solve. It can take a variety of forms, such as questioning users, observing dictionary use, or setting up experiments in which users take part. It is useful to divide the field into *market research* (carried out by publishers) and *academic research* (carried out by teachers, researchers in universities, and sometimes lexicographers).

2.3.2.1 *Market research* Dictionary publishers regularly carry out (or claim to carry out) market research. This can take many forms, ranging from detailed questionnaires or surveys to informal conversations with teachers, students, and other users. These are usually 'internal' operations and results are rarely made public. On the other hand, publishers are alert to the PR benefits of being seen to be responsive to their customers' needs, so will often publicize the fact that they have carried out market research without being too specific about its methods or results. But there is no doubt that good market research often has direct and visible consequences for editorial policy (see Box 2.1).

In an interesting recent development, some publishers are using the internet as a medium for user research. The Macmillan Dictionaries website, for example (www.macmillandictionaries.com), provides supplementary materials such as lesson plans for using dictionaries in the classroom, a 'Word of the Week' feature, and a monthly e-zine with articles on a range of language issues. The service is free, but students and teachers register for it, thus creating a community of dictionary users. In planning the second edition of the *MED*, the publishers asked users to fill in a quite detailed online questionnaire. People seem more ready to cooperate with research conducted online, and this particular exercise got 1,331 responses – a significant amount of data about users' needs and preferences.

Monitoring the 'log files' of online dictionaries (which show exactly what people have looked up) may provide an even more direct way of

identifying users' needs and reference habits. In a fascinating paper, de Schryver and Joffe (2004) describe a project in which data of this type is 'directly integrated into the compilation of a reference work': the dictionary is available online as a 'work in progress' and an analysis of users' searches (including failed searches) has fed into a number of revisions. Using the same technique, the publishers can measure the effects of these revisions – which include an improvement in the percentage of successful searches.

Box 2.1 Market-research and its practical outcomes: two examples

LDOCE-2 (1987)

The General Introduction describes a market-research programme:

'We have conducted several research projects with schools and universities in various countries, including Belgium, Britain, France, Germany, Mexico, Nigeria, Japan, and the United States, to try to find out how effectively students make use of the information [in dictionaries]... This has enabled us to build up a clearer picture of learners' needs.'

The passage mentions a number of findings, among them this: 'although grammatical information is sometimes sought, most users found mnemonic codes off-putting and impenetrable'. In the new edition, the alphanumeric grammar codes found in *LDOCE-1* (such as [T5a] and [X9]) are abandoned in favour of more explicit ways of showing complementation. The point made in the introduction is that this change is a direct response to market research, which suggested that users did indeed need to know about grammar, but couldn't understand the codes in *LDOCE-1*.

Bloomsbury Concise English Dictionary (2005)

The introduction explains the genesis of the dictionary's usage notes:

'We assembled an Advisory Board of academics and teachers from around the [English-speaking] world... We sent our Advisory Board questionnaires eliciting their responses to broad questions like these: What is the most pervasive usage problem that you see in your students' writing?... What types of spelling problem do you see in your students' writing?...' Findings are reported and recurrent problems identified. The introduction goes on to describe the publisher's response: 'All these problems... are dealt with in the Dictionary's 600 Usage Notes, its A-Z list of 700 commonly misspelt words, and its 400 "Spellcheck" notes', all of which are said to be 'grounded in the classroom and reviewed and edited by English teachers'.

2.3.2.2 *Academic and lexicographic research* There is a large and growing body of user research by academics and (more rarely) by practising lexicographers, and several books have been devoted to the subject. Academics tend to focus on dictionary use in educational environments. Subjects are sometimes native speakers, as in Miller and Gildea (1985), a seminal paper on how well (or badly) American fifth and sixth graders understand dictionary definitions, or the two studies of college students' use of dictionaries, McCreary (2002) and McCreary and Amacker (2006). More often, they are language-learners of varying degrees of proficiency, cf. Bogaards (1992, 1998a). Lexicographers, in their research, have tried to discover how actual users use their actual dictionaries in as near natural settings as possible. An account of several such projects is to be found in Atkins (1998).

2.3.3 Know your user: conclusions

With characteristic gloominess, Samuel Johnson noted 'They that take a dictionary into their hands, have been accustomed to expect from it a solution of almost every difficulty.' A 'good dictionary' was once memorably defined by lexicographer Janet Whitcut in a conference intervention[6] as 'one that's got in it what you're looking for'. Users typically expect their dictionary to include every word they are ever likely to encounter, but in practice this can't happen, even with the best or biggest dictionary. Shortly after the publication of the *MED* (2002), football star David Beckham injured the metatarsal bone in his foot. Suddenly the word was everywhere, but *MED* had no entry for *metatarsal* (and neither did any other dictionary of its type). In subsequent updates, some of the learners' dictionaries added entries for the word. Yet, unless high-profile sportspeople continue to sustain such injuries, *metatarsal* will probably revert to its earlier status as a term used mainly among specialists, and the case for including it in a general-purpose learners' dictionary will be weak.

What this shows is that it is impossible to predict all the questions that users will ask of their dictionary, so we need to take a pragmatic view about what we can achieve. A realistic goal is to meet the needs of most users most of the time. And to achieve this, we have to get the clearest possible picture of who these users are and what kinds of question they will ask of their

[6] The First Fulbright Colloquium (on the emerging of lexicography as an international profession), London, 13–16 September 1984.

dictionary. Creating a user profile and taking careful note of relevant user research will help you to make well-informed editorial decisions.

2.3.4 Decisions affected by user profiling and user research

To sum up: with a clear idea of your users and their needs, you are well-placed to make decisions on a range of editorial and design issues, covering both content and presentation. In this section we set out a few questions to ask yourself when making editorial planning decisions.

2.3.4.1 *Content*

- Which headwords (and which meanings) should the dictionary include? Other questions in this area:
  - How many headwords does the dictionary need to contain?
  - Will users want to look up literary, dated, or obsolete words?
  - Should the dictionary include dialect words?
  - Should it cover specialist terms, and if so, which domains are most relevant to users?
- And, for each headword, which information categories are most important? Here, too, other questions arise:
  - Do your users know about (or need to know about) how words combine grammatically?
  - Do they need information about pronunciation or the stress patterns of phrases?
  - Do they already know how regular verbs inflect, or will you need to tell them this?
  - Do they need to know about typical contexts of the headword?

The answers to these questions may also impact on your corpus development programme. For example, editorial planning for the *Macmillan School Dictionary* (2004) – a book aimed at non-native speakers studying the full range of school subjects through the medium of English – started with an analysis of a built-for-purpose corpus. School textbooks and exam syllabuses for relevant subjects were collected from countries where the book would be sold, and frequency data from the resulting 20-million-word corpus provided the basis for headword selection.

2.3.4.2 *Presentation: metalanguage*

- What linguistic skills can you expect your users to have? Other questions that follow from this one:
  - Will definitions need to be written in simplified language?
  - Can we use IPA to show pronunciations?
  - Are users familiar with terms relating to transitivity, countability, and collocation?
- What reference skills can you assume in your users? Here we ask:
  - Will they understand 'standard' abbreviations (such as *adj*, *phr vb*, or *AmE*)?
  - Can you use 'codes' to indicate syntactic behaviour, or should this information be carefully spelled out?

2.3.4.3 *Presentation: design and layout*

- What is the best way to set out the material so that the dictionary is easy to use but still contains enough information?

The way information is presented makes a big difference to how easily users find what they are looking for, and how confident they feel about consulting their dictionary. Decisions in this area are generally made by the publisher and designer, but some input from the editorial team is essential and it is worth being aware of the issues.

Good design 'is intended to serve the reader by making the structure of the author's text clear in a visual form, and also by making the book pleasant to handle' (Luna 2004: 847). Traditionally dictionaries have shown certain worrying tendencies:

- They pack the maximum information into the smallest possible space, giving the page a very dense look (as for instance in the entry shown in Figure 2.3 above).
- They rely on variations in typeface to signal different information types: thus linguistic labels are often indicated by a change to italic type, cross-references are often shown in small capitals, and multiword expressions are usually in bold type.

Contemporary dictionaries have improved on the almost impenetrable layouts of earlier models through the use of more 'white space' and the practice of starting new meaning blocks – in longer entries at least – on a fresh line. But reliance on typeface variation remains heavy, and dictionary planners

must try to be realistic about whether their target users can recognize intended differences. As always, the test of the system is not whether it satisfies lexicographers' desire for order, but whether users actually understand the information being offered.

2.4 Tailoring the entry to the user who needs it

Once you have done your user profiling and have a good idea of the needs and skills of the typical user of your dictionary, you have to set about devising entries that meet these needs and build on these skills. What this means in practice can be seen from the comparison of three types of monolingual dictionaries (in §2.4.1 below) and from an analysis of a bilingual entry (in §2.4.2).

2.4.1 Monolingual dictionaries

In this section we look at the three major types of monolingual dictionary:

- for adult native speakers, represented by *CED-8* (2006)
- for school children, represented by *Collins School Dictionary* (2006)
- for adult learners, represented by *MED-1* (2002).

We'll compare their approaches to the same material, the verb *disturb* and its relatives *disturbed* and *disturbing* (omitting *disturbance*); the three entries are shown in Figure 2.5. Our focus will be on the content and the presentation of the entries.

When you are comparing dictionary entries it's a good idea to go about it systematically, so in comparing the *disturb* entries we'll look at the following features of these three dictionaries, and see how they reflect the user profile of each:

- content:
  - amount of information
  - type of facts in entry
  - wording of definitions
- presentation:
  - treatment of the 'word family' (as headwords or otherwise)
  - the way the words are divided into senses.

disturb (dɪ'stɜːb) *vb* (*tr*) **1** to intrude on; interrupt **2** to destroy or interrupt the quietness or peace of **3** to disarrange; muddle **4** (*often passive*) to upset or agitate; trouble: *I am disturbed at your bad news* **5** to inconvenience; put out: *don't disturb yourself on my account* [C13 from Latin *disturbare*, from DIS-[1] + *turbare* to confuse] > **dis'turber**.

disturbance […]

disturbed (dɪ'stɜːbd) *adj psychiatry* emotionally upset, troubled, or maladjusted.

disturbing (dɪ'stɜːbɪŋ) *adj* tending to upset or agitate; troubling; worrying > **dis'turbingly** *adv*

CED-8 (2006)

disturb, **disturbs, disturbing, disturbed**

1 (VERB) If you disturb someone, you break their rest, peace, or privacy. **2** If something disturbs you, it makes you feel upset or worried. **3** If something is disturbed, it is moved out of position or meddled with.

disturbing (adjective).

disturbance […]

Collins School Dictionary (2006)

disturb /dɪ'stɜːb/ verb [T] **

1 to interrupt someone and stop them from continuing what they were doing: *I didn't want to disturb you in the middle of a meeting.* ♦ *Sorry to disturb you, but do you know where Miss Springer is?* ♦ *Her sleep was disturbed by a violent hammering on the door.*

2 to upset and worry someone a lot: *Ministers declared themselves profoundly disturbed by the violence.*

3 to make something move: *A soft breeze gently disturbed the surface of the pool.* **3a**. to frighten wild animals or birds so that they run away.

4 to do something that stops a place or situation from being pleasant, calm, or peaceful: *Not even a breath of wind disturbed the beautiful scene.*

disturb the peace *legal* to commit the illegal act of behaving in a noisy way in public, especially late at night

do not disturb a sign that you hang on a door, especially in a hotel or an office, to say that you do not want to be interrupted

disturbance […]

disturbed /dɪ'stɜːbd/ adj *

1 affected by mental or emotional problems, usually because of bad experiences in the past: *These are very disturbed children who need help.*

2 extremely upset and worried: *I am very disturbed by the complaints that have been made against you.*

disturbing /dɪ'stɜːbɪŋ/ adj * making you feel extremely worried or upset: *I found the book deeply disturbing.* ♦ *disturbing images of war and death.*

—**disturbingly** adv: *The crimes were disturbingly similar.*

MED-1 (2002)

Fig 2.5 The *disturb* entry: three different approaches

2.4.1.1 *Comparing the content of entries* As befits their respective functions, these three dictionaries are all different sizes and formats: the *CED* is a large collegiate dictionary with about 120,000 headwords; the *CSD* is a concise dictionary of around 14,500 headwords, with larger print and a lot of white space; and the *MED* is a standard volume of about 46,000 headwords and quite compressed text. All three of the entries compared are good entries from good dictionaries. Nonetheless, some anomalies are apparent in this brief analysis. As we all find on reading reviews of our

dictionaries, anyone shining a spotlight on a single entry will find something to complain about. No one can guarantee total consistency throughout a set of (say) 50,000 entries compiled by (say) a dozen lexicographers over a period of (say) three or four years.

Amount of information

- A quick comparison of the three entries is enough to show that the adult learner clearly needs more help than either the adult or the young native speaker. That figures – native speakers have their own linguistic instincts to rely on when it comes to putting a 'new' word into a sentence, or trying to use it.
- Most of the additional information in the *MED* has to do with encoding rather than decoding.

Type of facts in entry

- The most interesting point is the fact that the learners' dictionary identifies by the bold typeface two 'multiword expressions' (cf. §7.2.7.1 for an explanation of this term). These are the legal idiom *disturb the peace* and the sign *do not disturb*, and the dictionary explains each of them. The others do not, although an adult native speaker at least might reasonably be supposed to need a definition for the legal phrase.
- There are no examples in the children's dictionary (possibly owing to length restrictions), while the learners' dictionary is understandably very rich in good informative examples to help learners slot the word into their passive – and hopefully active – vocabulary. The function of the examples in the *CED* is probably rather to help users sort out the various already known meanings of the word.
- A justifiable omission from the *CSD* is the word *disturbingly*.
- The dictionaries for adults, whether native speakers or learners, give the pronunciation in the International Phonetic Alphabet (IPA); the editors of the schools dictionary rightly believed their readers couldn't handle IPA and mostly omit pronunciation, including it only for some of the more difficult words in a form of respelling, shown in the *CSD* entry in Figure 7.3 in Chapter 7.
- The only grammatical information in the children's dictionary is the wordclass of the headword, and even then you wonder whether 'verb' and 'adjective' mean very much to the majority of its readers; transitivity is specified in the adults' dictionaries ('*tr*' and 'T'); and the fact that

in the 'upsetting' sense *disturb* is often passive is given in the dictionary for native speakers (who presumably don't need this) but not in the one for learners (who probably do).

- Both the dictionaries for adults include a domain label ('psychiatry' in the *CED* and 'legal' in the *MED*); it would have made sense for both of them to include both labels, but it's realistic to omit them in children's dictionaries.
- The adult native-speaker dictionary (*CED*) contains etymologies; the *MED* (in common with most learners' dictionaries) does not, although these are starting to appear in some electronic versions.
- Only the learners' dictionary mentions corpus frequency (using asterisks to indicate how common the words *disturb*, *disturbed*, and *disturbing* are). It's reasonable to think that this information would hold little interest for native speakers.

Wording of the definitions

- *CED* tends to define by means of semi-synonyms instead of a paraphrase (senses 1, 3, 4, and 5).This technique is rightly eschewed in *CSD* and *MED*, where the paraphrase definitions are longer, but much more user-friendly, and allow the editors to describe the meaning of the headword for the most part in simpler and easier words, as is appropriate for children and language-learners.
- As a result, the definitions in *CED* contain some words (*agitate*, *maladjusted*) that might be expected to challenge a user who needs to look up *disturb*; those for children and adult learners are more instantly comprehensible.
- The conversational format of the 'if…' definitions in the children's dictionary (pioneered in *COBUILD* from which the *CSD* is derived) make for a much more natural-sounding description of meaning, as though the dictionary were answering its users' question 'What does this word mean?'.[7]

2.4.1.2 *Comparing the layout of entries* As was the case with the content of the three entries, their various layouts reflect the editors' awareness of the intended users.

[7] This type of defining has its champions and detractors: defining is discussed in detail in Chapter 10.

Treatment of the 'word family'

- The most user-friendly way of setting out word meanings in a dictionary is to make every searchable word a headword: that way, your user is less likely to overlook it. Not every dictionary however can find the space to do this.
- While the words *disturb* and *disturbed* are given headword status in all three dictionaries, *disturbing* occurs as a headword in *CED* and *MED* but not in the schools dictionary. The *CSD*, smaller in format and subject to harsh length constraints, presumably decided not to devote space to making headwords of semantically transparent adjectives in *-ing*. (But it would have been more user-friendly to make all searchable words into headwords.) The word *disturbingly* is a run-on with only wordclass information in *CED*, while in the *MED* it is given an example sentence as well; it doesn't appear at all in the schools dictionary. The rare word *disturber* is attested in the *CED*, presumably to reassure people who suspect it might exist or don't know how to spell it (or both). All these decisions seem quite in keeping with the dictionaries' various user profiles.

Senses of the headword

- As often with sense differentiation – not an exact science (cf. the discussion in Chapter 8) – the division of the headword *disturb* into senses is hard to reconcile with the various dictionaries' targeted users, although the fact that the largest dictionary (*CED*: for adult native speakers) splits the word into the most senses (5), and the smallest dictionary (*CSD*: for children) into the least (3) is fairly typical.
- Without examples, it's difficult to see the difference that *CED* draws between its senses 1 and 2 (although *CED*'s sense 1 is probably *MED*'s sense 1, which also covers *CED*'s sense 5, and *CED*'s sense 2 is probably *MED*'s sense 4). *CSD* is very adequate for its young readers, who would probably be confused by more detail.

2.4.2 Bilingual dictionaries

Before we can say much about bilingual dictionaries, there are a few concepts and terms that we should clarify. Consider your own bilingual dictionary, or one you know well.

- As we saw already, it may be a 'unidirectional' dictionary, i.e. it consists of a single text from Language A (the source language, or SL) to Language B (the target language, TL).
- It may be a 'bidirectional' dictionary, i.e. it contains two distinct texts in one volume:
  - one from Language A to Language B, and
  - one from Language B to Language A.

Consider now a single unidirectional text (i.e. from Language A to Language B): this will be half of a bidirectional dictionary, and the whole of a unidirectional one.

- If your own language is the SL then your dictionary is an 'encoding' dictionary (sometimes called an 'active' dictionary).
- If your own language is the TL, then your dictionary is a 'decoding' dictionary (or a 'passive' dictionary).

If your dictionary is a bidirectional dictionary, selling to speakers of Language A and those of Language B, then each of the two sections within the one volume will have to serve a double purpose:

- The A→B section must be simultaneously
  - an encoding dictionary for Language-A speakers, i.e. speakers of the SL
  - a decoding dictionary for Language-B speakers, i.e. speakers of the TL.
- The B→A section must be simultaneously
  - an encoding dictionary for Language-B (SL) speakers
  - a decoding dictionary for Language-A (TL) speakers.

What does this mean in practice? It means that all but the shortest entries have a high level of redundancy for both sets of speakers. §2.4.2.2 looks at this in more detail.

➔ The SL speakers need all the help they can get. Don't short-change them. They're trying to write in a foreign language. The TL speakers can wing it if the worst comes to the worst. They're writing in their own language – as long as they can understand what the foreign word means, they'll manage.

2.4.2.1 *For one language group* The simplest bilingual dictionary to write is a decoding dictionary for one language group, i.e. one destined for speakers of a single language (the TL) who want to translate into their own language. The next simplest is an encoding dictionary for one language group, i.e. speakers of the SL who need to translate into or express themselves in a foreign language.

Targeting a dictionary at a single language group means essentially that the dictionary is designed to be sold in a single market. Not too many English speakers learn Finnish, so a bilingual English and Finnish dictionary would normally be produced in Finland for speakers of Finnish. It might very well be sold in English-speaking countries, but it's not likely to be designed for use by English speakers, as that would make the entries much more complex, they would take longer to compile, the book would be bigger, the whole thing would cost much more, and the publishers would never sell enough copies in the English market to make it worthwhile. So if you're an English speaker having problems with your English-Finnish dictionary, it may not be all your fault. All the metalanguage will be in Finnish; the English-Finnish text will be written specifically for the decoding Finnish (TL) speaker and the Finnish-English text for the encoding Finnish (SL) speaker.

The result is a dictionary entry like the one shown in Figure 2.6, which is an example of the simplest bilingual entry: *a decoding entry for one language group*, taken from the *English-Finnish General Dictionary* (published by Werner Söderström Oy, Helsinki, 1998).

disturb /…/ *tr* **1** a) häiritä *(I hope I'm not ~ing you);* b) sekoittaa, järkyttää *(the balance)* **2** panna sekaisin, sotkea; siirrellä *(he found that the papers on his desk had been ~ed)*; muuttaa, muutella; koskea jhk *(do not ~ the screws).* […] **~ed** *a* **1** levoton **2** sielullisesti häiriintynyt; *~ ward* rauhattomien potilaiden osasto.

Fig 2.6 A decoding entry for TL (Finnish) speakers

Pretend you speak English, but no Finnish. Look at the entry and decide which Finnish word would you choose for *disturb* in the translation of these sentences from the British National Corpus:

She ensured that others did not disturb him when he was at his books.
I'm sorry if my questions disturb you.
Sorry to disturb you, but I have to ask…

The contents of each drawer had been disturbed.
The ground had been freshly disturbed.
The animal will only attack if it is disturbed.
He didn't seem unduly disturbed.

Not much hope of success there! And yet with this entry, no Finnish speaker would have problems with understanding *disturb* in these sentences, and finding and using the correct Finnish equivalent.

Figure 2.7 contains an entry for *hopefully* from the *CRFD-8* (2006), written expressly to help English speakers translate this word into French, or express the concept in French. That is to say, it is an encoding entry for one language group, the SL speakers.

hopefully /ˈhəʊpfəlɪ/ ADV 1 (= *optimistically*) [*say, look at*] avec espoir ♦ **... she asked** ~ **...** demanda-t-elle pleine d'espoir. 2 (* = *one hopes*) avec un peu de chance ♦ ~ **we'll be able to find a solution** avec un peu de chance, nous retrouverons une solution ♦ ~ **it won't rain** on espère qu'il ne va pas pleuvoir ♦ **(yes)** ~ **!** je l'espère !, j'espère bien ! ♦ ~ **not!** j'espère que non !

Fig 2.7 An encoding entry for SL (English) speakers

This entry has been carefully thought out, with due regard to the contexts in which *hopefully* occurs, and an English speaker should have no problem finding the correct French for this word in the following BNC sentences:

We waited hopefully.
Hopefully he'll recover well and be back to normal.
The four horses gazed at them hopefully.
Hopefully I will be fighting fit in two weeks.
It could be it disappearing . . . hopefully it is!
Ros looks at him hopefully.
Did it leave a lot of marks? No . . . hopefully not.
There is a tour of Ireland, Wales and hopefully South Africa . . .

2.4.2.2 *For two language groups* However, the *CRFD* entry in Figure 2.7 actually comes from a dictionary which sells in two markets: the English-speaking and the French-speaking. Its entries must therefore serve a dual purpose, and this one must act as:

- an encoding entry for English speakers, and
- a decoding entry for French speakers.

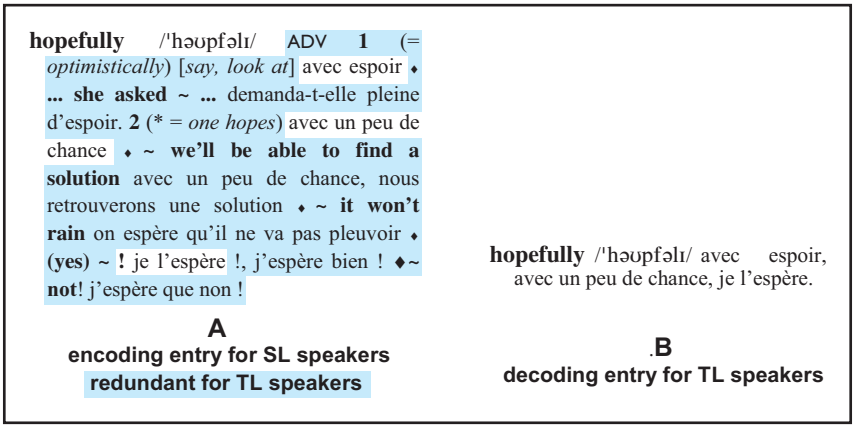


Fig 2.8 Encoding and decoding versions of the same entry

But when we consider the amount of information in the entry, it's clear that there is far more than the French speaker needs. In Figure 2.8, entry (A) has the redundant information shaded, and it has been removed entirely from version (B). All French speakers need is to be told how to pronounce the word and to be given one or two equivalents in their own language. They know how to use them.

These contrastive versions of the bilingual entry for *hopefully* are proof of how much the users' skills can influence the essential information in the entry. This is true of every type of dictionary, but of course – as in so many respects – the bilingual dictionary is more complex, and less amenable to clear explanations, than all but the most scholarly and sophisticated of the monolinguals.

Exercise

Choose a dictionary you are familiar with. Then . . .

1. Describe the dictionary in terms of its properties:
   - Make a list of the properties. (cf. §2.2.1)
   - Which dictionary type best matches your list of properties? (cf. §2.2.2)
2. Draw up a user profile for this dictionary in terms of the following:
   - types of user (cf. §2.3.1.1)
   - ways in which they will want to use the dictionary (cf. §2.3.1.2)
   - the skills they bring to the task (cf. §2.3.1.3).

3. Select one page of the dictionary, and on the basis of that page . . .
   - List as many points as you can which are good in the light of the user profile.
   - Make a note of any feature which could prove difficult for the dictionary's intended users.
   - Suggest ways of making the dictionary more suitable for the intended users.

Reading

Recommended reading

Dictionary types: Atkins 1985.
Dictionary use: Atkins and Varantola 1997, 1998; Hulstijn and Atkins 1998; Miller and Gildea 1985.

Further reading on related topics

Dictionary types: Hausmann and Wiegand 1989.
Dictionary use: Bogaards 1990, 1992, 1996, 1998a, 1998b; Bogaards and van der Kloot 2001; de Schryver and Prinsloo 2000; Lew 2002, 2004; Mackintosh 1998; Marello 1998; Martin and Al 1988; McCreary 2002; McCreary and Amacker 2006; McCreary and Dolezal 1999; Nesi 2000; Nesi and Haill 2002; Nuccorini 1994; Varantola 1998.
Dictionary design: Luna 2004.

Websites

Yukio Tono's Bibliography of Dictionary User Studies: http://leo.meikai.ac.jp/~tono/userstudy/userbiblio.htm

3

Lexicographic evidence

This chapter (see Figure 3.1) explains how to design, acquire, and process a collection of linguistic data which will form the raw material for your dictionary. We will look first at *citations* and then – in greater detail – at lexicographic *corpora*. Software for querying the data in a corpus is discussed in the next chapter (§4.3.1), and the process of analysing corpus data to create dictionary text is covered in Chapters 8 and 9.

3.1 What makes a dictionary 'reliable'?

Dictionaries describe the vocabulary of a language. For any given word, a good dictionary tells its readers the ways in which that word typically contributes to the meaning of an utterance, the ways in which it combines with other words, the types of text that it tends to occur in, and so on. Clearly it is desirable that this account is reliable. A reliable dictionary is one whose generalizations about word behaviour approximate closely to the ways in which people normally use (and understand) language when engaging in real communicative acts (such as writing novels or business

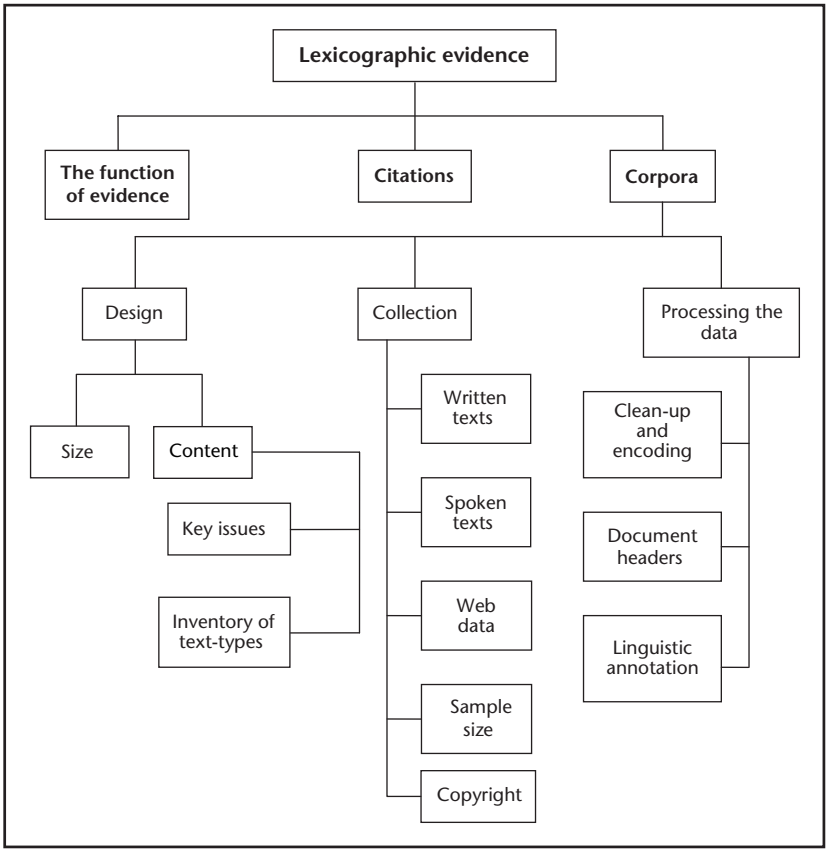


Fig 3.1 Contents of this chapter

reports, reading newspapers, or having conversations). But how can we feel confident that we *know* how people normally use words, and hence that the description given in our dictionary is reliable? Reliability depends on the kind of evidence that underpins our account of the language – and evidence comes in several forms.

3.1.1 Subjective evidence and its limits

'Introspection' is a form of evidence. It describes the process in which you give an account of a word and its meaning by consulting your own mental lexicon (all the knowledge about words and language stored in your brain), and retrieving relevant facts. Introspection is a useful device which we use all the time – for example when a child asks us what something means, or

when a friend from another speech community asks us whether we also use a particular expression which is familiar to her or him. But introspection alone can't form the basis of a reliable dictionary. Even if we assume that we have full access to the contents of our mental lexicon, one individual's store of linguistic knowledge is inevitably incomplete and idiosyncratic. At best it will furnish a moderately accurate, but necessarily partial, account of language use. At worst, we may find that there is a significant disparity between how we think words are used and how people actually use them. This is easily demonstrated. If you try, through introspection, to retrieve everything you know about the meanings and combinatorial behaviour of a fairly complex word, and then check your findings against a dictionary (or better, a corpus), you will almost certainly find there are gaps in your account, and there may be some misconceptions too about how the word is really used.

For similar reasons, informant-testing, in which speakers of a language are questioned about their use of words, is also of limited value for mainstream lexicography. It is a method that has been used extensively for cataloguing the vocabulary of languages which exist only in oral form. But, like introspection, it is essentially a subjective form of evidence. For the purposes of this chapter, 'evidence' refers not to people's reflections or intuitions about how words are used, but to what we learn by observing language in use. Objective evidence, in other words. This means looking at what speakers and writers actually do when they communicate with listeners and readers. Creating a reliable dictionary involves a number of challenging tasks, but the observation of language in use is the indispensable first stage in the process.

3.1.2 The scope of the dictionary

Language in use, however, is a moving target. It is a dynamic system which tolerates a good deal of variation, creativity, and idiosyncrasy. Speakers of English comprise a very large and very diverse speech community. Not only that, we know that individual members of any speech community will sometimes use language in eccentric ways. In his award-winning novel *Vernon God Little* (Canongate Books, 2003), the writer D. B. C. Pierre describes the weather in a small town in Texas as 'bitterly hot', and in a later passage he tells us that 'silence erupted'. Both combinations are highly

atypical[1]; indeed, they depend for their effect on the reader recognizing that Pierre is deliberately violating the norms of the language. For a variety of reasons, individual speakers and writers may consciously depart from 'normal' modes of expression. How do we cope with this as lexicographers? As always, the answer will depend to some extent on 'users and uses' (§2.3): the kinds of people the dictionary is designed for and the reference needs which the dictionary aims to cater for. But a good basic principle is that (with the possible exception of large historical dictionaries), the job of the dictionary is to describe and explain linguistic *conventions* – the ways in which people generally use words – rather than trying to account for every individual language event. Our focus, in other words, must be the probable, not the possible.[2]

If our goal is to provide 'typifications', then how do we know whether a given utterance is typical (and therefore worth describing) or merely idiosyncratic (and therefore outside our remit)? A typical linguistic feature is one that is both *frequent* and *well-dispersed*. Any usage which occurs frequently in a corpus, and is also found in a variety of text-types, can confidently be regarded as belonging to the stable 'core' of the language. It is part of the climate, rather than the weather, to use Halliday's illuminating analogy[3] – and this is what we will focus on as lexicographers.

3.2 Citations

3.2.1 What are citations and how do you find them?

Until about 1980, the main form of empirical language data available to lexicographers was the *citation*. A citation is a short extract from a text which provides evidence for a word, phrase, usage, or meaning in authentic use. The use of citations as lexicographic evidence pre-dates Samuel Johnson, but Johnson was the first English lexicographer to use citations

[1] Intuition suggests this, statistics confirm it: a count using Google shows that 'bitterly cold' is about 3,000 times more common than 'bitterly hot'.

[2] This point has been made most eloquently by Patrick Hanks (2001), who speculates that lexicons of the future will 'focus on determining the probabilities, and associating them with prototypical contexts, rather than seeking to cover all possible meanings and all possible uses'.

[3] M. A. K. Halliday, 'Corpus studies and probabilistic grammar', in Aijmer, K. and Altenberg, B. (Eds), *English Corpus Linguistics*. London: Longman (1991), 30–43.

Box 3.1 Rationalism and empiricism: two approaches to understanding language

Lexicographers (and corpus linguists generally) are empiricists. What we are interested in is describing 'performance' (what writers and speakers do when they communicate). We do this by observing language in use and – on the basis of this – attempting to make useful generalizations that will account for phenomena in the language which appear to be recurrent. Another major tradition in linguistics is represented by the rationalists, whose goal is to describe linguistic 'competence': the internalized, but subconscious, knowledge we have of the rules underlying the production and understanding of our mother tongue. This tradition is associated most obviously with Noam Chomsky. For linguists working in this paradigm, 'data' derives from introspection rather than observation. Until the 1950s, there was a thriving empiricist tradition in American linguistics, but 'in a series of influential publications [Chomsky] changed the direction of linguistics away from empiricism and towards rationalism in a remarkably short time' (McEnery and Wilson 2001: 5). It is easy to caricature this major division, and there are lively debates (for example on the CORPORA discussion list) in which Chomskyites are demonized as 'the enemy' of corpus-based approaches. As always, the truth is a little more nuanced than this neat, binary characterization implies. Nevertheless, Chomsky[4] is on record as being sceptical about the value of corpora, and a recent interview shows that his stance has not shifted. He says:

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides . . . that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights.
>
> (p.97 in Andor, Jozsef. (2004) 'The Master and his Performance: An Interview with Noam Chomsky', in *Intercultural Pragmatics* 1.1: 93–111)

With Chomsky's star in the ascendant, early corpus linguists like the team responsible for the Brown Corpus (§3.4.1) were working very much against the grain of the prevailing orthodoxy. But now that technology can provide us with very large bodies of linguistic data, the empiricist tradition has moved closer to the mainstream.

[4] Thanks to Ramesh Krishnamurthy for the quotation from Chomsky.

systematically. The description of English found in the *OED* famously draws upon the many millions of citations that were collected (mainly by volunteers) from the 1860s onwards. Until the late twentieth century, the *OED*'s citations would be written in longhand on index cards (known as 'slips'). These were filed alphabetically according to the keyword of the citation, then retrieved from the files to be used by James Murray and his colleagues and successors as the primary data source for every entry in the dictionary. Figure 3.2 shows what a typical citation looks like.

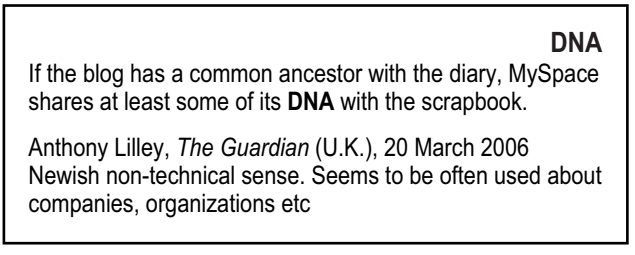


Fig 3.2 A citation for the non-technical use of 'DNA'

➔ To find some citations, read a page or two of text – for example from a newspaper, a contemporary novel, or a blog. Make a note whenever you come across a word, phrase, or meaning which strikes you as novel or unusual, and which you suspect is not currently accounted for in your dictionary. Then record (either on a card or on your computer) the sentence containing the usage you are interested in. Do this in a form that identifies the headword where this usage would be entered (assuming it makes it into the dictionary), and indicate the source of the citation. Almost everyone who tries this is surprised by how easy it is to find instances of language in use which have not yet been recorded in any dictionary.

3.2.2 Setting up a reading programme

Some dictionary publishers provide online 'forms' to enable members of the public to contribute citations. Most publishers' experience of this data-collection model is that the ratio of unusable citations to good ones is high, so that a great deal of activity yields relatively little in the way of genuinely new and useful data. A well-planned 'reading programme', on the other hand, will often have great value. A reading programme is an organized data-gathering exercise, in which the publisher identifies target

texts, recruits and trains readers who will scour these texts for citations, and provides a structured way of recording the resulting data. Traditionally, incoming citations would be filed in the form of slips, but nowadays they will typically be recorded using a web-input form, with a database behind it for storing and sorting the citations. The amount of information that readers are required to supply depends on the type of dictionary that will use the citations, but you will need at least four main data fields (all of which are illustrated in Figure 3.2):

- keyword or phrase: the usage that your citation illustrates, filed under the headword to which it relates
- the citation itself: usually a single sentence is adequate but you may sometimes need more
- information about the source of the citation: the date, title, and author's name are all important; additional information (such as the page number where the citation appears, or full bibliographic details on the source text) may be useful for specialized or historical dictionaries, but are generally not needed
- a comment field: this gives readers the option of adding a note to clarify the citation; it may, for example, be a new meaning that needs explaining, or it may be characteristic of one particular dialect (as in the case of the expression 'the guards', a common way of referring to the police in Ireland, but virtually unknown in the rest of the English-speaking world).

Storing the data in a computer database, of course, will greatly enhance its value: citations can be grouped according to any of the input parameters (their date, for example), and the entire content of any citation (rather than just the keyword) can be retrieved and can, in turn, be used as Linguistic data.

3.2.3 Citations: advantages and disadvantages

The benefits of a reading programme include:

- Monitoring language change: even in the age of Google, citation reading remains an efficient way of tracking developments in the language. It's easy for computer programs to spot completely new words (like *blogosphere*), but a high proportion of 'new' vocabulary consists of compounds, multiword expressions, and novel uses of existing words

(like *DNA*) – and this is where human readers still have a significant edge.

- Gathering terminology from a specific subject field or a particular variety or dialect: a publisher can give a reader a collection of titles relating to basketball, or titles written in Jamaican English, and rapidly acquire a body of relevant citations.
- Training lexicographers: collecting citations requires you to think about what 'counts' as an item to be described in a dictionary, and to distinguish genuinely new usages from *ad hoc* coinages. This makes it a good way of raising awareness of many of the issues that lexicographers have to make judgments on.

The disadvantages of this form of evidence include:

- Collecting data in this way is labour-intensive, so volumes will always be low. It is true that the two great historical English dictionaries (the *OED* and *Merriam-Webster*) have many millions of citations between them, but these have been collected over more than a century. Even so, the evidence they provide for contemporary language is relatively thin compared with what a large corpus will deliver.
- Although instances of usage are authentic, there is a big subjective element in their selection. As Noah Webster and James Murray both observed, human readers tend to notice what is remarkable and ignore what is typical, and this creates a bias towards the novel or idiosyncratic usages which inevitably catch the reader's eye. When reviewing the data for the letter A, Murray remarked on the imbalance between rare and common uses: 'Of Abusion, we found in slips about 50 instances: Of Abuse not five.'[5] In the 'Additional Notes' to his 'Directions for Readers' (1879), he rather tetchily asks readers to 'kindly remember that the Dictionary is to contain all English words ordinary and extraordinary included'.[6] Of course, Murray's concerns about the poverty of data for common words are resolved by modern corpora.

The arrival of the web gives a new angle to citation reading: a manually collected citation can be checked against the vast resources of the internet.

[5] Eighth Annual Address of the President to the Philological Society, *Transactions of the Philological Society* (1877–79), 561–586.

[6] Murray's various 'Appeals to Readers' can be found on the OED's website: www.oed.com .

If you encounter an unfamiliar idiom and want to find out whether it is frequent or rare, widespread or region-specific, a search on the web will usually provide the answers. And if you aren't sure whether a particular usage is still current, a site like Google News will show how recently it has been used (which usually turns out to be within the last 24 hours).[7]

Citation reading continues to have value, especially as a form of lexicographic training. But now that most written texts (including very old texts) are available in digital form, it has become a more marginal way of collecting linguistic data. The corpus has moved to centre stage.

3.3 Corpora: introductory remarks

English corpora designed for use in lexicography have been around since the beginning of the 1980s. Anyone embarking on the creation of a lexicographic corpus can therefore draw on a set of guiding principles and a body of good practice which have evolved during the intervening period. All of these issues will be discussed here. We also need to be aware that – just as the advent of corpora transformed the way lexicographers work – the arrival of the web, and its rapid growth and penetration, changes the landscape once more, often in quite far-reaching ways. The rest of this chapter will deal with the three major aspects of corpus creation:

- design: selecting the texts that will make up your corpus
- data collection: acquiring these texts
- encoding: converting constituent texts to a common format, and making them ready for use in a corpus-querying system.

And as we deal with the different phases in the corpus creation process, we will show how ideas and methods developed in the pre-web era may need to be modified in light of changing circumstances.

3.3.1 The central role of the corpus

Objective evidence of language in use is a fundamental prerequisite for a reliable dictionary. Traditionally, such evidence was found in collections of

[7] One might have imagined, for example, that the phrase 'Beam me up, Scotty' had fallen into disuse, but web data shows that it is alive and well.

citations, but these have their limitations (§3.2.3). If the dictionary's function is, as we have argued, to focus on 'normal' language events, it follows that you need very large volumes of data: normal language events are those which are *recurrent*, which can be observed to take place frequently and in different types of text. So we can only confidently distinguish what is conventional from what is idiosyncratic if we have plenty of data at our disposal. Citation banks alone – even the largest ones – can't usually supply language data in the required volumes, so the case for a large corpus is clear.

What do we mean by the term 'corpus'? One well-known definition comes from John Sinclair, who pioneered the use of corpora for lexicography in the early 1980s:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.
>
> (Sinclair 2005: 16)

This is not without its problems. The idea that a corpus can 'represent a language' is contentious (§3.4.2.2), and this in turn calls into question the theoretical validity of 'external' selection criteria. Arguably, therefore, the term 'corpus' should be extended to *any* collection of text in electronic form when it is viewed as a source of data for linguistic research. Our focus here, however, is not on corpora in general but on the *lexicographic* corpus – a collection of language data designed specifically for use in the creation of dictionaries. And at the very least, lexicographers need to know what sort of data they are using and where it comes from. So Sinclair's definition is a good starting point, even if we find that we need to modify it to take account of recent developments on the web.

3.3.2 Some inescapable truths

There is no such thing as a perfect corpus for lexicography, and it is important to be clear about this from the outset. So we will begin with a few caveats, noting some of the constraints within which corpus developers have to work. We will then outline the characteristics of a corpus that will – within those limitations – provide the best possible raw materials for writing a dictionary.

3.3.2.1 *The corpus is a sample* For a few languages (such as Ancient Greek or Old English), it is possible to collect and examine every extant example

of usage. Such a corpus would provide a complete record of the surviving evidence for its target language. But in most cases this is impossible – how could you collect every instance of Japanese in use, for example? So most corpora will comprise a subset (usually a very small subset) of all of the communicative events of the language under investigation. It must, in other words, be a sample. To create a sample that fairly reflects the wider population, you need clear selection criteria, and these will be determined by your corpus's intended function. People use corpora for all sorts of purposes, many of them highly specialized.[8] As a general rule, the more precise and well-defined the application, the easier it is to establish criteria for selecting texts. Lexicography, however, lies at the other end of this spectrum: a corpus designed for use in dictionary-making must cover a very wide range of text-types, and devising a sample that achieves this aim involves significant challenges.

3.3.2.2 *The corpus does not favour 'high quality' language* When Samuel Johnson was assembling the raw materials for his dictionary, one of his stated objectives was 'to preserve the purity . . . of our English idiom'.[9] Given his aim of reversing a perceived decline in the quality of written English, it was, for Johnson, 'an obvious rule' that his source texts should come only from 'writers of the first reputation'. This idea that dictionaries exist in order to uphold standards, and to adjudicate between 'good' and 'bad' usage, has widespread popular appeal. But it is fraught with difficulty. Selecting texts on the basis of their 'quality', and excluding those which fail this test, is fundamentally at odds with the descriptive (as opposed to prescriptive) ethos of corpus linguistics. Who is to judge which texts are 'good', and on what basis? The whole point of using corpora is to avoid pre-judging the data and choosing texts because you approve of them in some way.[10] In fact, even Johnson relaxed his initially didactic stance in the

[8] The Proceedings of the Corpus Linguistics conferences held biennially in the UK give a good idea of the range of topics for which researchers create and use corpora: see the archive at http://www.corpus.bham.ac.uk/conference2007/index.htm .

[9] Johnson, *Plan* (1747: 4).

[10] For further discussion, see Kilgarriff, Rundell, and Uí Dhonnchadha (2007: 131f.). The importance of this point was grasped long ago by Leonard Bloomfield, who observed (in *Language* (1933) Chapter 2, §2.9): 'He [the linguistic observer] must not select or distort the facts according to his views of what speakers ought to be saying'.

course of writing his dictionary.[11] A century later Richard Chenevix Trench, one of the founding fathers of the *OED*, argued convincingly for a non-judgmental approach to the description of language.[12] In characterizing the lexicographer as 'an historian, not a critic', Trench helped to establish the basic principles within which modern English lexicography operates. If we follow these principles, it is clear that a lexicographic corpus must be a genuine – and *inclusive* – snapshot of a language, not a set of texts that have been specially chosen to advance someone's notion of what constitutes 'good' usage.

3.3.2.3 *Pragmatism and compromise* Corpus creation is a pragmatic enterprise. For all sorts of reasons, corpus developers will find themselves making compromises between what they would ideally like to do and what is feasible within normal time and budget constraints. The need for compromise extends to all three phases described in this chapter: design, data-collection, and encoding.

Take design: the texts for a corpus should be selected using criteria which are transparent and well-argued, but we should not delude ourselves that this selection process is (or can be) 'scientific'. The British National Corpus (BNC) is the best pre-web corpus of English. Well-balanced, meticulously encoded, and with the highest level of copyright clearance, it has (rightly) been seen as a 'gold standard' for corpus developers everywhere. The content of the BNC (its individual texts, broad text-types, and proportions of each) was specified by a committee of academics and publishers. They considered relevant theoretical arguments, took account of previous work in the field, and generally went to great lengths to ensure a good range and balance of texts. The resulting configuration is thoroughly reasonable. It nevertheless represents no more than the subjective decisions of one group of people – albeit a well-informed group – about what a good corpus should look like.

[11] 'When first I collected these authorities, I was desirous that every quotation should be useful to some other end than the illustration of a word . . . Such is design, while it is yet at a distance from execution. When the time called upon me to range this accumulation of elegance and wisdom into an alphabetical series, I . . . was forced to depart from my scheme of including all that was pleasing or useful in *English* literature' (*Preface* 1755).

[12] Trench's seminal paper, 'On some deficiencies in our English Dictionaries' (1857), can be found on the *OED*'s website: http://www.oed.com/archive/paper-deficiencies .

Once we have decided which texts we ideally want to include, various non-linguistic factors may force us to change our minds: some authors may refuse to allow their books to appear in the corpus; one text may be substituted for another if the first doesn't exist in digital form but a reasonable alternative does; collecting good spoken data may turn out to be more labour-intensive than we thought; and so on. And once we have our texts, the level of detail to which they can be encoded (for bibliographic data or linguistic features) is, as we shall see later, almost infinitely variable, so issues of finance will come into play. It should be clear, then, that pragmatic choices have to be made all the way through the process.

3.4 Corpora: design issues

Designing a corpus means making decisions about:

- how large it will be
- which broad categories of text it will include
- what proportions of each category it will include
- which individual texts it will include.

This section discusses the factors that bear on these decisions and the arguments that inform them.

3.4.1 Size: how large is large enough?

For major languages like English, data sparseness is a thing of the past and corpus size has almost ceased to be an issue. Language data for most types of text is now available in vast quantities,[13] and the technical constraints that once made corpus-building such a daunting enterprise have largely disappeared. Most texts already exist in digital form (relieving us of the costly and labour-intensive business of keyboarding or scanning), while the requirement for large-scale data storage and powerful data-processing can easily be met by the average personal computer. In any case, the most usual arrangement now is that the corpus sits on a remote server, so the only technology you need is a web browser and fast internet connection.

[13] For more on mega corpora created from web data, see §3.5.3 below.

But it was not always like this. In the early days of corpus lexicography, there was never enough data, and processing even a few million words of text stretched the available technology to its limits. So corpus size was a major preoccupation, and two of the most important developments in the field – the Birmingham corpus in the early 1980s, and the BNC a decade later – were driven primarily by a desire to provide lexicographers with linguistic data in much greater volumes than anything currently available. As Figure 3.3 shows, corpora have increased in size by roughly one order of magnitude in each decade since the 1970s, and there are now no technical limits to further growth.

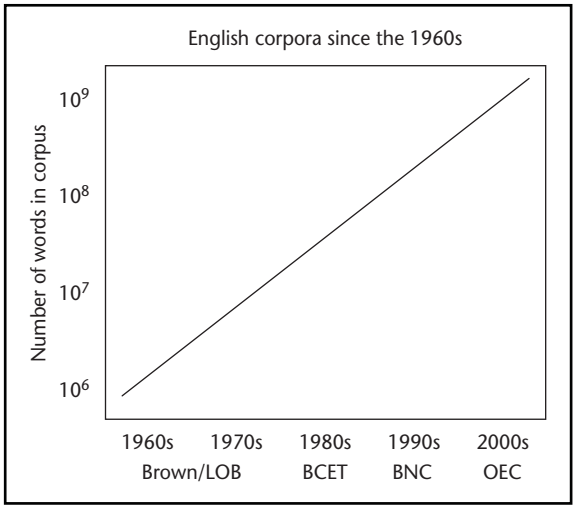


Fig 3.3 Corpus size: growth since the 1960s

When the Brown Corpus – the first electronic corpus of English – was developed in the early 1960s, its goal of collecting one million (10^6) words of text was immensely ambitious and technically demanding (Kučera and Francis 1967). Brown was a collection of written texts in American English, and a British English equivalent – the Lancaster-Oslo-Bergen Corpus – was created a decade later. The Birmingham Collection of English Text (BCET) was compiled as part of the *COBUILD* project in the 1980s (and it later morphed into the Bank of English). It raised the stakes by an order of magnitude, with its initial collection of 7.3 million words rising to 20 million by the middle of the decade (Renouf 1987). The standard for the 1990s was set by the 100-million-word British National Corpus (BNC), and in the 2000s, the Oxford English Corpus (OEC) broke through the one-billion-word (10^9) barrier, and is still growing.

Now that we no longer have to ask 'how large a corpus can we afford to acquire?', we can ask a more interesting question: 'how large a corpus do we need in order to write a good dictionary?' To understand this question better, it is useful to know a little about word frequencies, and specifically about 'Zipf's Law'.

3.4.1.1 *Zipf's Law and its implications* As long ago as the 1930s, the Harvard linguist G. K. Zipf studied manually gathered word-frequency data for English, German, Chinese, and Latin, and observed what he called 'the orderliness of the distribution of words' (Zipf 1935). Zipf found that 'a few words occur with very high frequency while many words occur but rarely' (ibid.: 40). Languages, in other words, consist of a small number of very common words, and a large number of very infrequent ones. Starting from these general observations, Zipf went on to formulate his now-famous 'law', which states that the frequency with which a word appears in a collection of texts is inversely proportional to its ranking in a frequency table.

What exactly does this mean? In essence, Zipf's Law predicts that the tenth most frequent word[14] in a corpus will occur about twice as often as the 20th most frequent word, ten times as often as the 100th most frequent word, and 100 times as often as the 1,000th most frequent word. As Figure 3.4 shows, this is reasonably well borne out in the BNC. Here we take the frequency and rank of *was* (ranked 10th in the corpus) as our baseline, and predict the other figures from that.

word form	ranking in BNC	actual frequency in BNC	frequency predicted by Zipf's Law
was	10th	923,957	–
at	20th	478,177	461,978
made	100th	91,659	92,396
advice	1000th	10,316	9,240
quiet	2000th	5,295	4,619

Fig 3.4 Word frequencies illustrating Zipf's Law

One of the consequences of the 'Zipfian distribution' of words in a language is that a few words occur so often that they account for a very high proportion of any text. The 100 most frequent words in English make up

[14] Or, more precisely, word-form. The figures here are for single forms (like *made*), not whole lemmas (like *make*).

around 45 per cent of the BNC's 100 million words. Look at any sentence in this chapter, and you will usually find that around half the words are high-frequency 'grammatical words' like *and*, *you*, *will*, and *that*.

The converse of this is that most vocabulary items occur only rarely. Consider, for example, a verb like *adjudicate*: not exactly a central item of English vocabulary but by no means a rarity either. In all its forms, *adjudicate* occurs 121 times in the 100-million-word BNC – a little more than one occurrence for every million words. BNC data enables us to make the following statements with a fair amount of confidence:

- *adjudicate* sometimes take a direct object (*their purpose is to adjudicate disputes between employers and employees*);
- more often, it has no object but is followed by a prepositional phrase with *on* or *upon* (*had the sole power to adjudicate on claims of privilege*); almost 40% of cases show this pattern;
- it is occasionally used with an object followed by a complement (*eight years since he was adjudicated bankrupt*);
- it has an unusually strong tendency to be used in the infinitive (51 out of 121 instances), an example of what Hoey (2005) would call a 'colligational' preference;
- the nouns that appear most frequently as its direct object, or following a preposition, are *dispute*, *matter*, and *question*;
- the subject of the verb is typically a specially appointed official or an official body;
- the context is almost invariably a public or official one (rather than a private or domestic one).

This analysis supports Hanks's claim (2002: 157) that 'in a corpus of 100 million words, a simple right- or left-sorted corpus clearly shows most of the normal patterns of usage for all words except the very rare'. For *adjudicate*, at least, with its 121 hits in the corpus, there seems to be enough data to underpin a useful description. But what about words like *temerity* (73 hits in the BNC), *exasperating* (45), *inattentive* (31), or *barnstorming* (20)? Though infrequent, none is so rare as to fall outside the scope of a standard learners' dictionary. If we are looking for data on a range of linguistic features (like the ones for *adjudicate*, above) 20 corpus examples doesn't give us a great deal to go on.

So far we have been talking only about *words*, but similar distributional patterns apply to word meanings and word combinations. A lemma like the

verb *break*, with almost 19,000 hits in the BNC, appears to be well supplied with data. But there is a strong correlation between a word's frequency and its complexity. Thus *break* has at least twenty different meanings and up to a dozen phrasal verbs (some of them polysemous). It also participates in numerous phrases, grammatical patterns, and lexical collocations. And just as some words are frequent and some rare, the same applies to meanings, phrases, patterns, and collocations. Consequently, for some uses of *break*, we may find the evidence surprisingly thin. When we find that the BNC has only *eight* examples of the combination 'break someone's serve/service' (in tennis), our 19,000 hits no longer look so impressive.

3.4.1.2 *Corpus size: conclusions* There is no definitive minimum size for a lexicographic corpus, but the frequency characteristics observed by Zipf indicate that you need very large amounts of text in order to get adequate information for the rarer words and rarer usages. If we are thinking of a corpus that will support the compilation of a dictionary with (say) 80,000 headwords, it's clear that we will need a lot of data to yield enough instances of those items at the lower end of the frequency range. We don't actually know how much data we need in order to account for a given linguistic feature, be it a word, a meaning, or a word combination. What we do know is that the more data we have, the more we learn.[15] And with large volumes of text at our disposal, new kinds of corpus-querying tools come into play: lexical-profiling software, for example (discussed in §4.3.1.5) only works well for lemmas with at least 500 corpus hits (preferably far more).

3.4.2 Content: preliminary questions

We have established that we need to collect very large quantities of text in order to build a lexicographic corpus. But how do we decide what *kinds* of written or spoken material our corpus should include? In this section, we look at the issues that need to be considered when selecting the texts that make up a corpus.

3.4.2.1 *Different texts, different styles* One easy way of collecting text in large quantities is to focus on journalism. For example, the catalogue of the Linguistic Data Consortium – the leading supplier of data for use in

[15] In particular, it is difficult to identify and describe features such as colligation (§8.5.2.3) and semantic prosody (§9.2.8) without really large amounts of data.

language research – includes vast collections of newspaper text in many languages. Its English holdings are taken from sources such as Associated Press newswires and the *New York Times*, and a single DVD can provide us with well over a *billion* words of English.[16] This solves the size problem at a stroke – but how well would such a corpus serve the needs of lexicographers?

A corpus assembled in this way will be fairly homogeneous. Though each individual file may deal with a different topic or event, every constituent text shares certain properties:

- They are all written (as opposed to spoken) texts.
- They all belong to the category 'journalism'.
- They are all examples of American English.
- They all come from a small number of source publications.
- They all originate from a specific, rather short, time-frame (e.g. 2005–2007).

Does this matter? Common sense tells us that American speakers use English in subtly different ways from speakers of British, Australian, or Indian English; that journalism has certain stylistic and rhetorical features not found, say, in academic monographs or face-to-face conversations; that newspapers cover certain subjects (such as politics and business) more fully than others; and that language changes over time and new vocabulary appears. We could predict that when the word *party* appears in newspaper text, it is more likely to refer to a political grouping than a social event (and the position would probably be reversed in a popular romantic novel). We know, in other words (and there is plenty of empirical research to back this up), that different kinds of text have their own distinctive *styles* and deal with their own distinctive *subject matter*.

Experience in using corpora supports these intuitions. Stylistic differences are discussed by Sinclair, who notes that the original Birmingham corpus included a high proportion of fiction. One consequence of this was that certain features of fictional narrative were very prominent. Thus for example 'the broad range of verbs used to introduce speech in novels came out rather too strongly – *wail*, *bark* and *grin* are all attested in this grammatical function' (Sinclair 2005: §4). Meanwhile, differences in subject matter mean that certain words and meanings will be well represented in some texts, and poorly represented (or not present at all) in others.

[16] The LDC's homepage is at http://www.ldc.upenn.edu/ .

largely free of the enclosing **matrix**, and look now much as they wou
poetry from the whole social **matrix** and milieu in which such a subj
h ward. The result was a data **matrix** giving pixel counts for five l
vision by a matrix, when the **matrix** happens to be zero. what does
are summarized in the Payoff **matrix** in Figure A. Now, why the
shing on its own, green is a **matrix** in which to set other colours l
erent crystal structure. The **matrix** is a yellow limestone common in
t eight. The team competency **matrix**. it's upside down. There it
archy to a flat hierarchy, a **matrix** model or a team-based structur
, and are acquainted with the **matrices** of Derrida 's thought in Heg
e a substance is in the lipid **matrix** of olfactory cells, the more i
cultural text, which is the **matrix** or master code that the literar
are well-preserved in a clay **matrix**, some of which adheres to one
ar tissue, connective tissue **matrix**. Which particular technique di

Fig 3.5 Extract from a concordance for *matrix* from the BNC

This is nicely illustrated by the concordance in Figure 3.5. As the data shows, *matrix* is a highly polysemous word, and its various meanings (in geology, anatomy, ceramics, social sciences, and other fields) only emerge here because the corpus from which this concordance is taken includes texts from each of these subject areas. Conversely, a corpus consisting of a single type of text will reflect only the stylistic and subject-matter features of that particular genre. It will, as corpus linguists say, be a 'skewed' corpus, which fails to represent the diversity of style and content in the language as a whole. Since a dictionary has to account for *all* the main meanings and uses of the headwords it includes, it follows that a lexicographic corpus must provide evidence for all these uses.

All of this argues for a corpus whose constituent texts are drawn from a wide range of sources. But this is not a very precise objective. We have established that our corpus – however large – can only be a subset of all the linguistic data in our target language. It is clearly desirable that this subset should reflect the wide variety of ways in which the language is used. In other words, we want to design a sample (a corpus of English, say) that is representative of the broader population (all of the 'communicative events in English'). How far is this possible?

3.4.2.2 *Can a corpus be representative?* Making inferences from samples is a common procedure in many social and applied sciences. There is a well-established body of theory underpinning the collection of samples from which researchers can make generalizations about the population as a whole. The standard way of avoiding bias is to collect a 'random sample',

one in which every possible member of the broader population has an equal chance of being selected. The theory is that any observations we make about the sample will support inferences about the whole population. But we immediately run into problems when we apply this approach to the study of language. In most fields, samples are selected from a population which (even if very large) is both well-defined and of limited extent: for example, your population might be 'all registered voters in the state of California', or 'all items auctioned on e-Bay on 1st October 2007'. Natural languages don't fit well into this model because it is difficult to define what the total population is, and because the population is continually growing. Even if we could satisfactorily define 'English', it is such a vast and diverse entity that 'it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample' (Atkins, Clear, and Ostler 1992: 7).

One partial solution is to apply *stratified sampling*. This involves breaking up the total population into a number of subcategories or types, then creating independent random samples from each of these groupings. This has been a popular strategy among corpus-builders because a stratified approach tends to lead to more representative samples (cf. Biber 1993: 244). But this immediately raises two new questions:

- How do we define these subcategories?
- How do we decide what *proportions* of each subcategory the corpus should include?

Dividing 'language' into discrete, relatively homogeneous groupings is by no means straightforward, because there is no universally agreed classification of text-types. And even if we have a robust set of text-types to sample from, we still have to decide *how much text* we want from each type. The usual approach with stratified sampling is to allocate a percentage to each stratum that reflects its proportion in the total population. But how is this possible when the strata are text-types in a language? Suppose our corpus is to include novels, academic writing, conversations, and newspaper text: is there any objective way of deciding what proportions of each type are appropriate? A few questions will illustrate the complexity of this issue:

- Almost every member of a speech community takes part in face-to-face conversations many times every day; it follows that the majority of communicative events in any language are spoken rather than written. So should a representative corpus consist predominantly of spoken

text (which just happens to be one of the most difficult forms to capture)?

- On the other hand, most spoken encounters are ephemeral, so does the greater 'longevity' of many written texts imply that they are in some way more valuable?
- The number of people involved in the average conversation is small (compared, say, with the number of people watching a popular TV show). So should we take account of 'audience size', and decide that language events involving large numbers of people are more important than small-scale ones?
- Following this argument, should our corpus give more weight to *The News of the World* (a popular tabloid, and the UK's best-selling newspaper) than to *The* (London) *Times* (which is read by far fewer people)?
- Pursuing this last point, should we (or can we) make any allowance for the 'influentialness' of a text? A work of popular romantic fiction may be read by millions, but a serious novel by an admired author may be felt to be a more influential set of language data, and may (unlike more 'popular' titles) continue to be read over many years and studied in schools and universities.
- And if we decide to focus only on published written texts, how do we decide what the total population consists of? Suppose, for example, that there were 2,000 daily newspaper titles published in the US in 2005, and 10,000 books. How do we sample from this population? Do we count each separate edition of a newspaper as a different title? If so, the population of newspapers greatly exceeds the number of books, so our corpus will be dominated by newspapers – yet common sense suggests that 'daily newspapers' represent a single text-type, whereas the category 'book' encompasses many.

It turns out that even defining what a 'language event' constitutes can be extremely difficult. If I pass a sign on my way into work saying 'All visitors must go first to Reception', does this count as a language event (and is it a different event each time I pass it)? Or if I overhear a conversation on the train, is that a language event?

Questions like these have exercised corpus developers for many years, and since the mid-1990s the situation has become even more challenging, as a whole range of new text-types has arisen. Chat rooms, blogs, emails, SMS

messaging, and 'social networking' websites like Second Life or FaceBook generate new and important forms of language data, but they don't always conform neatly to text categories established in the pre-web era.

In the next section we outline an approach to text classification which is well tried and intuitively satisfying, but we should be under no illusion that this will magically deliver a truly representative corpus. There is no obvious way of creating a 'representative' corpus of a widely used living language because:

- it is almost impossible to define the population that the corpus should be representative *of*, and
- since the population is unlimited, it is logically impossible to establish 'correct' proportions of each component.

3.4.2.3 *An achievable objective: a 'balanced' corpus* Even if 'representativeness' is unattainable, it remains a good aspiration. We know that words behave differently in different contexts of use, so a corpus drawn from a single source (e.g. 500 million words from *The Wall Street Journal*) won't provide all the data we need to support a general-purpose description of English. Somewhere between the two extremes of a perfectly representative text collection and a 'monolithic' one lies a more modest goal. We might describe this goal as a 'balanced' corpus. A balanced corpus is one that conscientiously seeks to reflect the diversity of the target language, by including texts which collectively cover the full repertoire of ways in which people use the language. We have to accept that creating a balanced corpus can never be a scientific process: designing our ideal sample involves too many subjective decisions, and even then the eventual selection of texts will be constrained by practical and financial factors (more on this in §3.5). Nevertheless, with a good set of criteria we can establish a useful typology of text-types. And if we then apply a stratified sampling approach to identify specific texts within each main category, we should end up with a corpus that systematically reflects the range of available text-types. Finally (and we will discuss this further in §3.6.2) if every text is carefully described in terms of its key features (genre, authorship, date of publication and so on), corpus-users will have the information they need to assess the significance of any given instance of a word, phrase, or meaning.

3.4.2.4 *Selecting texts: internal and external criteria* Before we discuss the parameters to be used in selecting texts for a corpus, it is worth making a

distinction between the internal and external properties of texts. Internal properties refer to linguistic or stylistic features that some texts share with others. External properties reflect the situational or functional attributes of a text, and refer to categories such as 'newspaper', 'novel', 'instruction manual', and 'conversation'.

A good deal of work has been done[17] to investigate and identify the (internal) linguistic properties of different kinds of text. We can observe, for example, that a number of features – such as verbs in the past tense or passive, first and third person pronouns, and prepositional phrases – appear with varying frequency distributions in different texts. To give a simple example, Biber (1993: 251) shows that noun + preposition sequences are significantly more common in technical writing than in fiction. The argument is that text-types can be identified (and then collected for a corpus) according to the particular ways in which clusters of these features are distributed. The corpus-collection model here is a recursive one:

- First you gather some texts from a range of sources.
- Next you analyse them to identify recurring clusters of linguistic features.
- This enables you to establish provisional categories of texts, grouped on the basis of shared linguistic features.
- Then you collect more texts to reflect these feature-distributions.
- Then you repeat the analysis on your enlarged corpus, refine your typology, and collect more texts.
- And so on.

The process thus 'proceeds in a cyclical fashion' (ibid.: 256) until you have collected a large corpus whose contents reflect the proportions in which the various key features are observable in large bodies of text.

This is a labour-intensive way of developing a corpus. But aside from the practical issues, there is an important theoretical objection. As Sinclair has pointed out, if texts are collected in this way, 'the results of corpus study would be compromised by circularity in the argument' (Sinclair 2003: 171). If we collect 'humanities' texts, for example, on the basis of an observation about the way language is used in such texts, and then 'discover' these same features when we analyse the resulting corpus, is this a genuine feature of

[17] Notably by Douglas Biber: see esp. Biber 1990, and 1993: 248–255.

such texts, or does the finding merely reflect the criteria we used to select these texts in the first place?

The more usual corpus design model, therefore, is one based on *external* criteria. There are several well-established typologies of texts to guide us: for example, the Lancaster-Oslo-Bergen (LOB) corpus used the categorizations in the *British National Bibliography*, whose subject index is based on the Dewey Decimal library system. These categories enable us to create a 'sampling frame', and we can then apply a stratified sampling method by randomly selecting texts from each part of the frame.

3.4.2.5 *Spoken data: a special case* With spoken data, the goal of achieving balance presents special challenges. For written texts, library classifications and similar typologies are a good place to start when creating a sampling frame. But 'with a corpus of spoken language there are no obvious objective measures that can be used to define the target population' (Crowdy 1993: 259). The way the BNC addresses this issue – and it is as good a model as any – is to use a 'demographic' approach in order to collect samples of ordinary, face-to-face conversation, and to supplement this with a set of 'context-governed' spoken texts. Demographic sampling – a well-known technique in social-science research – entails defining the population (in this case, speakers of English living in the UK) in terms of features such as gender, social class, age, and region, and then creating a sample that reflects all these variables. (The data collection methods used are discussed later: §3.5.2.) It was felt, however, that a spoken corpus consisting *only* of conversation would not adequately reflect the diversity of the spoken language. There are several types of language event which – though far less abundant than conversation – are nevertheless significant forms of spoken discourse. These belong to the 'context-governed' component of the corpus, and include:

- educational and informative events, such as lectures, seminars, and news broadcasts
- business events such as consultations, interviews, and meetings
- public events such as political speeches and parliamentary proceedings
- leisure events, such as meetings of clubs, chat shows, and phone-in shows.

With this twin-track approach, the BNC makes a creditable job of representing 'the full range of linguistic variation found in spoken language' (Crowdy 1993: 259).

3.4.2.6 *A note on 'skewing'* Skewing refers to a form of bias in data whereby a particular feature is either over- or under-represented to a degree that distorts the general picture. The BNC – though generally well-balanced – includes one egregious case of skewing. It contains several large samples (almost 750,000 words in all) from *Gut: the Journal of Gastroenterology and Hepatology* (a highly specialized source). Most of the time this doesn't cause problems, but frequency counts from the corpus throw up some anomalies: the word *mucosa*, for instance, has the same number of hits in the BNC (1,031) as the word *unfortunate*, and in this case the statistics are clearly misleading. But as corpora grow larger, problems with skewing gradually recede. In a small corpus, a single 'rogue' text may distort the overall picture, but in a large corpus the risks are reduced. Take for example the novel *Saturday* by award-winning UK author Ian McEwan (2006). This looks a perfect candidate for inclusion in the 'literary fiction' component of a lexicographic corpus – except for one detail. The book's central character is a neurosurgeon, and parts of it include highly technical vocabulary describing areas of the brain and surgical procedures (such as *transsphenoidal hypophysectomy*). While a small corpus might give undue prominence to these eccentricities, they have little impact on a large one. Large corpora are more 'forgiving' and less likely to be affected by skewing. We still need to be careful about the categories of text from which we source our corpus documents (and this is addressed in the next section), but the requirement for careful selection of individual texts diminishes as corpus size increases.

3.4.3 Content: an inventory of text-types

The text selection criteria we describe here consist of a number of *attributes* which a text has, and for each attribute, there are two or more possible *values*. The attributes we discuss are:

- language
- time
- mode
- medium
- domain

These attribute/value combinations enable us to classify any text and situate it in a particular part of our sampling frame. The parameters described

here are neither the only ones you can use nor necessarily the best ones you can use, but they do provide a workable model for building a lexicographic corpus.[18]

3.4.3.1 *Language* Will your corpus be:

- monolingual?
- bilingual?
- multilingual?

The term 'parallel corpus' denotes a set of corpora (two in a bilingual parallel corpus, more in a multilingual version) in which the texts in Language A correspond in some way to those in Language B (and perhaps C and D and so on). There are two types of parallel corpus of value to lexicographers: the 'translation corpus' and the 'comparable corpus'. Their use is discussed in §11.3.2.1. A translation corpus consists of a set of texts in one language with translated versions of the same texts in another language or other languages. Perhaps the best-known bilingual translation corpus is the English and French Canadian Hansard corpus, while a typical example of a multilingual translation corpus might be a set of European Union documents translated into every official language in the community. In a comparable corpus, texts from two or more languages (or language varieties) are collected using an identical sampling frame. A good example is the International Corpus of English (ICE), which consists of fifteen corpora, each of one million words, for varieties of English from places such as New Zealand, India, the Philippines, and Jamaica.[19]

For present purposes, let's assume your corpus consists of texts from a single language. Even so, decisions have to be made:

- Does the corpus represent one, several, or all varieties of the target language? Compare, for example, the BNC (explicitly a corpus of British English) and the Bank of English (which seeks to cover several of the major varieties, including American, Indian, and Australian English).
- How far does the corpus account for dialectal variation? For example, the spoken component of the BNC used volunteers from 38 different locations across the UK, with the aim of capturing regional variation.

[18] Variations on the typology outlined here can be found in Renouf (1987); Atkins, Clear, and Ostler (1992); and Sinclair (2003).

[19] The ICE's homepage is at http://www.ucl.ac.uk/english-usage/ice/ .

- Do we restrict ourselves to texts produced by native speakers of the target language? (The BNC does this, but the Irish component of the New Corpus for Ireland does not.) If so, we will need a good operational definition of native speaker. (This is far from straightforward: consider Joseph Conrad, whose novels belong to the English literary canon – yet English was Conrad's third language.) And when we collect texts in English from the web, it will often be impossible to determine the mother tongue of the writer.

3.4.3.2 *Time* Will your corpus be:

- synchronic?
- diachronic?

In a synchronic corpus, the constituent texts come from one specific period of time, whereas the texts making up a diachronic corpus come from an extended period. The best examples of synchronic corpora are the Brown and LOB collections, which consist exclusively of texts published in 1961. At the opposite extreme, the Oxford Historical Corpus spans twelve centuries (from *Beowulf* to the early twentieth century). Between these poles, there is a continuum of 'diachronicity': the BNC, for example, includes texts dating from 1975 to 1992, while the Irish-language component of the New Corpus for Ireland covers the period 1883–2003.[20] Essentially, corpus-builders have to decide 'how diachronic' their corpus needs to be in order to support the kind of lexicography they will be doing. While a historical dictionary like the *OED* clearly requires a fully diachronic corpus, dictionaries designed for learners deal mainly with contemporary language, so they need a (broadly) synchronic corpus which provides a snapshot of the language as it is used at the time of compilation.

3.4.3.3 *Mode* Will your corpus include:

- written texts?
- spoken texts?
- both?

The Brown and LOB corpora consist only of written material, whereas the BNC includes a substantial component of spoken data. As with most of

[20] For details see Kilgarriff, Rundell, and Uí Dhonnchadha (2007), and www.focloir.ie/corpus/ .

these binary categories, the boundaries are not always completely clear: fictional dialogue, for example, is a written form of text, yet it aspires to replicate the way people talk to one another. Newer 'hybrid' forms of text associated with the web complicate the matter further. The text found in chat rooms is, strictly speaking, 'written' (it consists of keystrokes rather than sound waves). On the other hand, it is typically produced (as conversations are) in real time, and displays many of the characteristics of spontaneous spoken dialogue. As we shall see later (§3.5.1), spoken text is – for a number of reasons – more difficult to collect than written, so practical issues will often influence your corpus design model.

3.4.3.4 *Medium* Medium refers to the 'channel' in which the text appears. A simple classification here would distinguish print media and spoken media. The former include (*inter alia*) books, newspapers, magazines, learned journals, dissertations, movie scripts, government documents, and legal statutes. Spoken media include face-to-face conversations, broadcasts and podcasts, public meetings, and educational settings (seminars, lectures, etc.). Once again, traditional categories become blurred when we add the web to the mix. Some 'new' text-types (blogs and social networking sites, for example) are exclusive to the web, but many documents exist in both print and electronic media. Most newspapers are published in both channels – but online versions often include additional material that doesn't appear in the printed edition. Similarly, serious, refereed journals and conference proceedings are increasingly published online (either exclusively so, or in conjunction with print versions).

3.4.3.5 *Domain* Domain refers to the subject matter of a text: what the text is *about*.[21] It will be immediately clear that – unlike Language, Time, Mode, and Medium – Domain is not a 'universal' parameter because not all kinds of text can be classified in these terms. For example:

- Although some spoken encounters (such as academic seminars or public meetings) may focus on a particular topic, most ordinary conversations do not.
- Some works of fiction or drama may be set in a particular period or may deal with a particular subject, but they could rarely be said to

[21] Sometimes also referred to as 'topic', e.g. by Sinclair (2003: 172).

belong to a single domain in the way that most academic monographs could.[22]

- Newspapers, despite their primary focus on current affairs, are far from homogeneous in their subject matter.

Taking account of issues like these, the BNC applies the 'domain' criterion only to written material, and within that broad category only to 'informative' (as opposed to 'imaginative') texts. Its design allocates these texts to one of eight major domains,[23] and each domain is subdivided into more specialized categories. Assigning texts to a specific domain and subdomain is reasonably straightforward, but deciding which subdomains to sample from raises interesting questions.

3.4.3.6 *Dealing with sublanguages* When we think about the vocabulary of a language, it is useful to make a broad distinction between 'core' usages and 'sublanguages'. The word *deuce* is part of a sublanguage: it belongs to the vocabulary of tennis. A word like *important*, on the other hand, belongs to the core vocabulary of English. We know this intuitively, and empirical data supports our observation: *deuce* (leaving aside its use in old-fashioned interjections like 'what the deuce . . . ?') appears in only 9 of the BNC's 4,124 texts, whereas *important* is found in 3,810 of them. (Less obviously, words like *serve*, *set*, *game*, and even *love* are core in some meanings, but they too – in specific senses – belong to the sublanguage of tennis.) Though sublanguage terms may crop up anywhere (the characters in a novel might have a game of tennis, for example), the only way of ensuring they are systematically represented is to include specialized texts. This raises interesting questions (which are discussed in more detail in Kilgarriff and Grefenstette 2003: 10). Do we:

- include no sublanguages? (This would give us an impoverished view of language.)
- include all sublanguages? (How do we know what they are?)
- include some sublanguages? (An unsatisfactory compromise.)

[22] Think for example of Pat Barker's *Regeneration* trilogy published in the 1990s: the novels are set in World War I but deal with subjects such as psychiatry, love, and poetry, as well as warfare. Similarly, Michael Frayn's play *Copenhagen* (1998), though focusing on an imagined meeting between Werner Heisenberg and Niels Bohr, can hardly be said to be a text 'about' theoretical physics.

[23] Namely: applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs.

Despite its arbitrariness, corpus-builders have tended to take the latter route. Thus the BNC's texts cover subjects as diverse as making cakes and soup; gliding; taking care of dogs; hotel management; photography; village life in Nepal; and bereavement counselling (among many others). On the other hand, it has no texts about badminton or volleyball, car maintenance, heavy-metal music, carpentry, or (surprisingly) Islam. This is perfectly understandable – there are serious practical issues in attempting to sample 'all' sublanguages – but it is clearly not ideal. In fact, however, improvements in technology will enable us to overcome most of the obstacles to broad coverage of sublanguages, because:

- Data storage is no longer a major issue.
- Figuring out what the relevant sublanguages *are* is becoming less difficult.
- Collecting appropriate texts has become far easier because of the web.

The web now plays host to vast numbers of 'cyber-communities': namely groups of people with a common area of interest who generate and share content on the internet. For example, Yahoo! has a 'Recreation' forum, which includes tens of thousands of online 'clubs' devoted to every conceivable sport, game, hobby, and recreational activity. Research into communities of this type may provide the basis for a comprehensive inventory of subject fields (and hence of their characteristic sublanguages).[24] Meanwhile, software for the rapid creation of corpora for specialized subjects is already in place.[25] A corpus that systematically samples 'all' sublanguages may still be some way off, but there is no longer any compelling reason not to move in this direction.

3.4.4 Corpus design: some conclusions

We have shown that a lexicographic corpus should be as large and diverse as possible, and that the technical constraints which once made these objectives so challenging have to a large extent disappeared. A truly representative corpus is an impossible goal because we are sampling from a population

[24] See for example Kumar et al. 'Trawling the Web for Emerging Cyber-Communities' (1999), available online at http://citeseer.ist.psu.edu/kumar99trawling.html .

[25] A good example is WebBootCaT, one of the components of the Sketch Engine: see Baroni et al. (2006).

whose nature is unknowable and whose extent is unlimited. Nevertheless, we know that the description of language in a dictionary cannot be complete if the dictionary's source data doesn't reflect the full repertoire of language events. Our goal, then, is a 'balanced' corpus, though we recognize that there is no single, scientific methodology for achieving this. Texts can be categorized in a variety of ways, but even the very broad categories discussed here have fuzzy boundaries and are not always mutually exclusive.

Beyond these major categories, texts have numerous other attributes which we may want to record. In Atkins, Clear, and Ostler (1992: 11ff.), for example, the following features are proposed (and most of these found their way into the BNC's text-description schema):

- Authorship: was the text produced by one person, or is its authorship joint, multiple, or corporate? Was the writer (or speaker) male or female?
- Preparedness (an attribute especially of spoken data): is the text spontaneous, based on notes, or fully edited?
- Function: not all texts have a specific function (in which case it is 'unmarked') but many can be characterized as narrative, informative, expository, or hortatory/persuasive.
- Audience: is the text aimed at children, teenagers, adults, or any other specific group?
- Technicality: some texts are produced by specialists for specialists, some by specialists for laypeople, and some by non-specialists. This can be a useful way of distinguishing texts in the same subject-field: within the area of 'computing', for example, we may have an advanced manual for programmers on the one hand, and on the other a popular article giving simple guidelines for beginners.

Variations on these attributes can be found in the design specifications of other corpora.[26] But even when you have established a good, workable sampling frame, you still have to decide how much text to acquire for each part of the frame. The BNC, for example, uses the following (approximate) proportions:

- mode: 90% written, 10% spoken
- written texts: 75% informative, 25% imaginative
- spoken texts: 42% demographic, 58% context-governed

[26] See for example Kučera and Francis (1967), Renouf (1987), Summers (1993).

- written medium: 60% books, 30% periodicals of various kinds, 5% ephemera (brochures, advertising material, etc.), 5% unpublished material.

Other corpora balance their constituent texts and text-types in different ways. Each approach is generally well thought through and well argued – but also adjusted to cope with the practical challenges that rear their heads as the corpus-gathering process gets under way. None of the various approaches is obviously 'better' than the other, and none can claim to be scientific. But a corpus that combines high volumes of text with a design which conscientiously reflects the diversity of the language will provide excellent raw materials for mainstream lexicography.

3.5 Collecting corpus data

In this section we discuss the issues involved in acquiring texts for your corpus. We address the following topics:

- collecting written data
- collecting spoken data
- collecting data from the web
- sample size
- copyright and permissions.

3.5.1 Collecting written data

It goes without saying that the texts in an electronic corpus have to be in digital form. For synchronic corpora this is rarely a problem, now that most written material starts life in some kind of digital format. Life wasn't always so easy. Earlier corpora made extensive use of scanning and keyboarding (both slow, labour-intensive processes) to convert printed pages into usable data. Renouf's account of the development of the original Birmingham/COBUILD corpus gives some idea of the heroic efforts involved (Renouf 1987: 5–6). Some texts were captured by keyboarding, but many were scanned using a 'Kurzweil Data Entry Machine' (KDEM) – a state-of-the-art (and phenomenally expensive) piece of technology in 1980. The KDEM first had to be trained, a laborious procedure in itself. After about nine months of intensive work on both fronts, 'we had KDEMed and keyboarded sufficient material to allow us some choice in putting together

six million words for concordancing'. By the time the BNC was getting under way about ten years later, it was expected that reliance on scanning and keyboarding would be greatly reduced, because 'the corpus designers believed that many texts would already exist in electronic form' (BNC website: http://www.natcorp.ox.ac.uk/corpus/creating.xml.ID=electronic). This turned out to be an optimistic assessment: 'texts in electronic form which fitted the corpus design were far fewer than had been supposed', so scanning and keyboarding still had a big part to play. For today's corpus-builders, these problems have largely disappeared, though converting electronic texts to a standard format may still involve significant effort (as we discuss later: §3.6.1).

3.5.2 Collecting spoken data

If a corpus aims to provide a snapshot of contemporary language, it is clearly desirable that it should include significant quantities of spontaneous, unscripted speech. But spoken data has, traditionally, been difficult and expensive to collect. Consequently, although the majority of communicative events in a language occur in spoken mode, few corpora include high proportions of spoken material. Only 10 per cent of the BNC is spoken, and less than half of its 10-million-word speech component consists of ordinary face-to-face conversations. These were captured on tape by a large group of volunteers, recruited by the British Market Research Bureau to form a representative cohort of the British population – with 'equal numbers of men and women, approximately equal numbers from each age group, and equal numbers from each social grouping' (BNC website). Volunteers kept detailed logs of every conversation in which they were involved, and their recorded conversations were painstakingly transcribed. This is valuable data, but the costs of collecting it were high. It is a fair bet that nothing on this scale will be attempted again until speech-recognition technology can cope with ambient noise, overlaps between speakers, and the many other challenges that multi-participant conversations present.

But if this kind of data represents a 'gold standard', there are plenty of sources on the web for spoken material which still has considerable value. For example, transcripts of broadcast interviews (from the Larry King Show, Voice of America, and so on) are available in enormous volumes. In most cases, the material is not 'sanitized' to look like written text, so it retains the feel of spontaneous speech, as this short extract illustrates:

Well, it's, you know, really the whole week, it's just been such an emotional week for the town and the communities around. Obviously, the town of Cheshire, just down the road Route 10 a little bit from us, very emotional.

Sites like 'Everyzing' (www.everyzing.com) provide access to huge libraries of broadcast material, advertising, and home-made video clips, in domains as diverse as politics, entertainment, business, sport, and health. Everyzing comes with a powerful search engine that enables users to retrieve countless examples of any given word or expression, and for each clip on view, there is a full transcript of the text. We discuss later (§3.7.1) the trade-off that corpus-builders often have to make between quality and quantity, and sites like these illustrate the point well. Though the BNC's spoken material is of very high quality, it is a small subcorpus by present-day standards, and its value inevitably degrades as time passes, so that what was once current language starts to look dated. Against this background, web-derived spoken data – which offers up-to-date material in large quantities and at low cost – begins to look like an attractive alternative.

3.5.3 Collecting data from the web

The web itself, of course, is a vast and ever-growing repository of texts of every conceivable type. It is probably pointless to speculate how much linguistic data it holds, but for the world's major languages (and even some less major ones) the volumes of available text are far in excess of lexicographers' needs. The question of 'whether the web is a corpus' is a hotly debated topic in language-engineering circles. For lexicography, it is better to see the web as a *source* of texts from which a lexicographic corpus can be assembled. Collecting data in this way is not without its problems, but a good deal of work has already been done, especially within the NLP community. The biggest challenge has been to develop reliable methodologies for automatically separating large tranches of continuous, cohesive text from all the other data-types in this medium (such as navigational aids, lists, images, sound files, and other varieties of 'noise'). The technology is now fairly mature and 'web corpora' have been assembled for several languages.[27]

[27] For example, researchers in Leeds have created large and balanced web corpora for Chinese, English, French, German, Italian, Polish, Portuguese, Russian and Spanish: http://corpus.leeds.ac.uk/internet.html . See also Baroni et al. (2006) on WebBootCaT,

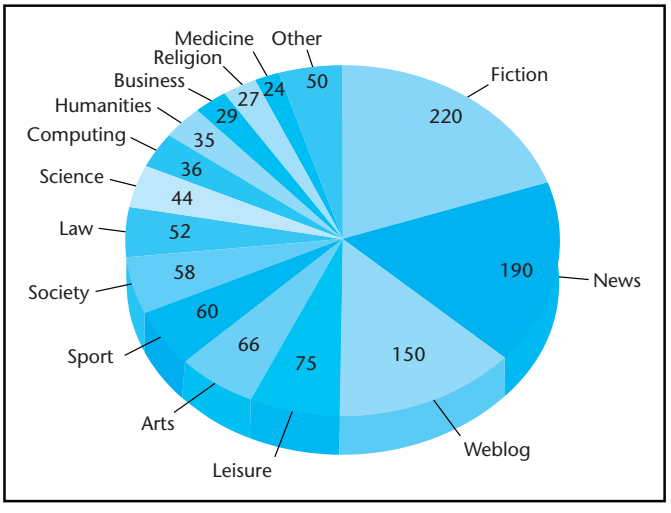


Fig 3.6 Composition of the *Oxford English Corpus* (OEC)

The New Corpus for Ireland – collected in 2003–4 – is an early example of a corpus which includes web data. Though most of its texts come from 'conventional' sources, about 20 per cent of the Irish and Hiberno-English components consist of material from the web (Kilgarriff, Rundell, & Uí Dhonnchadha 2007: 134f.).

The Oxford English Corpus (OEC) – launched in 2006, and still growing – is the first lexicographic corpus of English sourced entirely from the web. The figures in each 'slice' of the piechart in Figure 3.6 indicate the number of words, in millions, in each major component of the corpus. The OEC has clearly made serious efforts to cover a wide range of text-types and domains, and its one billion words come from a huge variety of sources. For example, the OEC's 60 million words in the general domain of 'Sport' are made up of subcorpora that deal with about forty individual sports. Large and impressive though it is, the OEC has already been overtaken – in terms of sheer volume, at least – by web corpora such as the ukWaC corpus, which is bundled with the Sketch Engine package and weighs in at just over 2 billion words.

It goes without saying that there are some downsides here. It is often harder to be sure about the exact provenance of a text on the web. Which part of the English-speaking world does the author come from? Is the

a software tool for the rapid creation and processing of web corpora in specialized domains.

author a native speaker? Do we even know who the author is? And has the text gone through the kind of editorial process we associate with conventionally published material? Questions like these inevitably raise concerns about the 'integrity' of web data. On the other hand it is undeniably true that the web allows us to gather large quantities of text far more efficiently (and cheaply) than was ever possible using conventional methods. Concerns about the range of text-types available on the web (could the web really supply the variety of registers found in a corpus like the BNC?) have proved largely unfounded. Research done in comparing web-derived corpora with 'benchmark' collections like the BNC is (thus far) encouraging, and suggests that a carefully designed web corpus can provide reliable language data. Fletcher (2004), for example, compared word-frequency data in a 5.4-million-word web corpus with counts from the BNC. He notes a number of differences in frequency ranking which reflect 'the biases and gaps' in each data-set. Obvious examples are the BNC's bias towards British forms, institutions, and place-names, while the web data has a clear orientation towards the US. Also, 'in the BNC texts the language of news and politics stands out, while in the web corpus academic concepts are quite salient'. In general, though, the differences are not unduly great, and Fletcher believes that research of this kind will 'help dispel doubts about the representativeness of selected web documents for English as a whole'.[28]

3.5.4 Sample size

The Brown Corpus was made up of short, 2,000-word extracts from each of its constituent texts. Given its goal of collecting one million words in all, this was a sensible way of ensuring that a wide range of text-types could be represented, *and* that no single text was large enough to upset the overall balance (cf. §3.4.2.6 above on 'skewing'). The BNC had a lot more words to play with, but even so there were concerns that including very large documents might affect the reliability of corpus-derived frequency counts and give undue weight to the idiosyncrasies of a single writer or source.

[28] Similarly, investigations in the frequencies of various types of bigram in web data led Keller, Lapata, and Ourioupina (2002: 236) to conclude that 'the counts obtained from the web are highly correlated with the counts obtained from the BNC, which indicates that web queries can generate frequencies that are comparable to the ones obtained from a balanced, carefully edited corpus such as the BNC'.

So for that part of the corpus made up of books, the BNC set a limit of 40,000 words per individual document. If a source text was larger than this, a sample was taken.

This approach is not without risks, and there are good arguments for using complete texts rather than extracts. In many registers (notably academic writing), the discourse structure and rhetorical features of a text may vary as it proceeds from its opening paragraphs, through its central sections, to the concluding chapters.[29] The BNC's 'solution' to this was to ensure that 40,000-word samples were taken variously from the beginning, middle, and end of its source documents, and a similar approach was used in the New Corpus for Ireland. But as we enter the era of mega-corpora, the 'skewing' argument begins to lose its force. A very large document such as Vikram Seth's novel *A Suitable Boy* (1994) – a massive book of over half a million words – would have fatally skewed the Brown Corpus if included in its entirety, and seriously unbalanced the original (7.3m word) Birmingham corpus. But the larger the corpus, the less of an issue this becomes – Seth's text would have little impact on a billion-word corpus. There remains, however, a pragmatic reason why samples may be preferred to complete texts: copyright owners are more likely to allow an extract of their text to be used in a corpus, rather than the full version. This is the subject we address next.

3.5.5 Copyright and permissions

As Atkins et al. noted, 'one of the serious constraints on the development of large text corpora and their widespread use is national and international copyright legislation' (Atkins, Clear, and Ostler 1992: 6). Unless a corpus is made up of much older texts, most of its source material is likely to be protected by copyright. The basic principle is that a text is usually the intellectual property either of its creator or of the person or organization that paid for it to be created. So corpus-builders need to ensure they have permission from the copyright owner to include it in their corpus. The 'level' of permissions needed will depend to some extent on the breadth of its availability: who is entitled to use the corpus, and for what purposes? The Bank of English, for example, is not widely available: though a small subset

[29] Sinclair refers to 'the marked differences that have been observed between different parts of a text', and notes that 'not many features of a book-length text are diffused evenly throughout' (1991: 19).

is publicly available through subscription, access to the complete corpus is generally restricted to employees of HarperCollins (its joint creator) and bona fide researchers at Birmingham University. The BNC, on the other hand, has a much higher level of permissions. Access is not completely unrestricted (as in the case of WordNet, for example), but the licence agreements entered into by donors of text have proved flexible enough to allow the corpus to be used very widely, for commercial purposes as well as academic. A great deal of work is involved in securing permissions from copyright-holders and, not surprisingly, this turned out to be one of the most time-consuming aspects of the whole BNC project. As Sinclair observes (1991: 15), 'the labour of keeping a large corpus in good legal health is enormous'.

➔ In order to ensure that your corpus is 'in good legal health', you will first need to find out who owns the copyright of each text that you plan to include (this isn't always as straightforward as it sounds). When you approach the copyright owner for permission, it's important to be transparent about the intended use of their text. Publishing rights managers will generally be wary about requests for permission to disseminate their intellectual property – which is hardly surprising: it's their job to be protective. Furthermore, they won't necessarily know much about corpus linguistics, so it's a good idea to accompany your request with a short, simple explanation of what a corpus is, how and why people use it, and how their own data will eventually form part of a much larger body of texts. The key thing is to reassure rights-holders that their data will be safe in your hands, and it does no harm to imply that selection for inclusion in a major corpus says something positive about the value and significance of their text. During the creation of the New Corpus for Ireland, for example, copyright owners – after an initial approach – were sent three documents:

- A permissions request describing the precise terms on which a text would be included. As well as specifying the computational processes involved, the letter noted that 'All users of the New Corpus for Ireland will sign an End User Licence, limiting the uses that may be made of the data'. For this corpus (and the same point applies to the BNC) it was made clear that samples would be used (of up to 60,000 words, in this case) and that no text would be included in its entirety.
- A copy of the End User Licence referred to, which included a clause explicitly prohibiting licensees 'from publishing in print or electronic

form or exploiting commercially in any form whatsoever any extracts from the Corpus other than those permitted under the fair dealing provisions of copyright law'.

- A short PDF document giving a simple explanation of corpora and how they are used, illustrated by screenshots from a corpus query system, showing concordances, lexical profiles, wordlists, and so on.[30]

One final (and very important) recommendation: never offer to pay for permission to include a text, and never agree to such a request from a copyright owner. Once money starts changing hands (even if for a single text in a single corpus), a precedent would be established that could have fatal consequences to corpus-creation efforts worldwide.

However well this part of the project is planned, obtaining permissions for the hundreds or thousands of texts in a corpus will always be a time-consuming administrative job. It was certainly one of the biggest overheads on the BNC project, and it is hard to imagine anyone attempting to do this again on the even larger scales that are now seen as normal for a lexicographic corpus. This can only increase the appeal of sourcing texts from the web. Most legislation governing intellectual property rights (IPR) was framed before the internet came into existence, so we are now in a transitional phase where the law has not yet caught up with the technology. This is a classic grey area, and it would be imprudent to attempt a definition of what is and is not allowable. Kilgarriff and Grefenstette (2003: 335, footnote 2) make the following point:

> Lawyers may argue that the legal issues for web corpora are no different to those around non-web corpora. However . . . a web corpus is a very minor sub-species of the caches and indexes held by search engines and assorted other components of the infrastructure of the web: if a web corpus is infringing copyright, then it is merely doing on a small scale what search engines such as Google are doing on a colossal scale.

At any rate, several reputable language institutions and publishing companies have already gone down this route after taking legal advice, and there are good reasons for believing that web corpora – provided access

[30] The excellent BNC website also provides useful information on this subject: go to http://www.natcorp.ox.ac.uk/corpus/ and follow the link to the 'Permissions Clearance' page.

is controlled by the same kinds of End User Licence used for 'traditional' corpora – are for the time being on the right side of the law.

3.6 Processing and annotating the data

In this section, we discuss the processes involved in taking corpus texts from their raw state to a final form in which they can be used efficiently by lexicographers and linguists. This operation entails three fairly distinct stages:

- clean-up, standardization, and text encoding: essentially, the process of taking a heterogeneous collection of input documents and converting them all to a standard, usable form
- documentation: providing each input text with a unique 'header document' which records its essential features
- linguistic annotation: enriching raw text by adding grammatical information which will enable corpus users to frame sophisticated queries and extract maximum benefit from the data.

3.6.1 Clean-up, standardization, and text encoding

A large and diverse corpus will include many thousands of individual texts from a wide range of sources and in a wide range of media. Input texts may have been keyboarded, transcribed from recordings, scanned, dumped from typesetters' tapes, or downloaded from websites. The first thing we need to do is standardize this disparate collection of data, so that we end up with a single body of text in a uniform format. This makes the resulting corpus maximally portable, and ensures that the data can be used in a corpus-query system (on which, see §4.3.1). A number of formats have been used in the past (the BNC, for example, was originally encoded using an SGML standard), but there is now a generally accepted standard in the form of XCES, the XML Corpus Encoding Standard, details of which can be found at http://www.xml-ces.org/ .

3.6.1.1 *Written data* Corpus-builders are in a much more fortunate position than the pioneers of the late twentieth century (§3.5.1), because most texts from written sources – books, newspapers, journals, and so on – will

already be in digital form. But although scanning and keyboarding won't generally be needed, we are still some way from an ideal world in which most input texts come in ready-to-use form. Different organizations deliver their texts in different software packages (some proprietary, some home-grown), typically using their own forms of mark-up. Consequently, input formats may vary enormously, and could include documents in the form of HTML, RTF, PDF, Microsoft Word, PostScript, QuarkXPress, and many others. All this diversity needs to be squeezed out, so that we can start the next stage with clean plain text.

Decisions also need to be made about what to do with those parts of a corpus document that have limited value for lexicography. Most books, for example, include components such as:

- acknowledgements, copyright information, names of authors, page headers or running titles, tables of contents.

Many also have:

- indexes, glossaries, tables and diagrams, scientific or mathematical formulae, bibliographies.

What should be done with these? (The usual approach is to remove them, at least for a lexicographic corpus.) Newspapers and magazines have their own flavours of 'non-text' material, such as:

- lists of share prices, crosswords and sudoku puzzles, advertisements, TV listings, lonely-hearts columns, racing results.

Web data is even messier. Web pages have a natural bias against the kind of data that lexicographers most need: long stretches of continuous text. They tend to be broken up by (among other things):

- frames, navigational data, copyright notices, captions, images, and lists of various kinds.

But after a decade or more of work devoted to extracting text from the web, a number of well-tried methodologies are now in place. There are routines for automatically stripping out unwanted material, as well as ruses for identifying uninterrupted text. For example, if the ratio of textual characters to HTML tags is high, there is a good likelihood we have located a stretch of unbroken text – because in discursive text, tags are usually needed only for signalling things like paragraph breaks or the occasional change in typeface.

Similarly, a stretch of data where the proportion of grammatical words is high (and similar to what is found in printed documents) is a good candidate for being usable text. We have seen how the most frequent words in the language – like *the*, *in*, *but*, and *of* – make up a high percentage of most texts (§3.4.1.1), but words like this would be scarce or non-existent in a table of football scores or a list of team members. There is still much to be done in this area. Duplication, for example, is endemic in much web data and remains a difficult problem to solve. But the various strategies are being steadily refined, and there is an active community of researchers who are pooling ideas and expertise to improve automatic data-extraction techniques.[31]

3.6.1.2 *Spoken data* Spoken data has its own special challenges. Transcribing recorded speech is an inherently difficult task (§3.5.2), and the transcription system used will need to cope with phenomena such as pauses, vocalized pauses (like *erm*, *mhm*, *ooh*), overlaps, contractions (how do we deal with things that sound like *dunno*, *gonna*, and *cos*?), paralinguistic features (like laughing or whispering), or with words that are too unclear to allow for confident transcription, or simply inaudible. A good place to start is Crowdy's (1994) paper on the transcription scheme used for the BNC. Spoken data, and the 'depth' to which it is transcribed, nicely exemplifies an issue that we discuss at the end of this chapter: the trade-off between size and granularity.

3.6.1.3 *Encoding* Once the entire corpus has been converted into clean, raw text in a standard format, we are ready for the next phase. Encoding corpus texts typically consists of the following stages:

- tokenization
- marking textual structure
- lemmatization.

Each stage entails a form of mark-up. 'Marking up' a text means enriching the raw data by adding information of various kinds. This is done by means of 'tags' enclosing those strings of characters which embody a particular textual feature. To give a simple example, a run of words that constituted a

[31] The main player here is a special-interest-group called CLEANEVAL, set up specifically to evaluate technologies for 'cleaning' web pages: http://cleaneval.sigwac.org.uk/ .

paragraph in the original text will typically be marked by a tag at the start and a tag at the end.[32]

Tokenization refers to the task of identifying all the tokens in the corpus – which effectively means marking the boundaries of each word and punctuation mark. As Sinclair notes, it is reasonable – for most kinds of linguistic analysis – to see the word as 'the simplest building block' of any text (2003: 179), so routines are needed to automatically find and mark each word. This is relatively easy for Western writing systems, because word boundaries are typically marked by spaces. (It is much harder in the case of Chinese, for example.) Even so, it is not entirely straightforward. Hyphenation can sometimes be a problem. Some hyphens are intrinsic to the word they appear in (as in *trade-off* or *once-over*), but many are just an arbitrary product of layout. (Most news text, for instance, appears in short columns and this leads to a lot of hyphenation.) Similarly, apostrophes can be ambiguous in terms of their function – they may signal a quote mark, a contraction (*wasn't*), a possessive (*the boys' changing room*), or occasionally a plural (*NGO's*). So if our goal is to replicate the structure and wording of the source texts, we need software routines that address issues like this. The output of a tokenization process will look something like this, with words enclosed by a <w> tag and punctuation marks by a <c> tag (note that 'didn't' is treated here as two 'words': *did* and *n't*):

Input text
She really didn't like him.

Output text
<w>She</w><w>really</w><w>did</w><w>n't</w><w>like</w><w>him
</w><c>.</c>

Next, the *textual structure* of corpus documents needs to be recorded: it is important for corpus-users to be able to see the boundaries of chapters, sections, paragraphs, and above all sentences (since knowing where a sentence begins and ends is critical to understanding a text). Again, this is not always as easy as it sounds. Although most sentences end with a full stop, there are plenty of exceptions:

[32] This book is primarily concerned with how lexicographers use language data, so the account we give here of text-encoding processes covers only the basics. For fuller information, see for example Sinclair (2003: 181ff.), Kilgarriff, Rundell, and Uí Dhonnchadha (2007), and the BNC website.

- Some sentences end with question marks, exclamation marks, or closing quotes.
- Full stops don't invariably signal sentence boundaries: they can also appear in abbreviations (*i.e.*), numbers (*38.4 degrees C.*), references (*Sinclair 1991.124*) or initials (*D. H. Lawrence*).

Finally, the text needs to be lemmatized. Continuous text consists of 'word-forms' (like *permitted* or *permits*), but the headwords in a dictionary are generally 'lemmas' like *permit*, and this is the object that we usually want to study. A lemmatization program takes as its input the various word-forms and maps them on to the lemma they belong to. When the lexicographer is researching *permit*-verb, a single query will find corpus instances of *permit*, *permits*, *permitting*, and *permitted*. Once again, English – with its simple morphology – is one of the easiest languages to lemmatize, but things are far more challenging in the case of morphologically complex languages like Finnish, Irish, or Xhosa.

For all three processes, a great deal of work has been done over several decades, and mature technologies for automating these tasks are already in place for many languages (and above all, for English).

3.6.2 Textual annotation: the document header

In our discussion of 'skewing' (§3.4.2.6), we noted that the word *mucosa* had the same frequency in the BNC as the word *unfortunate*. But it turns out that *mucosa* appears in only nine of the BNC's source-texts, whereas *unfortunate* is much more evenly distributed, occurring in 648 different texts. For lexicographers and other corpus analysts, this information is vital to any assessment of the relative 'importance' of each word – and this is where the document header comes into play. We know exactly which texts – and which kinds of text – *mucosa* appears in because each individual text (or 'document') in the corpus includes a unique header that tells the computer (and hence the user) what kind of text it is. For example, is it a written text, or a sample of speech? Is it a piece of fiction or an academic monograph? The headers have to provide 'whatever information the user might need about a text, including feature-values which would be potentially used in corpus queries' (Kilgarriff, Rundell, and Uí Dhonnchadha 2007: 141). Headers typically give bibliographic information (title, author's name, date and place of publication, and the like), and precisely locate each text in

whatever typology is being used. Thus if we come across the word *sacrifice* in our corpus, in what looks like an unfamiliar use, it will be helpful to know that in this case the example comes from a book about baseball written in the US in 1999. The header might classify this source text as being *informative* (as opposed to *imaginative*), *written* text belonging to a broad *sports and leisure* category and in the subdomain *baseball*, and as *American English* from the *1990s*. Recording features like these for every corpus text ensures that lexicographers are not misled by counter-intuitive frequency data (as in the case of *mucosa*) and enables them to apply labels (such as *American English*, *literary*, or *journalistic*) with a degree of confidence. It also facilitates the process of generating subcorpora – of American English, for example, or of texts about sport. The types of information encoded in header documents will vary according to the intended uses of the corpus. In the Irish component of the New Corpus for Ireland, for example, the header states whether the author is a native speaker of Irish, and which (if any) of the three major dialects of Irish is used in the text. The obvious rule is that the more information we encode in the header, the broader the range of questions we can ask the corpus. If the header includes the author's gender, for example, we can compare the ways that women and men use language.[33]

3.6.3 Linguistic annotation

One of the more unlikely success stories in UK publishing of the last decade was a short, rather prescriptive book about language called *Eats, Shoots and Leaves*. The title alludes to an old joke popular in linguistic circles, about a panda who finishes his meal in a restaurant, guns down the waiter, and walks out – thus conforming to his dictionary definition: a large bear-like mammal which 'eats shoots and leaves'. Though the book mainly addresses issues of punctuation, its title is equally relevant to grammatical categories. The two possible readings for *shoots* and *leaves* depend on whether these word-forms represent plural nouns or present-tense verbs in the third person singular. A lexicographer compiling an entry for the noun *leaf* needs a corpus system which displays every instance of the noun (*leaf* and its plural form *leaves*), but also filters out occurrences of the verb-form *leaves* – which, for present purposes, would constitute

[33] See also http://www.natcorp.ox.ac.uk/docs/userManual/hdr.xml for details of the BNC's approach to header documents.

'noise', or non-relevant data. This is perfectly possible as long as the text in the corpus has been 'part-of-speech tagged' (or POS-tagged), and this is the most common form of linguistic annotation applied to corpus data.

A POS-tagger is a software tool that automatically assigns every word in the corpus to a wordclass. The term is slightly misleading because most taggers (for English, at least) go well beyond simply saying 'this word is a noun, this one's a verb, this is an adjective'. The tagger used for the BNC, for example, has different tags for singular and plural nouns, proper nouns and common nouns, and it distinguishes various forms of lexical verb (such as the third person singular, past tense, past participle, or *-ing* form). The system (known as 'CLAWS-5') has fifty-seven main grammatical tags, as well as several others for punctuation marks. POS-tagging has been a major research topic in the NLP community for a long time, and taggers which perform with a high degree of accuracy are available for most major languages (and for many less major ones). This is not the place to explain how POS-tagging works – there are plenty of good sources for this kind of information.[34] Our focus here is on the lexicographic value of a well-tagged corpus.

Our input text from above (§3.6.1.3) has now been tokenized and its sentence-boundaries are marked by <s> tags. So it looks like this:

<s><w>She</w><w>really</w><w>did</w><w>n't</w><w>like</w><w>him</w><c>.</c></s>

After it has gone through the POS-tagging process, it looks like this:

<s><wt="PNP">She</w><wt="AV0">really</w><wt="VDD">did</w><wt="XX0">n't</w><wt="VVI">like</w><wt="PNP">him</w><c>.</c></s>

Each word tag <w> has now been enriched by a POS-tag. Thus *She* is tagged 'PNP' (personal pronoun), *really* is tagged 'AV0' (general adverb – there are separate tags for adverbial particles like *off* and *out*, and for wh-adverbs like *why*), and so on. With this information built into the text, a corpus-query system will be able not only to find all the noun uses of *take* (and thus save the labour of identifying them from the far larger set of verbal uses), but also to conduct quite sophisticated searches and generate results that are relatively free of noise.

[34] See for example Grefenstette (1998), and McEnery and Wilson (2001, chapter 5).

A common way of analysing a word is to start with a set of, say, 500 randomly sampled concordances. You may then want to perform some more complex (and more focused) queries. For example, the verb *seem* has almost 60,000 hits on the BNC, a daunting number. But we quickly notice (or retrieve from our mental lexicon) its tendency to be followed by an adjective (*the dog seemed distressed*), a noun phrase (*he seemed the embodiment of the new age*), or a prepositional phrase (*they seemed in good spirits*). A well-tagged corpus allows us to focus on each pattern in turn and view a manageable number of examples, because we can specify queries like '*seem* + noun' within a given span. Similarly, we could use the tagging in the corpus to collect evidence for the verb *train* (excluding all the noun uses) when directly followed by a prepositional phrase, thus enabling us to investigate the distinctive uses of the patterns 'train *as*' (*train as a nurse, an engineer, a teacher*), 'train *in*' (*train in management, medicine, first-aid techniques*), and 'train *for*' (*train for a diplomatic career, a job in sales, the probation service*).

For lexical-profiling software, too (which we discuss in §4.3.1.5), the first requirement is a POS-tagged corpus. Word Sketches (a well-established type of lexical profile) produce a statistical summary of a word's grammatical and collocational behaviour. A Word Sketch for the verb *exercise*, for example, will quickly tell us the kinds of object the verb usually takes: words like *restraint*, *discretion*, *caution*, and *vigilance*. In order to do this, the software needs to be told how to recognize the object of a verb, and this is done by identifying the various sequences of POS-tags that can instantiate the V + O relation. At its simplest, this could be a combination such as 'verb + determiner + adjective + noun' (where the noun is the object: e.g. *committed a serious crime*). By aggregating the various definitions of this grammatical relation, the software is able to find most instances of the relevant pattern and will then identify the specific lexical items that most regularly appear in the object slot.

It is important to point out that neither approach (lexical profiling, or complex concordancing) is perfect. The results will always include the occasional false positive (*She had trained for six years* is not the same as *She had trained for an acting career*), and will also miss valid examples in cases where the system's grammar fails to capture every possible instance of a given relation. But it is equally important to be clear that – for the purposes of lexicography – *this doesn't matter*. What lexicographers are concerned with is identifying the regularities in the language, and a large,

well-annotated corpus enables us to see these. Lexical-profiling software in particular shows only those patterns that occur frequently, so the technology is tolerant of the occasional glitch. There are, to be sure, a number of methods that will deliver 'cleaner' results: the various types of 'corpus query language' (CQL) provide powerful tools for creating very precise queries which eliminate most of the noise. These are invaluable for corpus linguists and other researchers, but rarely needed for lexicography. Our job is to look for the norms, not the oddities, and dictionaries would never get written if we agonized about the imperfections in the results of our queries.

Even when the software is working optimally, it may still be unable to distinguish sentences like these:

guidelines for treating patients with AIDS
guidelines for treating patients with antibiotics

Though the surface grammar is the same (and POS-tagged corpora deal only in surface grammar), it is obvious to the human reader that these sentences are different, and the difference relates to what the prepositional phrase attaches to (*treating with antibiotics* vs. *patients with AIDS*). The problem is intractable unless our corpus is not only tagged but parsed, but this raises other issues. In the first place, automated parsing – the process of identifying the grammatical structure of a sentence – has a significantly lower success-rate than POS-tagging; the latter claims over 97 per cent success, while estimates for the reliability of parsers rarely go above 75 per cent. There is, moreover, a philosophical argument against using parsed corpora for lexicographic analysis. The categories we use when parsing a sentence or text reflect a model of sentence constituents which was developed in the pre-corpus era. But part of the point of using a corpus is to discover facts about language that we didn't know before. So it can be argued that if we apply existing notions to the analysis of sentence structure, we may be boxing ourselves into a corner and perpetuating a particular *a priori* approach to sentence analysis. As Sinclair points out: 'The theoretical position and descriptive strategies of those who performed the analysis . . . provide the only perspective through which the language can be viewed' (2003: 187). As long as concordancing and lexical-profiling software continue to improve, it seems unlikely that lexicographers will feel the need for fully parsed corpora.

3.7 Corpus creation: concluding remarks

3.7.1 The trade-off between size and granularity

We have shown how the results of even carefully specified searches will generally include a certain amount of noise. We have also argued that – for skilled human analysts – this is not a problem because irrelevant data can be rapidly discounted. If you get a Word Sketch from the BNC for the noun *adult*, you will notice it often modifies other nouns (in combinations like *adult suffrage*, *adult education*, and *adult literacy*). But the software also tells us that the fourth most statistically significant combination is *adult worm*. The intelligent lexicographer does not waste time pondering whether this combination should feature in a dictionary entry for *adult*: we know at once this is an aberration. If we follow up the link and generate all the concordances for *adult* + *worm*, we'll see that all but two corpus instances come from a single file (not surprisingly, a book about parasitology). This reassures us that the combination is not sufficiently part of the general language to be accounted for in a dictionary – but in practice, few lexicographers would even bother seeking such reassurance. Part of the skill of analysing corpus data lies in knowing how not to get sidetracked. To be sure, there is scope for the software to be improved: in the present case, the search algorithm could be refined, with an additional weighting to reflect how widely (or narrowly) an item is dispersed across the corpus. Something on these lines would eliminate *adult* + *worm* from the list of significant combinations. This sort of dialogue between users and designers of corpus-querying tools is what ensures that the software goes on improving. But human intuition still has a vital part to play – not in pre-judging our analysis by (for example), applying *a priori* beliefs about what the senses of a word are; but rather by sensibly ignoring things that we know to be unimportant for our purposes.

All of which raises an important point about the difference between what lexicographers want from a corpus and what other kinds of corpus linguists may want. As a general rule, lexicographers prefer size to granularity. That is, if the choice is between high volumes of data with the occasional bit of noise, or very 'clean', carefully annotated data in much smaller quantities, they will always go for the former. And this is a choice which often needs to be made. There are at least three areas in which this size/granularity trade-off might have implications for corpus design:

- text-selection parameters
- level of detail in document headers
- linguistic annotation.

Taking these in turn: it is obvious that we can categorize texts (when selecting material for a corpus) using typologies that fall anywhere on a spectrum from very broad to very fine-grained. Thus, the text-type 'newspaper' could be thought of as subsuming 'reportage', 'editorial' material, and articles or features in specific domains. The design of the Brown Corpus reflected distinctions like these; in a corpus of a million words, it was feasible to collect 2,000-word samples of various subtypes. But what about the OEC, which includes 190 million words of news text? Here, the task of separating the various sections in a given newspaper – such as those devoted to health, arts and culture, or personal finance – cannot be performed manually. Automated routines will help up to a point, but compromises are usually necessary, and it is unrealistic to expect that every item in every newspaper can be distinguished in terms of domain.

The same point applies to document headers: in an ideal world, the header will record every potentially relevant fact about the text it is attached to and the text's author. The more information in the header, the more kinds of query become possible. But the relevant information has to be found and verified, and recording it is a highly skilled, labour-intensive operation.

And finally, annotation. As we have seen already, automatic POS-tagging is reasonably reliable; the usual claim is that systems like this assign POS-tags with an accuracy of about 97 per cent. But it is by no means unusual to find mis-taggings (which impact on search results), and a program of manual post-editing could raise the accuracy level to close to 100 per cent. For a large corpus, however, this would be a major undertaking. When handling spoken data, we face an even wider range of options. The amount of information about speakers and speech segments is almost endlessly expandable, and there are indeed spoken corpora (such as the spoken parts of the Survey of English Usage and its successors[35]) which are finely annotated to take account of prosodic features like stress, intonation contours, and rhythm. There is no doubt that the more information we encode in our corpus, the more sophisticated the searches we can conduct. This will

[35] Information on the DCPSE (Diachronic Corpus of Present-Day Spoken English) – which is both parsed and finely annotated in terms of speech features – can be found on the Survey's website: http://www.ucl.ac.uk/english-usage/ .

yield fuller information *and* less noise. All this is to state the obvious. But the question we have to ask is not whether these things are possible, but whether they are worth doing. For many kinds of research, a corpus with meticulously detailed headers and fine-grained linguistic annotation (manually checked to guarantee high levels of accuracy) is precisely what is needed. But when building a lexicographic corpus, we need to keep in mind the kinds of information that dictionary-makers actually need, and to ask whether additional processing is worth the cost and effort involved. The fact is that – for most purposes – standard POS-tagging together with broad categorization in headers is perfectly adequate. Once these basic requirements are met, we face a choice between acquiring text in very large volumes (and tolerating the odd imperfection) or focusing on the creation of corpora of unimpeachable quality. For lexicographers, the choice is simple, and it is worth noting that the granularity of the major corpora (Brown, then BNC, then OEC) has declined as the volume of data has increased.

3.7.2 Final thoughts

In this chapter we have outlined a methodology for building a corpus for use in lexicography. The approach we describe takes account of relevant theoretical work (in text-type analysis and patterns of vocabulary distribution, for example); of many decades of research and practice within the language-engineering community; and above all of the practical experience of corpus-developers and corpus-users since the earliest days of data-driven lexicography.

It will be clear that there is no such thing as a 'perfect' corpus – natural language is just too diverse and too dynamic to allow us to think we can create an impeccably representative sample of it. Furthermore, before we can say whether a corpus is 'good', we have to ask 'good for what?' The design of a corpus, and the delicacy with which it is annotated, depend critically on the uses to which the corpus will be put. And we have argued that, for lexicography, the best, most useful kind of corpus is one that combines very large volumes of data with diversity in a number of broad categories (like mode, medium, and domain), and a level of linguistic and textual annotation which aspires to high quality but does not seek perfection. The value of such a corpus for dictionary-makers will become apparent in later chapters of this book, but the single biggest benefit is the access it gives us

to the 'regularities' of the language – the typical and recurrent features and patterns which make up the norms that lexicographers seek to identify and describe. There is no longer any serious argument about whether or not to use corpora in creating dictionaries. The use of corpora can be taken as a given, and our main concerns now are with optimizing corpus-querying software in order to make it faster, more efficient at tracking down the information we need, more proactive in alerting us to lexicographically relevant facts, and better-adapted to helping us discover new and unsuspected information about the way language works.

Reading

Recommended reading

Biber 1993; Atkins, Clear, and Ostler 1992; Sinclair 2003; Grefenstette 1998; Kilgarriff and Grefenstette 2003; Kilgarriff, Rundell, and Uí Dhonnchadha 2007; Landau 2001: 273–342; Fillmore 1992.

Further reading on related topics

Corpus design and annotation: Biber 1988; Biber, Conrad, and Reppen 1998 (Part IV); Crowdy 1993, 1994; Kučera and Francis 1967; McEnery and Wilson 2001 (chapter 5); Prinsloo 2008; Renouf 1987; Rundell and Atkins 2008; Rundell and Stock 1992; Sinclair 1991 (chapters 1 and 2); Summers 1993; van Dalen-Oskam, Geirnaert, and Kruyt 2002.

Corpora and the Web: Baroni et al. 2006; Braasch 2004; Fletcher 2004; Grefenstette 2002; Keller, Lapata, and Ourioupina 2002.

Websites

Corpora: http://natcorp.ox.ac.uk/ (BNC); http://americannationalcorpus.org/ (ANC); http://corpus.leeds.ac.uk/internet.html (Web Corpora); http://www.cis.upenn.edu/~ldc (Linguistic Data Consortium); http://www.terminotix.com/eng/index.htm (Canadian Hansard bilingual corpus)

Corpora with querying tools: http://www.sketchengine.co.uk/ (the Sketch Engine: a complete corpus query system with numerous ready-loaded corpora and corpus-building tools); http://www.kwicfinder.com/KWiCFinder.html (KWiCFinder: online multilingual concordancer and research tool); http://corp.hum.sdu.dk/cqp.en.html (CorpusEye: online multilingual concordancer)

Discussion list: to join CORPORA – the premier discussion list for corpus linguists – go to http://nora.hd.uib.no/corpora/sub.html and fill in the form.

4

Methods and resources

4.1 Preliminaries

In Chapter 3 we discussed what is nowadays the most likely source of evidence for a team writing a dictionary of current language, the corpus. This chapter describes the role this corpus plays in a dictionary project, and the environment the lexicographers work in. Figure 4.1 gives an outline of the topics we cover.

4.2 The dictionary-writing process

There are many different ways of using a corpus in dictionary production. Often a publishing house owning a dictionary will want you to start from the text of that dictionary, adapt it to suit the new specifications, and update it according to corpus evidence, editing the wordlist itself as well as the entries. This is always enriching for the book, but can be very frustrating for the editing team, because the budget is often too tight to allow for true corpus analysis, and the resultant 'corpus-based' dictionary falls between two stools. More rewarding (if more labour-intensive) is to start afresh and work systematically from corpus to dictionary.

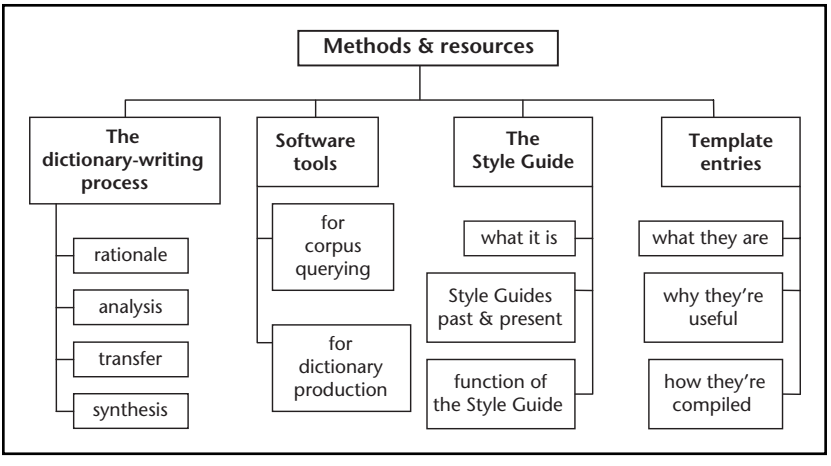


Fig 4.1 Contents of this chapter

The process described here is the ideal way to compile a corpus-based dictionary from scratch: it was developed during the editing of the *Oxford-Hachette English-French French-English Dictionary* (1994); it has been successfully applied in various adaptations on various projects since then; it uses the talents of different types of linguists and lexicographers to the best advantage; and it is, we believe, the most economical way of compiling, from corpus evidence, a dictionary which gives a true reflection of the language it describes. This method is twofold in the case of monolingual dictionaries (summarized in Figure 4.2), and threefold for bilinguals (see Figure 4.3).

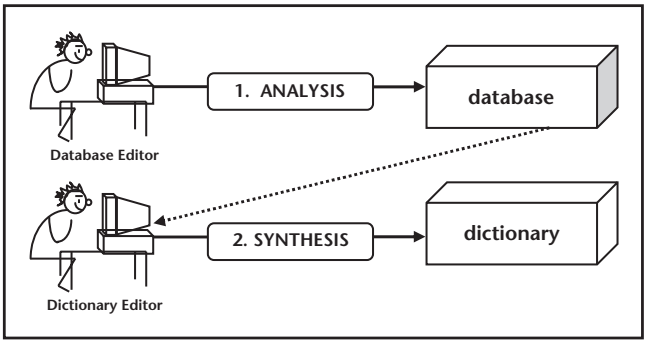


Fig 4.2 From corpus to monolingual dictionary: the twofold process

Lexicographers differ in what they do best: some are better at analysis, some at translating, and some at dictionary-entry writing. Separate these tasks and you use the whole team to best advantage. The various stages are outlined below, and form the focus of Chapters 8 through 12.

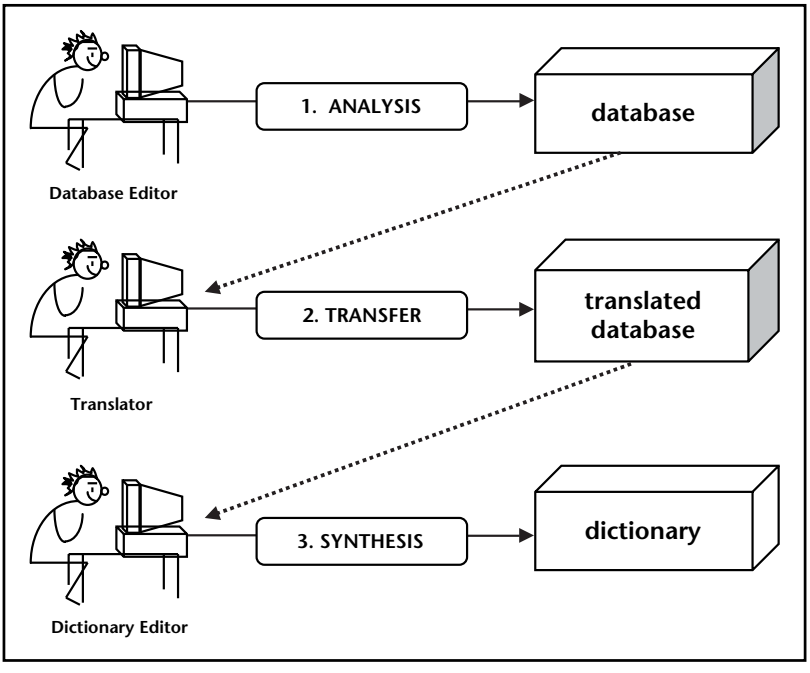


Fig 4.3 From corpus to bilingual dictionary: the threefold process

4.2.1 Rationale

The database produced by the 'analysis' process (described in detail in Chapters 8 and 9) is complex, and you might wonder why it's needed at all. Why can't we simply write dictionary entries straight from the corpus, especially since our corpus-querying tools are now so sophisticated? Well, we can, of course, and many dictionary-makers do exactly that. In a contemporary publishing environment, a good dictionary writing system (DWS: §4.3.2) provides a clear framework for editors to work within, and offers a lot of guidance on the content of dictionary entries. In this model, the compilation route is straight from corpus to dictionary, without a pass through a database. But although a dictionary entry produced in this way will reflect the evidence of the corpus, with this editorial approach you write your entry without ever having a systematic and comprehensive overview of the lexicographically relevant facts about your headword. This overview is what a pre-dictionary database provides. Moreover, a preliminary pass through the data is an essential part of writing a bilingual dictionary entry: it's impossible to supply adequate target language equivalents without

knowing a great deal more about the contexts in which the headword is found than can eventually be included in the actual entry.

The advantages of storing the facts about the headword in a relational database include the following:

- The structure of the database guides the analysis process: by specifying the types of fact to be identified and recorded, it reduces the risk of significant features being overlooked.
- The completed database holds a comprehensive record of how your headword behaves in the corpus.
- The database allows editors to scan the material in a systematic way, making it easier to decide how the word should be presented in the dictionary, and which of the many facts found in the corpus they should select for their purpose.
- The database speeds up the editing process: if the corpus analysis is done thoroughly, and the database is carefully designed, editors will rarely need to go back and look at raw corpus data.
- The database is re-usable: after the first dictionary has been drawn from it (say, a large monolingual), the database can be used as the basis for a bilingual dictionary; then perhaps a grammar book or other reference resource.

4.2.2 Analysis: compiling the database from the corpus

All dictionaries are in a sense databases. But we use the term here in a rather specific way, to refer to the structured collection of linguistic data assembled during the analysis stage of lexicography. The purpose of this database is to store selected facts about the word in a systematic way, so that by scanning them you can quickly and efficiently get a fix on the headword and extract the information you need for the final dictionary entry.

➔ The more detail the better, in the database, but avoid redundancies: remember the dictionary editors have to make sense of it all, as fast as they can.

The format of the database entry reflects that of a dictionary entry, but is much more detailed, as we shall see later (Chapters 8 and 9). It can hold a rich selection of corpus examples showing the headword in use in its various meanings and patterns (complete corpus sentences can be stored in the database, whereas they are often adapted for use in the dictionary);

the headword can be 'split' very finely into senses and subsenses, which can later be 'lumped' together for dictionary purposes;[1] information about the grammar of the headword is noted in special database fields, usually in formal codes; and its significant collocates (see §9.2.7 for discussion of what this term means) in the corpus are also noted in separate fields, each with its own example sentences. This highly structured format means that much of the information in the database is accessible to computerized searching and filtering. The database contents can therefore be re-used (and updated) after the primary dictionary has been extracted. From a commercial point of view, a well-designed and well-populated dictionary database represents a valuable piece of intellectual property. As well as being usable as a basis for all kinds of dictionary and as an information-source for linguistic research, it is likely to be attractive to builders of computer applications such as machine-assisted translation systems, information retrieval tools, and so on.

All the corpus searching, sense finding, collocate noting, and grammar coding is done in the analysis stage, freeing up the dictionary editors – who come along afterwards – to concentrate on fashioning entries that meet the needs of a specific target user. The value of the database lies in the fact that it is an unbiased record of what is happening in the one single language it is describing. It's often better for the database editors to have no knowledge of what kind of dictionary it will be used for – this stops them from making premature decisions about what is worth keeping and what isn't.

➔ At this point you're building a monolingual database: if your dictionary is a bilingual one, it's easy to be swayed by your knowledge of the target language and start picking out facts and examples on that basis – *don't do it!*

The analysis stage is discussed in detail in Chapter 8 (finding the senses, or lexical units, of the headword) and Chapter 9 (identifying what is worth recording for each of these lexical units, or LUs, and recording it systematically). For a monolingual dictionary, you go straight from analysis to 'synthesis' (see Chapter 10), and on the basis of the facts collected in the database, you write the most appropriate entry for your dictionary. For a bilingual dictionary, however, the 'transfer' stage (Chapter 11) necessarily precedes the dictionary-editing stage (Chapter 12), in which final entries are created from the accumulated data.

[1] More about lumping and splitting in §8.1.3.

4.2.3 Transfer: translating the database

The purpose of this stage is to build up a body of target language (TL) equivalents of the headword in as many contexts as possible, so that when the entry editors come to extract the final entry they have all their options assembled for them in one place. This work is best done by experienced translators with an excellent knowledge of both languages, and preferably native speakers of the target language: they are not necessarily trained lexicographers but are capable of fast and accurate translating. Lexicographers' skills are more effectively used when the entry is being constructed.

During the transfer, the database is partially translated – 'partially' because you don't want the translators to translate every phrase and corpus sentence in the database entry (that's a waste of time, it slows down the translation stage and often brings the editing stage to a grinding halt). You want the translators to come up with one or two 'general' translations for each LU, that is, the TL term that fits most of its corpus contexts. Then they work through all the corpus examples and offer headword translations *only* for the headword in contexts where the 'general' translation of that LU cannot be used. They don't translate the whole sentence, only the minimum necessary to make sense of the equivalence. They also provide TL equivalents of any MWEs in the database. This technique will become much clearer when you read the detailed description of the work in Chapter 11.

4.2.4 Synthesis: editing the entry

The purpose of this stage is to produce the final entry, the one most appropriate for the typical user of your dictionary. This work is best done by skilled and experienced lexicographers, though the shortest and most formulaic entries provide a good training ground for newcomers to the profession. (The fact that they are working from a coherent and systematic database entry makes it easier for them to pick out the most useful facts for their users; if they follow the Style Guide, they can't go far wrong here.)

The synthesis stage is reasonably straightforward when you are writing a monolingual dictionary: most problematic are the tasks of deciding on the dictionary senses, and devising definitions for these. This takes considerable skill at first, but this comes with experience (more on defining in Chapter 10). Extracting a bilingual dictionary entry from the partially translated database needs to be done by good bilingual linguists trained in

lexicography. Ideally, each entry will be drafted by a native speaker of the SL, checked through by a native speaker of the TL, then finalized by the SL speaker. All along the line you never lose sight of your typical user, or users if the dictionary is being prepared for both SL and TL speakers.
➔ A salutary thought – if careful users with adequate knowledge and skills have trouble with your dictionary entry, it's your fault not theirs.

4.3 Software

The processes described in §4.2 are supported by two types of software:

- a Corpus Query System (CQS): a computer program that enables you to analyze the data in a corpus in various ways
- a Dictionary Writing System (DWS): a program that enables lexicographers to compile and edit dictionary text, as well as facilitating project management and (later in the process) typesetting and output to printed or electronic media.

This section describes the features and benefits of both types of program.

4.3.1 The Corpus Query System (CQS)

In Chapter 3 we described the process of creating a lexicographic corpus, with the constituent texts encoded in a standard format (§3.6.1), linguistically annotated (§3.6.3), and enriched by 'document headers' providing information about each one (§3.6.2). This is the raw language data which you will study during the analysis process (§4.2.2), and the CQS is the program that makes this possible.

4.3.1.1 *Lexicographic needs and CQS functionality* A good way of evaluating a CQS is to start from the categories of information you want to include in your dictionary. As you compile the dictionary, you will have to:

- make decisions about what to include (headwords, variant forms, meanings, multiword expressions, and so on)
- identify word senses, explain their meaning, and decide what order to show them in

- describe combinatorial behaviour (syntactic preferences, collocations, phraseology, etc)
- assign labels to items that are characteristic of a particular region, style, subject-field or time period.

Your decisions on all these issues will depend on what the corpus reveals. And the question is not whether your corpus *contains* all this information (a large, well-constructed corpus certainly will), but how quickly and easily you can retrieve the information using your CQS.

4.3.1.2 *The KWIC concordance* KWIC (keyword in context) concordances are the basic tool of corpus lexicography. Figure 4.4 shows a concordance for the verb *taste*, which has been generated from the BNC, using the Sketch Engine CQS. What the CQS has done here is look for every

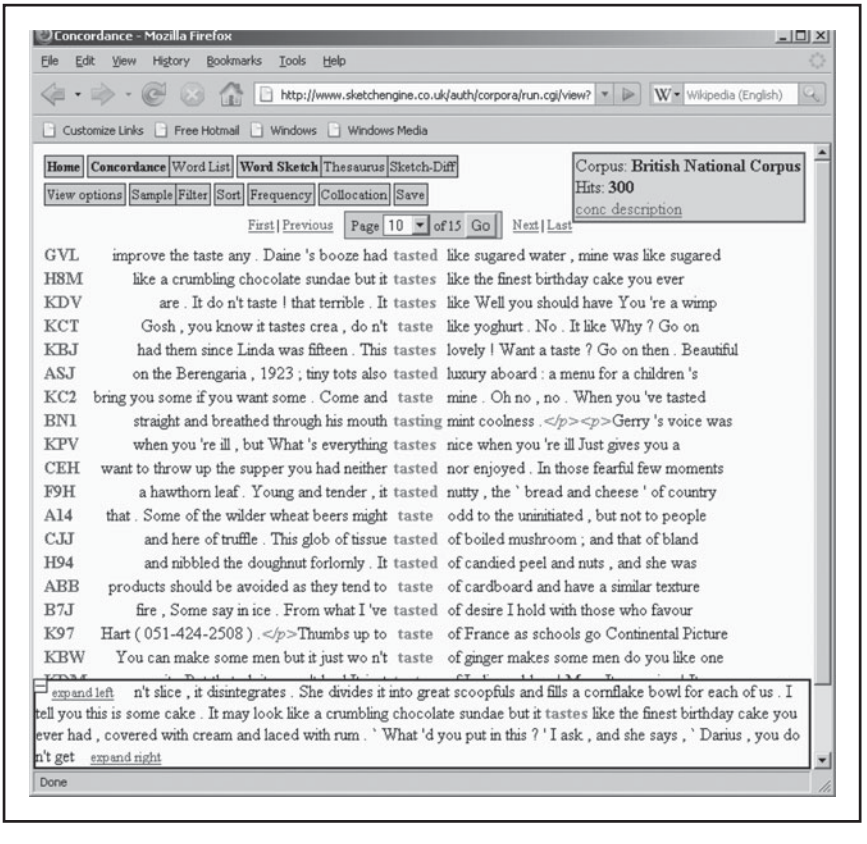


Fig 4.4 A KWIC concordance for the verb *taste* in the Sketch Engine

occurrence of the verb *taste* in the texts of the corpus, retrieve each instance along with about twenty words of surrounding context, and display them with the 'node word' (*taste*) in the centre of the screen. The software takes advantage of the following features of the BNC (cf. §3.6):

- Lemmatization: if you ask for a concordance for the lemma *taste*, the system automatically retrieves instances of *tastes*, *tasting*, and *tasted* as well as the base form.
- POS-tagging: this enables the CQS to retrieve data for *taste* as a verb, and to ignore instances of *taste* as a noun.
- Document headers: the alphanumeric codes in the left-hand margin indicate the source text where the adjacent corpus line occurs, and if you click on any of these, full bibliographical data will be displayed in a separate pane.

This concordance in Figure 4.4 is part of a sample of the available data, 300 lines for the verb *taste* randomly selected from the 1,408 occurrences in the BNC (the shaded box in the top right-hand corner gives frequency information). If the KWIC display doesn't provide enough information for a given line, you can see more of the source text by clicking on the node word: this opens up the pane at the bottom of the screen (which here shows more text for the second line in the concordance). Finally, the concordance here is right-sorted, meaning that lines are displayed following the alphabetical order of the word immediately to the right of the node. In a sorted concordance, different instances of the same pattern tend to cluster together, and in this case we see several examples of *taste like*... and *taste of*... One of the earliest revelations of corpus study was that right- or left-sorted concordances will often give a powerful, visual representation of a word's recurrent patterns – in a way that is impossible to ignore or overlook.

4.3.1.3 *How to frame a query* Scanning a sample concordance like the one in Figure 4.4 is a good starting point when you are investigating a complex word. As recurring features emerge from the data, you can use the CQS to conduct more specialized searches. Here, for example, a quick scan tells us that *taste* is often followed by an adjective (*tastes lovely*, *tasted nutty*, *might taste odd*, etc.), so we may want to see more examples of this pattern and get a clearer idea of how frequent it is. As Figure 4.5 shows, the query input form allows you to narrow down your search by requesting only those instances of *taste* which are followed by an adjective. In this query, the key

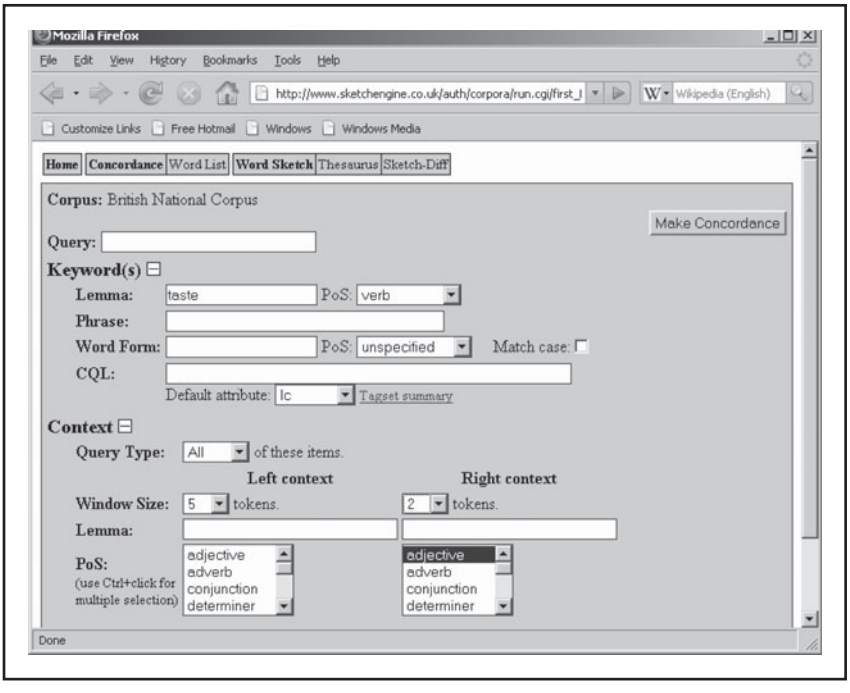


Fig 4.5 Inputting a query in the Sketch Engine

search word is (as before) *taste* as a verb; the drop-down list next to the lemma box allows you to select any of the main wordclasses. The 'Context' boxes in the lower half of the screen are used for specifying the text to the left ('Left context') or right ('Right context') of the node word, and you can either enter a specific word or simply select a wordclass; in this case, 'adjective' has been selected, in a 'Window Size' of two tokens to the right of the node. (The BNC has 420 instances of corpus lines that match this query.)

Endless variations are possible using the same processes. For example, you may notice – by scanning a random sample of concordance lines – that when *taste* is a noun, it often occurs in the pattern '*in* + ADJECTIVE + *taste*'. If you specify the word 'in' in the Left context 'Lemma' box, and select the wordclass 'adjective' in the 'PoS' box below it, the software will find expressions like *in (very/extremely) bad taste*, *in exquisite taste*, and *in the worst possible taste*. (The 'Phrase' box in the top half of the screen, incidentally, allows you to specify a precise search string like 'in the best possible taste'.)

Box 4.1 Making queries with CQL (Corpus Query Language)

When the user of a CQS enters search terms into the boxes provided, the program interprets these queries and converts them into instructions that the computer can understand. In the Sketch Engine (and some other types of CQS), these instructions take the form of a standard, widely used code called 'Corpus Query Language' (CQL). But you also have the option of using CQL directly (see the box labelled 'CQL' in the middle of the screen in Figure 4.5). A query in which the Lemma box contains the word *commit* and the Right context box selects 'noun' from the PoS list could be made equally well in CQL, like this:

[lemma = "commit"] [tag = "NN."]

The advantage of CQL is that it is powerful and flexible, and it allows you to make very complex searches; it is a good way of finding all instances of a particular grammatical pattern, for example. Should lexicographers familiarize themselves with this query language? On the whole, we suggest that this is unlikely to be useful. CQL has great value as a tool for linguistic research, but working lexicographers will rarely have time for in-depth investigations. When a CQS has been designed in consultation with lexicographers (as the Sketch Engine was), care is taken to ensure that all the most useful search routines – the kinds of search that lexicographers will need to make on a regular basis – can be made by filling in ready-made boxes like the ones in Figure 4.5.

4.3.1.4 *CQS functionality* A powerful CQS has a great many functions, the majority of which will be employed only rarely. A few will be in regular use, and the table in Figure 4.6 summarizes these, explaining the most important features of a CQS (the types of search it can perform), along with the corresponding benefits (the lexicographic tasks they facilitate).

4.3.1.5 *Lexical profiling: the 'Word Sketch'* Since the 1980s, the concordance has been the central tool of corpus lexicography. But this 'traditional' way of viewing lexical information begins to run into problems when data becomes very abundant. For lexicographers using the original 7.3-million-word COBUILD corpus, it was perfectly practical to scan all the available data for just about any headword. With the exception of function words like *the* and *out*, few items produced more than three or four hundred

Feature	What it does	Benefits
Basic KWIC concordance	Displays every instance of your search term as the node, with immediate surrounding context (see §4.3.1.2)	Provides a quick overview, helping you to make a provisional analysis of senses; reveals recurring patterns, providing cues for more specialized searches; a source of example sentences
Simple search	Finds occurrences of a particular lemma, word form, or multiword string	Allows you to see a concordance of *bargain* as a noun, ignoring verb uses; useful for retrieving data on multiword expressions like *into the bargain*
Complex search, specifying adjacent context	Finds occurrences of your keyword in particular lexical environments, within a user-defined 'window' to the left or right of the node	Allows you to search for (among other things) phrasal verbs (*bargain for*, *bargain on*); nouns modified by the keyword (*bargain offer*, *bargain prices*); and multiword expressions with variable elements, such as *strike a* [*tough*/*hard*/*excellent*] *bargain*, or a pattern like 'more . . . bargain for' (*got more than she bargained for*)
Sorting to left or right	The first (unsorted) view shows concordance lines in the order in which they are found in the corpus. You can then sort them several positions to right or left of the node, so that all occurrences of (for example) *run out of steam* cluster together	Especially useful for investigating recurring syntactic patterns (*decide to . . .*, *decide that . . .*, *decide on . . .*) and collocation (*pose a problem*, *pose a threat*, *pose a risk*)
Frequency information	A CQS will generally supply a wide range of frequency data. It will show at the very least how many matches the CQS has found for a query – whether this is a headword, phrase, or a recurring pattern like *taste* + any adjective. Most systems provide more detailed statistics	Helps you decide what to include, how much to say about it, and what order to put it in; provides the basis for explicit markers of frequency in the dictionary
Information on source texts	Shows the source of each line in the concordance (using the document headers of each corpus text); can give details of each source, or show a broad text-type to which the source text belongs (such as 'biography' or 'social sciences')	Supports labelling decisions e.g. *journalism* for an item that appears mostly in newspapers; shows whether a word is well-dispersed through many text-types or frequent in one text-type only, but otherwise rare
Longer extract of source text	Allows you to see more of the text from which a concordance line is extracted	The standard window of 20-odd words is usually adequate, but it is sometimes difficult to give a satisfactory account of an item without seeing more context. Words and phrases used as 'discourse organizers' are a good example: for expressions like *having said that* and *nevertheless* a standard concordance is unlikely to give you enough information, and the same may be true for expressions like *don't get me wrong* and *not to mention . . .*

Fig 4.6 Key features and benefits of a CQS

concordance lines. As corpora grew bigger, however, lexicographers faced the problem of how to handle very large concordances. In a corpus of 250 million words, even medium-frequency words like *coincidence*, *descend*, or *precise* will generate many thousands of lines. Having plenty of data is always better than not having enough, but scanning 5,000 concordance lines is neither practical (because publishing schedules won't allow it) nor efficient (because our short-term memories can't reliably process so much). Taking a sample is the simplest solution, but this entails the risk that valuable data will be lost in the process, and so reduces the benefits of having large amounts of data.

'Lexical profiling' offers a solution that maximizes the value of a large corpus while reducing the effort required by the human user. A lexical profile is a kind of statistical summary which reveals the salient facts about the way a word most typically combines with other words. We noticed earlier that *taste* is often followed by an adjective: a lexical profile will tell us exactly which adjectives most regularly fill this grammatical slot.

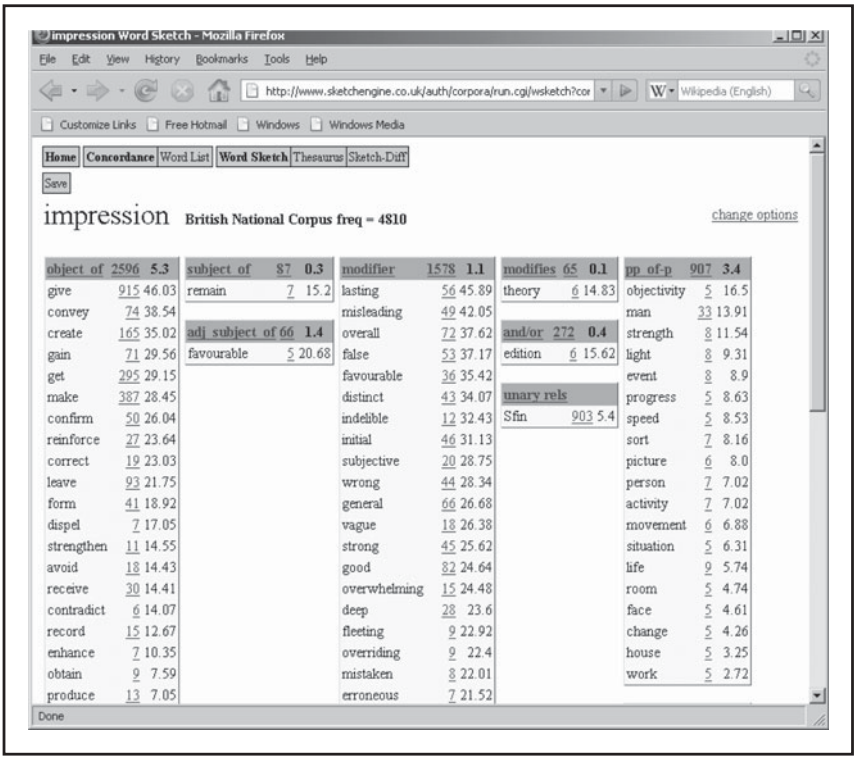


Fig 4.7 Part of a Word Sketch for *impression*

The Word Sketch is a type of lexical profile, and Figure 4.7 shows part of a Word Sketch for the noun *impression*. The first column, headed 'object of', shows that the BNC contains 2,596 instances of *impression* being used as the object of a verb. Below this heading is a list of the verbs that most frequently take *impression* as their object: *give*, *convey*, *create*, *gain*, and so on. It is, in other words, a list of collocates for a particular grammatical relation, in this case the V + O relation. (Similarly, the third column lists those adjectives which most frequently modify *impression*, while the fifth shows nouns that regularly appear after *of* in the pattern 'the/an impression of N'.) The first number following each listed collocate (e.g. 'convey 74') shows how many corpus lines instantiate this specific combination (and if you click on the number, you access – in a separate window – a concordance showing all 74 sentences in which *impression* appears as the object of *convey*). Collocates are ordered not by simple frequency (otherwise, *get* and *make* would come higher than *convey*). Rather, the order reflects the statistical significance of each combination, and this is indicated by the second number. (The higher the number, the greater the 'strength' of the collocation.)

The Word Sketch provides collocate lists for a wide range of grammatical relations (some of which appear in the screenshot in Figure 4.7), and achieves this by collecting every corpus instance of the search term and then subjecting this data to a further round of processing. Each grammatical relation is defined in terms of sequences of POS-tags: thus the software finds all the adjectives that frequently modify a noun by looking for any word with an 'adjective' tag within a certain span before or after the noun, with certain allowable POS-tags optionally present between adjective and noun. A certain amount of noise is tolerated, but the search routines are sophisticated enough to pick up collocates in non-obvious structures, such as the verb in a sentence like 'the *impression* she *conveyed* was . . . ' (Full details of how the software works and which probability statistics are used can be found in Kilgarriff et al. 2004.)

Lexical-profiling software adds a valuable resource to the lexicographer's repertoire. It was originally seen as a useful supplementary tool well-adapted for identifying collocational patterns – important information for pedagogical dictionaries. But the regular lexical environment of a word is one of the most reliable indicators of its senses (see §8.5.2.2) and editors have found that Word Sketches provide a compact and revealing snapshot of a word's behaviour and uses. For many lexicographers with access to this

type of software, the lexical profile has become the preferred starting point to their analyses of complex headwords.

CQS software is regularly improved and upgraded. In the Sketch Engine, for example, a further refinement to the Word Sketch program uses thesauric information to group collocates into sets of near-synonyms. Thus a Word Sketch for nouns occurring as objects of the verb *forge* can separate words like *relationship*, *connection*, and *alliance* from words like *passport*, *signature*, and *letter*, enhancing the program's value as an indicator of sense divisions. Graphic representations of word behaviour offer faster and more reliable ways of indicating lexicographically relevant facts. For example, bar charts or similar graphs can be used to show when a verb is passivized significantly more often than the norm, or when a word or phrase appears mainly in one particular type of text – thus enabling lexicographers to assign labels like 'usually passive' or 'mainly literary' with greater confidence. Additional refinements will no doubt add more value to these corpus-querying tools.

4.3.1.6 *The CQS: some conclusions* The CQS provides the link between raw corpus data and the dictionary. All the data you need to write your dictionary will be present in a good corpus, and a powerful and well-designed CQS allows you to retrieve relevant information efficiently and view it in a variety of ways. Improvements in the software open up new possibilities: they help us to perform existing tasks faster and more effectively, and they enable us to introduce completely new features. For example, the 'collocation boxes' in the *Macmillan English Dictionary* (an innovation when they first appeared in 2002, but now found in several dictionaries) were made possible thanks to an early version of the Word Sketch program running on the BNC. The information which these boxes supply was, of course, already embedded in the BNC when it was completed in 1993, but collecting it systematically would have been very labour-intensive without Word Sketch software. Thus additional functionality in the CQS can extend the scope of what dictionaries are able to do.

As new functions become available, lexicographers' search strategies evolve. A typical 'search cycle' might begin with a quick scan of the data (either through a sample of concordance lines or a lexical profile). Then, once you have provisionally identified senses, multiword expressions and any other LUs, each LU and its associated patterns and collocations can be investigated in greater depth. A good CQS offers a number of ways of uncovering information, and one of the skills a lexicographer develops

Box 4.2 Computers and lexicography: a brief history

The 1960s and 1970s

One of the pioneers of applying computer technology to lexicography was Laurence Urdang, Editor of the *Random House Dictionary of The English Language*, which was developed in the early 1960s. During the 1960s and 1970s, computers began to be used in the capture, storage, and manipulation of dictionary text. At this stage, lexicographers had no direct contact with the computer: they continued to write dictionary entries on paper (increasingly, using structured forms), and it was left to computer specialists to input the data. For the first time, computers were used to automate the arduous (formerly manual) process of checking cross-references. The first learners' dictionary to make extensive use of computers was *LDOCE-1* (1978). The dictionary's database included a comprehensive semantic coding system (which doesn't appear in the printed book). This added value to the text and made *LDOCE* popular with the natural-language-processing community. *LDOCE* also pioneered the idea of writing definitions using a limited set of high-frequency words (§10.6.5), and every definition was automatically checked to ensure conformity with this 2,000-word 'defining vocabulary'.

The 1980s

The first *COBUILD* project (1980–1987) 'placed great emphasis on the use of computers' (Clear 1987: 41), which were used not only for creating and exploiting a dictionary database, but also – for the first time in the development of an English dictionary – for generating concordances from the project's 7.3-million-word corpus. Nevertheless, lexicographers continued to work in traditional pen-and-paper mode. KWIC concordances were produced in microfiche form, and relevant alphabetical sections were printed out to form part of a lexicographer's 'workpack'. The concordances were right-sorted and of course completely static; there was no way of re-sorting, carrying out more specialized searches, or viewing the data in any other way. Dictionary text was written onto paper forms, with sections corresponding to fields in the dictionary database (see Clear 1987: 48 for an illustration). It was left to keyboarders to turn these entries into electronic text.

The 1990s

It wasn't until the beginning of the 1990s that lexicographers began to work directly on computers. The big dictionary publishers – in the UK, at

Box 4.2 (*Continued*)

least – acquired CQS and DWS software, either by developing their own programs or by buying them in. Hardware was still relatively expensive, so this method of working was possible only for staff based in publishers' offices. But by the second half of the decade, a combination of falling hardware prices, rising hard-drive capacity, and the arrival of email as a mainstream service made it possible for geographically dispersed editorial teams to work collectively on a single dictionary. The *Encarta World English Dictionary* (1999) was one of the first major dictionary projects to be run in this way, with a small in-house staff managing a large team of home-based editors on three continents.

From 2000 onwards

The availability of fast, always-on web connections underpinned the latest stage in computational lexicography, where both corpus data and dictionary text in progress are held on a server, which editors access online (from anywhere in the world) through the CQS and DWS.

over time is to know which functions to use in a given situation in order to get the most out of the corpus with least effort. Your CQS is probably capable of very fine-grained searches and you will sometimes be tempted to dig deeper. But time is always in short supply, and in practice you will tend to re-use a fairly small number of routines that maximize the value of your search. Knowing when to stop searching is as much a skill as knowing when a definition says enough.

4.3.2 The dictionary writing system (DWS)

Like any other written document, most dictionaries are now written on computers. When databases were first used for storing dictionary text, the job of inputting the data was left to technicians (Box 4.2). Nowadays, lexicographers compile dictionary text onscreen, and the software that allows them to do this is generally referred to as a dictionary writing system, or DWS.

The simplest form of text-input software – which has been used in a number of dictionary projects – is a generic XML editing tool such as Emacs (http://en.wikipedia.org/wiki/Emacs). Programs like Emacs can be customized for lexicographic work, and a (simplified) entry form might look

```
<headword>          </headword>
<POS>               </POS>
  <sense>             </sense>
    <definition>          </definition>
    <example>             </example>
  <sense>             </sense>
    <definition>          </definition>
    <example>             </example>
```

Fig 4.8 A simple entry form using an XML editing program

something like the one shown in Figure 4.8. Each entry component is signalled by a pair of tags (opening and closing), and the lexicographer inserts text between them. These are tried and tested programs, but they are not designed specifically for compiling dictionaries, and they won't necessarily be able to handle the special demands of a complex lexicographic project. Another option is to develop a 'homegrown' system, typically using an XML editor as a basis but customizing it for dictionary-making. Over the years, a number of major publishers have followed this route, often refining their systems in the course of several projects, with input from working lexicographers. But as specialized, off-the-shelf DWS packages have become available, publishers have tended to switch to these (the prestigious *Oxford English Dictionary* made this move in 2005).

A commercial DWS program is designed to manage the entire process of producing a dictionary, from compiling the first entry to outputting the final product for publication in print or electronic media.[2] A typical DWS consists of three main components, which are discussed below:

- a text-editing interface, in which lexicographers create and edit dictionary entries
- a database, in which the emerging dictionary text is stored
- a set of administrative tools which support the management of the project and the publication process.

4.3.2.1 *The lexicographer's interface: the editing tool* This is where dictionary text is compiled and edited. Lexicographers usually key text into boxes or spaces, as if filling a form, but the DWS will generally offer a number of

[2] Two of the best and most widely used packages are IDM's Dictionary Production System, or DPS (www.idm.fr/products) and the TshwaneLex dictionary compilation software (http://tshwanedje.com/tshwanelex/).

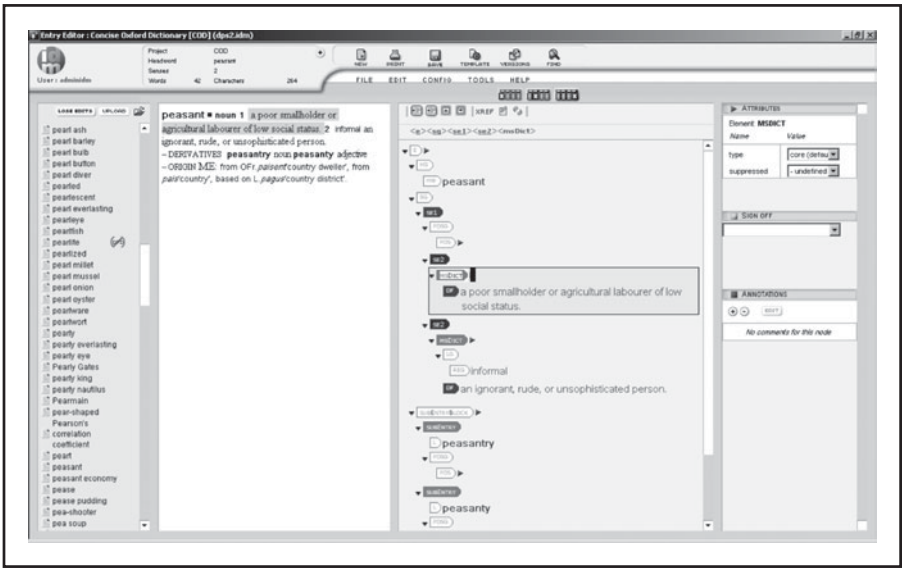


Fig 4.9 The lexicographer's interface in a DWS

ways of viewing the data. A typical screen might look like that from IDM's DPS shown in Figure 4.9. Here the left-hand pane lists the set of entries which have been assigned to the lexicographer or editor for working on, and the right-hand pane is for administrative functions. The central panes show the same data in two different views:

- a WYSIWYG view (or 'preview mode')
- a 'tree-diagram' view.

You can create or edit text in either of these panes (the other will automatically update itself when you do), but the usual approach is to enter text in the fields of the tree diagram. The WYSIWYG view is useful for getting an impression of how the final entry will look. The tree view reveals the structural elements of the dictionary entry (headword, wordclass marker, definitions, derived forms, and so on), and provides spaces where dictionary text is keyed in.

A good DWS maximizes the lexicographer's productivity by streamlining routine tasks and automating many of the 'administrative' procedures that used to be done manually. For example, drop-down lists may be used for entry components with a finite set of possible values (such as wordclass markers, grammar codes, or register labels); context-sensitive help is available if the Style Guide is integrated into the DWS (§4.4.2); and some systems

allow you to create 'templates', or generic entries, for common entry-types, containing ready-made configurations of structural elements which can be re-used whenever needed (cf. §4.5). Meanwhile – in background mode – the system can ensure that the syntax of each entry conforms to the dictionary's DTD (document type definition), which defines the constituent elements of the dictionary and the allowable sequences in which they can occur. Formerly arduous (and error-prone) tasks are now handled automatically: for example, if the senses of a polysemous word are re-ordered or a new sense is added in the middle of an entry, the system not only re-numbers the whole entry, but also makes appropriate changes to the sense numbers shown in any cross-reference to this entry from somewhere else in the dictionary. Real-time spellcheckers minimize the risk of typos, and if the dictionary uses a limited defining vocabulary, conformity to the words in this list will be checked too, as text is entered.

4.3.2.2 *The database* Text compiled and edited in the 'front end' of the DWS is stored in the dictionary's database.[3] On the whole, lexicographers won't interact with the dictionary in this format, but the DWS's database component makes it possible to run complex searches over the entire text. Using the system's query language, you could (for example) find:

- all entries written (or edited, or finalized) by a particular team-member between two specified dates
- every example sentence illustrating a particular syntax pattern
- every variant form that has an 'American' label
- any entry component that includes a specified word (which may be useful in cases of political or cultural sensitivity, or for the avoidance of litigation).

4.3.2.3 *Administrative tools* As well as providing an environment in which dictionary text can be written, edited, and stored, a DWS program will usually include 'housekeeping' tools that facilitate the management of a large dictionary project. The DPS software from IDM, for example, has a 'Workflow Manager' which allows the user to create 'workpacks' containing entries to be compiled or edited, and assigns them to a particular member of the team. The system keeps a complete record of the compilation cycle:

[3] Note that this is the generic use of 'database', and is not to be confused with the lexicography-specific use of that term as we use it in Chapters 8 and 9.

it knows who is doing what at each stage in the project, and it alerts the project manager to any batch of work that has overrun the time allotted for it. Once a batch is completed, it can be imported back into the database, and team members can always view the most up-to-date version of the text. Progress against the schedule and budget can be monitored continually, and senior staff can also keep a close watch on the extent of the text as it develops. (Ensuring dictionaries didn't overrun their agreed length was a major challenge in the pre-DWS era.). Systems like this – which are also a feature of the TshwaneLex package – ensure that only one person works on a given entry at any one time, and that individual fields in the database are 'lockable'. This means, for example, that all the pronunciation fields or all the etymology fields can be assigned to specialists, while regular team-members carry on working on the other parts of these entries. Once the text is complete, the DWS converts it into a form ready either for printing as a book or for publication in an electronic medium.

4.3.2.4 *Benefits* The DWS facilitates the entire process of creating and publishing a dictionary. As with CQS software, the trend is for the user's interface to be accessed online, and this environment makes it possible for widely dispersed editorial teams to work efficiently on the same project. A good DWS streamlines the editorial process and allows lexicographers to focus on lexicography, by relieving them of essential but fairly mechanical tasks. The system makes it easy for senior editors to review the text as it develops, monitor its quality, and give feedback to the editorial team. For project managers, the advantages are obvious. And for the publisher (who is also the budget-holder), the benefits include productivity gains, a smoother transfer of text to its eventual delivery format, and opportunities for re-using expensively created dictionary text, for example when updating an existing product or 'spinning off' smaller or more specialized versions from the database. For everyone involved in the project, the various features of the DWS help to deliver higher levels of quality, accuracy, and internal consistency.

4.4 The Style Guide

A dictionary is a complex object. It contains a wide range of information-types, many of which appear repeatedly throughout the dictionary. In

most monolingual dictionaries, for instance, almost every entry will have a headword, a wordclass label, and a definition, while many entries will additionally include style or regional labels, illustrative examples, multiword expressions, etymologies, or variant forms. Handling recurring elements in a consistent way is a basic principle of information management; a dictionary which labelled 'informal' items sometimes as *informal*, sometimes as *infml*, *inf.*, or *colloq.* would confuse a regular user and wouldn't inspire much confidence. For each entry component, therefore, the editorial team needs a set of guidelines. These guidelines show how the dictionary's style policies should be applied in individual dictionary entries. And the Style Guide – essentially a book of instructions for lexicographers – is the document in which all these guidelines are assembled. (For a good introduction to this topic, see also Landau 2001: 363–375.)

4.4.1 What kinds of information does a Style Guide include?

The information in a Style Guide will range along a continuum from unambiguous rules (for example, that in this dictionary *-ize* spellings, not *-ise*, will always be used in definitions), to general *principles* for handling elements such as example sentences, with plenty of detailed guidance in the middle. The Style Guide will show how each entry component should be dealt with, and will cover areas such as the following (though this list is far from exhaustive):

- Morphology: for example, will the dictionary show verb inflections, and if so when (in all cases, or only in irregular or semi-irregular cases)? How will the dictionary handle systematic varietal differences, such as consonant-doubling in British English but not American English (*travelled* vs. *traveled*)?
- Variant forms: in what circumstances can one word form be shown as a variant under another headword? For example, does the American word *aluminum* count as a variant of British *aluminium* (or vice-versa)? Should variant forms also have their own entries?
- Grammar: for example, what is the dictionary's policy on describing the syntactic behaviour of words, what codes or abbreviations are available for recording this, and will this information be provided for all headwords or only for high-frequency vocabulary?

- Labels and their use: for example, what criteria determine whether a word or meaning is labelled *offensive* or *dated*, and what is the difference between *American* and *chiefly American*?
- Definitions: for example, what defining styles are allowable, which abbreviations (if any) can be used in a definition, how can sexist language be avoided, and how should you define 'non-standard' entry types (such as affixes, expletives, or terms of address)?
- Examples: when should an example be shown, are sentence fragments (as opposed to complete sentences) allowable, and can corpus extracts be modified (and if so, in what ways)?
- Derived forms: should these be shown systematically, or should editors take account of the evidence of usage, and show only those forms which are in frequent use?
- Cross-references: what types of cross-reference are available (for example, 'compare X', 'see also X'), and when (and how) should they be shown?

To give a better idea of the level of detail found in a well-thought-out Style Guide, we will look briefly at a single entry component, the headword itself, and the complex issue of 'what counts as a headword'. To resolve uncertainties in this tricky area, the Style Guide will need to rule on questions such as:

- Plural nouns: in what circumstances do nouns such as *arms* or *customs* get full headword status, rather than being treated under the singular form?
- Participial adjectives: if *amazing* and *bored* are shown as headwords, what about *invigorated* or *irritating*, and what criteria can we use to decide?
- Compounds: should *point of view*, *in-your-face*, *upside down*, or *head and shoulders* get headword status, and if not, what other options are available for describing them in the dictionary? And if our corpus data shows that the forms *hardhat*, *hard-hat*, and *hard hat* are all valid, which do we show as the headword?
- Derived forms as headwords: if *hopefully* and *sadly* are shown as full headwords (because of their use both as manner adverbs and sentence adverbs) what about *fantastically* or *thinly* (the first being common as an intensifier, the second having very limited collocates, such as *disguised* and *veiled*)?

- Abbreviations and full forms: if the usual way of referring to something is an abbreviated form (such as *BBC* or *CIA*), should the main entry appear at the abbreviation or at the full form?
- Systematic spelling differences: for example, in an American dictionary, will the British form *harbour* be a headword, or merely a variant form at *harbor*?

For a bilingual dictionary, the Style Guide will also need to rule on such things as:

- the provision of translations, not only for headwords and meanings, but for other entry components too
- how to deal with items that have no direct target-language (TL) equivalent
- how to deal with abbreviations and their full-form, and their opposite numbers in the TL
- the labelling (for register, domain, style, etc.: see §7.2.8) of source-language (SL) items and their TL equivalents, both when they require the same label(s) and when they require different labels.

. . . and of course the many other aspects of producing a dictionary that deals with two languages which may be very far apart from each other both linguistically and culturally.

4.4.2 Style Guides past and present

Style Guides for teams without a DWS must include detailed instructions on issues such as:

- The correct order of the various parts of an entry: for example, does a regional label precede or follow a register label, and at what position in the entry does a variant form or etymology appear?
- The correct font to be used for each element: in some dictionaries, for example, ***bold italics*** may be used for showing collocational information, and SMALL CAPITALS for subject labels or cross-references
- The precise designation of a recurring element: for example, is an American English usage labelled as *AmE*, *NAm*, *American*, or *US*?

However, computerization, specifically the arrival of dictionary writing systems, has relieved lexicographers of many routine tasks that Style Guides traditionally ruled on. Data is entered in the form of plain text, and the software takes care of its eventual representation on page or screen. So, for example, the print edition of the *Macmillan English Dictionary* shows definitions in a plain 'roman' font and example sentences in italics, while the electronic edition uses a (non-italic) sans serif font for both components but shows definitions in black and examples in blue, but not in italics. But lexicographers don't have to worry about any of this: they simply key plain text into the relevant box, and the final output is generated by stylesheets. Similarly, the question of the order in which the entry components may appear almost ceases to be an issue because the writing system won't allow you to enter elements in the 'wrong' order.

The more options available to editors, the greater the risk of inconsistency among the members of a dictionary team. A good dictionary writing system (§4.3.2.1) will provide a list to choose from of items valid for any field where there is a finite set of options. The 'correct' forms are thus hardwired into the system. This allows lexicographers to concentrate on the linguistic issues (is this an informal usage or not?) rather than wasting mental effort on procedural ones (is the correct form 'informal' or 'infml'?).

Contemporary Style Guides no longer concern themselves with trivial issues like the correct form of a grammar code. However, they need to include instructions for inputting data in the right field, as the extract in Box 4.3[4] shows.

The traditional Style Guide was a printed document, and could easily run to several hundred pages. The inevitable changes in editorial policy during the course of a project would be implemented by means of update notices or appendices circulated to the editorial team. Nowadays, the Style Guide is an electronic document, typically accessed via a project intranet or wiki, and is thus easily updatable to accommodate policy changes. It may also be built in to the DWS, providing editors with context-sensitive help: thus a lexicographer working in the 'morphology' field, and uncertain of the rules for supplying verb inflections, can access the relevant Style Guide section with a single click.

[4] This extract from the Style Guide of the *New English-Irish Dictionary* is reproduced here by kind permission of Foras na Gaeilge.

Box 4.3 A Style Guide extract giving data-entry instructions

9.4.9 Framework Collocate Container (FwkCollCnt) in FwkSenCnt

When used directly within the Framework Sense Container (rather than within another container such as Framework Structure Container) the FwkCollCnt is used to show direct collocates of the headword which apply to a particular sense and not to all senses of the headword. In this case, the FwkCollCnt goes immediately after MEANING, expanding on and pinning down the information given in the MEANING field. For example, the headword *empty* in the sense of 'having no contents' goes straight into a FwkCollCnt containing:

- Collocate Type (COLLTYPE): *container*
- Collocates (COLL): *bottle, glass, can, tin, packet, box*, etc.

The Framework Collocate Container must contain:

- Collocate Group (CollGp)
- Example Container (ExCnt)

It can also contain:

- Collocate Type (COLLTYPE)
- Grammatical Information (GRAM) q.v.
- Label Group (LabelGp) q.v.

4.4.3 Why you need a Style Guide

The Style Guide is an essential resource in any dictionary project. It ensures that every member of the editorial team – even when the team is geographically dispersed – is 'singing from the same hymn sheet'. A clear, well-structured Style Guide resolves uncertainty in cases where straightforward rulings can be given, and provides advice in situations where lexicographers have to use their own judgment. All of this enhances editors' confidence and improves the efficiency of the compilation process. This in turn brings benefits to the dictionary user. Users gradually get to know how their dictionary works, and if it is well organized and internally consistent, they should find that unsatisfactory look-ups – which undermine confidence in the dictionary – are relatively rare.

➔ A final word of advice. A good Style Guide will provide the information you need in the great majority of cases. But the dynamic nature of human

languages means that there will always be situations where the Style Guide can't (or shouldn't) give a ruling. You will occasionally come across rare or anomalous linguistic features, where a blind adherence to Style Guide rules would result in dictionary text that is obscure or inelegant. Remember that the Style Guide is just that: a 'guide'. It is not set in stone, and you should feel free to challenge it where an existing policy doesn't appear to cover all relevant cases and an alternative solution may help to improve the dictionary.

4.5 Template entries

Template entries are 'pro forma' entries for use by the lexicographers writing either database or dictionary. In the latter case, the content varies significantly according to dictionary type. So we will revisit this issue in later chapters, specifically in §9.3 (templates for the database), §10.1.3 (templates for monolingual dictionaries), and §12.1.3 (templates for bilingual dictionaries). The purpose of this section is simply to explain the idea in general terms and to show how template entries are used.[5]

4.5.1 What they are and how they are used

A template entry is a framework designed to facilitate writing entries for words that belong to lexical sets. A 'lexical set' is any group of words that share a common element of meaning, such as the days of the week or months of the year, or birds, trees, flowers, and metals. In an ideal world, dictionaries would be compiled in lexical sets, so that the person who writes the entry for *lion* is also responsible for *tiger*, *cow*, *giraffe*, and *mouse*. But in reality this is rarely possible: so many words have multiple meanings (think of *mouse* as a small mammal, and the *mouse* you use with your computer) that their various senses could belong to several lexical sets, and this would necessitate an additional editing pass to assemble finished entries from their various parts.

[5] The templates we discuss here should not be confused with the 'entry templates' which you can create in a DWS, and which are specific configurations of entry-fields which you can 'grab' from a menu in order to speed up the compilation of recurring entry types (such as 'countable nouns' or 'phrasal verbs'): see §4.3.2.1.

Most of the dictionaries on your shelves will have been produced without the benefit of lexical-set compiling or template entries, and you can see that by comparing the entries for similar words. In the same English learners' dictionary, for instance, you can find these two definitions:

> **lion** *n*[C] a large strong African and Indian animal with four legs and light brown fur which eats meat and belongs to the cat family
> **tiger** *n*[C] a large wild cat which has yellowish orange fur with black stripes

Both definitions refer to:

- the animals' size, their fur, and their membership of the cat family.

The definition for *lion* also provides information about:

- diet, strength, number of legs, and habitat.

Yet these properties would be equally relevant in the definition of *tiger*. Conversely, the definition for *tiger* mentions the fact that it is a 'wild' animal (which applies equally to lions).

It's clear that these definitions were written by two different people, or by the same person at different times. In either case the two entries would have been planned separately, and the content and wording of each definition separately considered. But if you use template entries (in this case a template for 'animals', or perhaps for 'wild animals'), all this happens only once for each lexical set.

As a rule, the members of a lexical set pose the same kinds of lexicographic problem and should be handled in the same way in a dictionary. Once you've planned the entry for one word in the set, you can benefit from that work next time you meet another word from the same set. The template is designed to hold, in an ordered way, the essential facts about any word belonging to a specific category. The two principal aspects of entry-writing that benefit from the use of a template are:

- the structure and content of the entry
- the structure and content of the definition.

Here we look briefly at each type: what it is, and how it is used.

4.5.1.1 *Entry structure and content template* Figure 4.10 shows a template containing an entry structure and an outline of contents for the lexical set ANIMAL. It sets out the range of possible *lexical units* (LUs) that are likely

Lexical units	Notes, reminders etc.
1 – COUNT NOUN **LU meaning** **plural form** **definition**	Domain = zoology the specific animal (species) if irregular: check for 'collective plural' e.g. *they were after elephant* a [size] [wild / domesticated] [carnivorous / herbivorous] mammal, Latin name XXX, having fur / hide [colour, markings], found in [habitat]. Also called XXX. Remember: folk facts e.g. *lion is king of the beasts, cat has nine lives, mice are timid, foxes are sly, dogs are loyal . . .*
2 – COUNT NOUN	figurative uses, e.g. similes *(as) strong as a horse* idioms *to let the cat out of the bag* sayings *every dog has his day* etc.
3 – MODIFIER **LU meaning**	belonging to (*lion cub* etc) or made from some part of ANIMAL (*seal skin, fox fur, mink coat* etc.)
4 – COUNT NOUN **LU meaning**	figurative use of name to denote specific kind of person, *he's such a donkey, don't be a pig, it's a pig of a job* etc. (often informal)
5 – COUNT NOUN **LU meaning** **definition**	Domain = Zoology the genus e.g. *the cat family, the big cats* any [size] [wild / domesticated] [carnivorous / herbivorous] mammal of the genus (LATIN NAME), such as the (SPECIES NAME) or (SPECIES NAME), having fur / hide [colour, markings], found in [habitat].
6 – UNCOUNT NOUN **LU meaning** **definition**	Domain = cooking the flesh as food (*roast/boiled ANIMAL; I don't eat ANIMAL*) the meat of the ANIMAL
7 – MODIFIER **LU meaning**	Domain = cooking made with the flesh of ANIMAL (*lamb stew / turtle soup* etc.)
8 – NOUN SINGULAR **LU meaning** **definition**	Domain = Astronomy, Astrology constellation NB always capitalized and with 'the' (*the Lion* etc.) the less common name of the constellation LATIN NAME, the Nth sign of the Zodiac
9 – VERB	to do something in the manner of the ANIMAL *to dog sb's footsteps, to fox sb*
10 – VERB	to give birth to young of the ANIMAL *to foal, to pup* etc.
11 – PHRASAL VERB	consider any phrasal verbs *to rat on sb, to squirrel sth away, to wolf down one's dinner* etc.

Fig 4.10 Template entry for 'animals' lexical set

when the headword is an animal name, and suggests definition wordings that can be adapted to suit the needs of any specific headword.

Whenever you come to a headword belonging to a lexical set, a good first move is to check the relevant template entry. Instead of puzzling out the LUs for an animal-name headword, you can work through the ANIMAL template (shown in Figure 4.10), substituting your headword for 'ANIMAL' in the appropriate LUs. You then complete those LUs which are relevant to your headword, so, for instance:

- At an entry for *otter*, you would ignore LUs 4 through 11: there are no examples in the corpus of **what an otter you are!* or **have another plate of otter* or **stop ottering about!* and so on.
- If you are doing the entry for *elephant* or *lion*, you would want to record in LU1 any corpus use of the 'collective plural', such as *oxpeckers who live on the backs of large animals such as elephant, rhinoceros, and cattle in equatorial Africa*; *the decline of some species like lion and leopard*, and so on.
- Headwords like *lamb*, *goat*, and *horse* would probably require LUs 6 and 7.
- For headwords like *fox*, *mink*, and *ermine*, LU3 would come into play.
- In the entries for *lion*, *ram*, or *goat* you might want to complete LU8.
- If your headword is polysemous – *mole* for instance – then only one of your LUs is the name of an animal (the others could denote a long-term spy in an organization, a dark blemish on the skin, a breakwater, etc.); in such cases it is still worth starting from the ANIMAL template. Words like *emerald* (which is both a gemstone and a colour), *avocado* (fruit, colour), *pound* (weight, currency), and *cricket* (sport, insect) will activate more than one template.

Thus the 'animals' template acts as a checklist for use on entries in that lexical set.

Notice that the template in Figure 4.10 includes a proposed wording for some of the definitions (see LUs 1 and 5). This level of detail will usually be omitted from templates designed for *database* as opposed to *dictionary* compiling; if the database is designed to feed into a bilingual dictionary, the Latin names of animal species and genera are probably enough for the translators' needs (cf. §11.2.1). But definition models are invaluable in templates designed for monolingual dictionaries. We discuss these more fully later (§10.1.3), but they typically include:

- a choice of genus expressions (with guidelines as to which to use in which situations)
- a checklist of other possible defining features for you to choose from (along with recommended definition wordings).

4.5.2 Why templates are useful

Experience has shown that template entries have a useful part to play in any dictionary project, monolingual or bilingual. As many as 25 per cent of the LUs in a given dictionary could be written using some kind of template, and they . . .

- streamline the editorial process, by enabling you to extract and assemble relevant information much more quickly than you could otherwise
- ensure systematic and comprehensive coverage of the LUs involved: using templates, you are less likely to miss important facts, and when all the members of a team have access to them, the risk of producing widely differing entries for similar objects or entities is greatly reduced.

Writing template entries is also a useful form of lexicographic training for lexicographers. No project should be without them!

4.5.3 How templates are compiled

Compiling template entries is an excellent technique, not only for training novice editors in what is a lexicographically relevant piece of information (cf. §5.5), but also for helping everyone in an editorial team to reach a consensus about the kind of dictionary they are writing and what is important for that dictionary. Each person chooses one word from the lexical set being studied, and compiles the richest corpus-based entry they can for that word. These entries are then compared and collated in a discussion session, and the final version of the template entry is drawn up, with all possible relevant LUs included in it.

- First choose some ‘sample’ words in the lexical set (selecting words as far apart as possible on the scale), e.g. for the ANIMAL template we might choose *lion* (wild, large, dangerous, the stuff of myth), *unicorn* (imaginary), *rat* (small, often unwanted guest), *cow* (domesticated and

beneficial), *cat* (smallish and a pet), *fox* (wild and usually has a bad press), and so on.

- Share these words out amongst the lexicography team.
- Each person studies the corpus data for their particular word and outlines an entry. (At this point it's a good idea to have a brief discussion on how it's gone so far.)
- Everyone completes their own entry.
- The group assembles the facts into a template entry, discussing why something is good, why something else is not appropriate, the order of the LUs, etc.
- The resulting template can be used as a checklist entry for any word in the set.

A recent dictionary project identified over sixty categories of vocabulary amenable to template entry treatment, including the ones shown in Figure 4.11.

Template Category	Typical headwords
Academic Qualification	*BA, MSc, PhD*
Animal	*frog, poodle, lamb*
Bodypart	*knee, arm, heart, lung*
City Town Village	*London, Tokyo, New York*
Colour	*red, black, burgundy, ginger*
Currency	*euro, cent*
Day Of The Week	*Monday, Friday*
Festival	*Christmas, Passover*
Game Sport	*cricket, judo, hurling*
Gem	*diamond, sapphire, beryl*
Language	*English, Irish*
Letter Of The Alphabet	*A, B, C*
Measurements (various)	*centimetre, hour, litre, kilo*
Metal	*steel, gold*
Military Rank/Title	*General, Corporal*
Mineral	*granite, alum*
Musical Instrument	*viola, kazoo*
Nationality	*Pole, Polish*
Numbers (cardinal & ordinal)	*one, two, twentieth, hundredth*
Personal Name	*Samuel, Isabella, Ted*
Season	*autumn, spring*
Title & Form Of Address	*Mr, Doctor, Queen*

Fig 4.11 Some template entry categories

Reading

Recommended reading

Clear 1987; Joffe and de Schryver 2004; Kilgarriff et al. 2004.

Further reading on related topics

Atkins and Levin 1995; Atkins, Levin, and Song 1997; Kilgarriff 2006b; Kilgarriff and Rundell 2002; Kilgarriff and Tugwell 2002; Landau 2001: 398–401; Walter and Harley 2002.

Websites

Corpus Query Software: http://www.lexically.net/ (WordSmith Tools concordancer); http://www.athel.com/mono.html , http://www.athel.com/para.html (Monoconc and Paraconc: monolingual and bilingual concordancers); http://www.sketchengine.co.uk/ (the Sketch Engine: a complete corpus query system with numerous ready-loaded corpora and corpus-building tools).

Dictionary Writing Software: http://www.idm.fr/products (the Dictionary Production System, or DPS, from IDM); http://tshwanedje.com/tshwanelex/ (the TshwaneLex dictionary compilation program): both excellent examples of software for producing dictionaries.

5

Linguistic theory meets lexicography

5.1 Preliminaries

By the nature of the work they do, lexicographers are applied linguists. Yet many people working in the field have no formal training in linguistics. Does this matter? Our experience as editorial managers suggests that good lexicographers operate to a large extent on the basis of instinct, sound judgment, and accumulated expertise. A grounding in linguistic theory is not a prerequisite for being a proficient lexicographer – still less a guarantee of success in the field. But there are certain basic linguistic concepts which are invaluable in preparing people to analyse data and to produce concise, accurate dictionary entries. And as we noted earlier (§1.2.2), an awareness of linguistic theory can help lexicographers to do their jobs more effectively and with greater confidence. In short, a good lexicographer will become a much better one with an understanding of relevant theoretical ideas. This chapter reviews those linguistic theories which we have found to have direct application to our work as dictionary planners and dictionary makers, and Figure 5.1 gives an outline of the issues we cover.

In §5.2 and §5.3 we give a brief account of some relationships between word meanings which have proved helpful in sensitizing lexicographers to

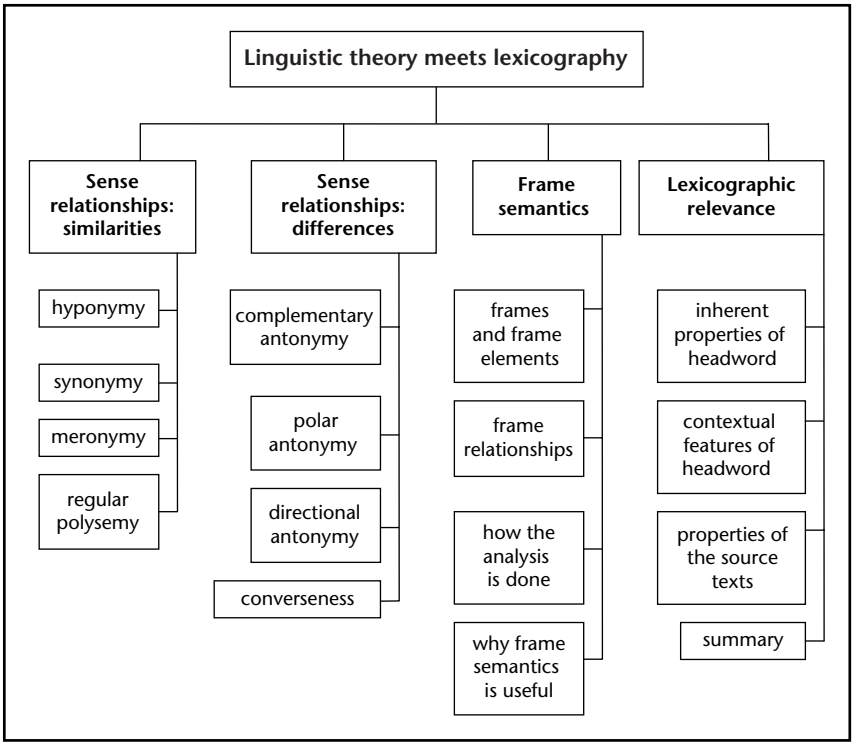


Fig 5.1 Contents of this chapter

the way words work.[1] It's important to remember that the paradigmatic relationships outlined in this chapter are all between *lexical units* (LUs),[2] that is to say word meanings and not words themselves. In the next section (§5.4) we discuss Fillmore's frame semantics, the theory that underlies the principle of lexicographic relevance set out in §5.5. That section also includes a brief overview of Mel'čuk's ideas on 'lexical functions'. Another theoretical field of great value for lexicography is the study of 'prototype' effects in language, an area associated especially with Eleanor Rosch and her colleagues, but also developed in the cognitive linguistics of George Lakoff, Ronald Langacker, and others. As we shall see later, their ideas

[1] This lexicographer's eye view of hyponymy, synonymy, meronymy and antonymy has been culled mainly from Lyons (1969: 400–435; 1977: 270–301; 1981: 136–151) and Cruse (1986: 84ff.; 2004: 143–171), two linguists whose accounts are eminently clear and inspiring. Thank you to both of them!

[2] This term is used as defined in Cruse (1986): essentially, a word or phrase in one of its meanings (see §6.1.1 for fuller explanation). An LU is the basic building block of dictionary entries.

have important implications for practical lexicography, especially in the areas of word sense disambiguation (§8.3.1) and definition writing (§10.5.3). Finally, and with regret, we admit that we haven't room for more than a mention of the WordNet Project, the large lexical database developed in Princeton NJ. WordNet, begun in 1986, is the brainchild of George Miller (see e.g. Miller et al. 1990). Under his guidance, English nouns, verbs, adjectives and adverbs were grouped into 'synsets': sets of near-synonyms each expressing a concept and linked together into a semantic network. The database, of some interest to lexicographers, has proved to be of great value to the natural-language-processing (NLP) community, and the idea has been carried forward into European languages in the EuroWordNet Project (Vossen 2004).

5.2 Sense relationships: similarities

This section summarizes different types of 'similarity' between LUs:

- those that share some semantic property or properties (hyponymy and synonymy)
- those that denote a part–whole relationship between objects in the real world (meronymy)
- those that allow similar metaphorical sense extensions (regular polysemy).

5.2.1 Hyponymy

The nodes of this hierarchy are the 'superordinate'[3] and the 'hyponym', as illustrated in the abbreviated classification tree in Figure 5.2, where the superordinates are first 'vertebrate' then 'mammal', 'canine', and so on down the column, and the rows show the dependent hyponyms of each, so that 'reptile', 'mammal', and 'amphibian' are all hyponyms of 'vertebrate', and so on. The terms in each row are cohyponyms. This relationship ('unilateral entailment') can be summarized as *if a hyponym then a superordinate*, so – working upwards in Figure 5.2 – we have the relationships set out in Figure 5.3.

Hyponymy is a relationship found in many nouns, in quite a number of verbs, and in some adjectives. Its major significance for lexicographers

[3] Also known as *hypernym*. Because of the similarity with *hyponym*, we prefer the term *superordinate*.

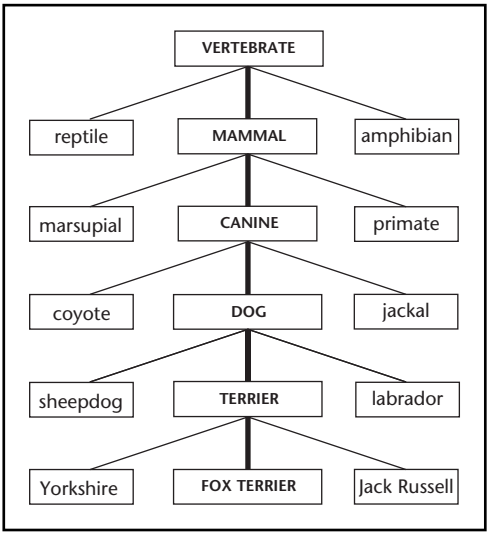


Fig 5.2 Superordinates, hyponyms, and cohyponyms

	hyponym		**superordinate**
if a	*fox terrier*	then a	*terrier*
if a	*terrier*	then a	*dog*
if a	*dog*	then a	*canine*
if a	*canine*	then a	*mammal*
if a	*mammal*	then a	*vertebrate*

Fig 5.3 Hyponyms and superordinates

is that the 'genus expression' (the 'central' word or words) in a definition should ideally be the superordinate of the headword (easy to say, but hard to do sometimes, especially in the case of adjectives and adverbs: cf. §10.5.1). Figure 5.4 gives some instances of this, in definitions of a noun, an adjective, and a verb.

These relationships can be expressed as in Figure 5.5. The dictionary relationships illustrated in Figure 5.5 represent the ideal – there are many, many words for which no precise genus expression exists.

➔ Hyponymy rule of thumb: X is a Y but Y is not only an X (*a terrier* is *a dog*).

Cohyponyms It is important to be able to distinguish between a pair of cohyponyms and a pair of synonyms (see §5.2.2). The definitions in

beret /…/ ▶ **noun** a round flattish cap of felt or cloth.

huge ▶ **adjective** (**huger**, **hugest**) extremely large; enormous: *a huge area* | *he made a huge difference to the team.*

punch[1] ▶ **verb** [with obj.] **1** strike with the fist: *he punched her in the face and ran off.* ▪ drive with a blow from the fist: *he punched the ball into his own goal.* **2** press (a button or key on a machine): *I punched the button to summon the lift.*

ODE-2 (2003)

beret /…/ noun [C] a round flat soft hat that fits tightly round the top of the head

huge /…/ adj ***
1 extremely large in size: *She arrived at the airport carrying two huge suitcases.*

punch[1] /…/ verb [T]*
1 to hit someone or something with your FIST (=closed hand), usually as hard as you can: *Two men punched him, knocking him to the ground. She punched her pillow angrily.* **1a.** to press a button or switch: *David punched a button on the television.*

MED-2 (2007)

Fig 5.4 Superordinates as genus expressions in definitions

	hyponym (headword)		**superordinate (genus expression)**
if a	*beret*	then a	*cap* or *hat*
if	*huge*	then	*large*
if	*punch*	then	*strike* or *hit* etc.

Fig 5.5 Headwords and genus expressions

Figure 5.6 show a fine display of cohyponyms. In many older dictionaries, definitions of adjectives were laden with cohyponyms, but current dictionaries make serious efforts to avoid this. The problem arises because finding a genus expression for a fuller definition is difficult: the hyponymy hierarchy is rarely found in adjectives, and consequently there is a real lack of superordinates (see also the discussion at §10.5.1).

➔ Cohyponyms rule of thumb: *X and Y are both Zs* (a *rose* and a *tulip* are both *flowers*).

5.2.2 Synonymy

Synonyms are words which have the same meaning. They bear a special relationship one to the other, which is defined in the formula shown in Figure 5.7; *pavement* in British English and *sidewalk* in American English denote the same stretch of walking space in city streets. The verbs *shut*

putrid /…/ *adj.* **1** decomposed, rotten. **2**. foul, noxious. **3.** corrupt. **4**. *slang* of poor quality; contemptible; very unpleasant.

strange /…/ *adj. & adv.* • *adj.* **1** unusual, peculiar, surprising, eccentric, novel. **2 a** (often foll. by *to*) unfamiliar, alien, foreign (*lost in a strange land; surrounded by strange faces; a taste strange to him*). **b** not one's own (*strange gods*). **3.** (foll. by *to*) (of a person) unaccustomed to; unfamiliar with. […]

nasty /…/ *adj. & n.* • *adj.* (**nastier, nastiest**) **1 a** highly unpleasant (*a nasty experience*). **b** annoying; objectionable (*the car has a nasty habit of breaking down*). **2**. difficult to negotiate; dangerous, serious (*a nasty fence; a nasty question; a nasty illness*). **3.** (of a person or animal) ill-natured, ill-tempered, spiteful; violent, offensive (*nasty to his mother; turns nasty when he's drunk*). **4**. (of the weather) foul, wet, stormy. **5 a** disgustingly dirty, filthy. **b** unpalatable; disagreeable (*nasty smell*). **c** (of a wound) septic. **6 a** obscene. **b** delighting in obscenity.

Concise Oxford Dictionary (1995)

Fig 5.6 Cohyponyms in use as definitions of adjectives

and *close* are synonymous in many uses. It is difficult to find convincing examples of synonyms, because true synonyms are extremely rare, if they exist at all.[4]

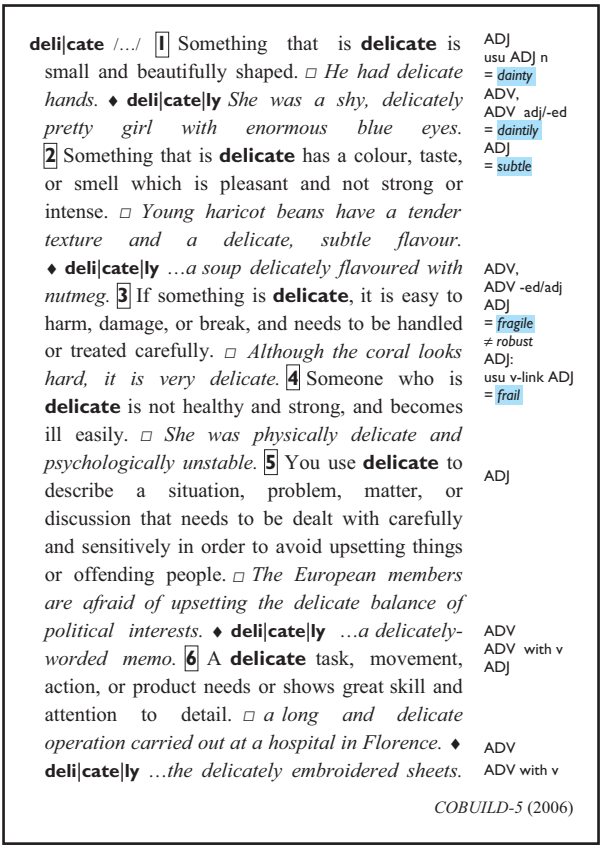
If	*X*	**then**	*Y*,	**if**	*Y*	**then**	*X*
If	*pavement*	then	*sidewalk*,	if	*sidewalk*	then	*pavement*
If	*shut*	then	*close*,	if	*close*	then	*shut*

Fig 5.7 The relationship of synonymy

The nearest you get is usually a pseudo-synonym, and 'synonyms' in dictionaries often turn out to be cohyponyms or superordinates. The relationship of synonymy should ideally hold between the headword and its target-language equivalent, but pure synonymy is rare across languages, except for the names of concrete objects which the two cultures share. Some learners' dictionaries include 'synonyms' as part of some entries: one of these is illustrated in Figure 5.8.

➔ Synonymy rule of thumb: X is Y and Y is X (*shut* is *close* and *close* is *shut*).

[4] In his *Preface* (1755), Johnson observes: 'Words are seldom exactly synonimous; a new term was not introduced, but because the former was thought inadequate: *names, therefore, have often many ideas, but few ideas have many names*.' (our italics)



deli|cate /…/ **1** Something that is **delicate** is small and beautifully shaped. □ *He had delicate hands.* ♦ **deli|cate|ly** *She was a shy, delicately pretty girl with enormous blue eyes.* **2** Something that is **delicate** has a colour, taste, or smell which is pleasant and not strong or intense. □ *Young haricot beans have a tender texture and a delicate, subtle flavour.* ♦ **deli|cate|ly** *…a soup delicately flavoured with nutmeg.* **3** If something is **delicate**, it is easy to harm, damage, or break, and needs to be handled or treated carefully. □ *Although the coral looks hard, it is very delicate.* **4** Someone who is **delicate** is not healthy and strong, and becomes ill easily. □ *She was physically delicate and psychologically unstable.* **5** You use **delicate** to describe a situation, problem, matter, or discussion that needs to be dealt with carefully and sensitively in order to avoid upsetting things or offending people. □ *The European members are afraid of upsetting the delicate balance of political interests.* ♦ **deli|cate|ly** *…a delicately-worded memo.* **6** A **delicate** task, movement, action, or product needs or shows great skill and attention to detail. □ *a long and delicate operation carried out at a hospital in Florence.* ♦ **deli|cate|ly** *…the delicately embroidered sheets.*

ADJ
usu ADJ n
= *dainty*
ADV,
ADV adj/-ed
= *daintily*
ADJ
= *subtle*

ADV,
ADV -ed/adj
ADJ
= *fragile*
≠ *robust*
ADJ:
usu v-link ADJ
= *frail*

ADJ

ADV
ADV with v
ADJ

ADV
ADV with v

COBUILD-5 (2006)

Fig 5.8 Synonyms in a *COBUILD* entry

5.2.3 Meronymy

Meronymy reflects the relationship of the part to the whole, and vice versa. 'X' is a meronym of 'Y' when you can say:

- X and the other parts of a Y, *or*
- the parts of a Y include the Xs.

See Cruse (1986: 168–179; 2004: 150–154) for a full discussion. The two formulae that quite adequately define this relationship for lexicographers are illustrated in Figure 5.9.

The role of meronyms in dictionary definitions is pretty constant: it's difficult to define the part without mentioning the whole. On the other hand, the part is only occasionally referred to in the definition of the whole, as may be seen from the pairs of definitions in Figure 5.10.

the parts of a	***Y***	**include the**	***Xs***
the parts of a	*wheel*	include the	*spokes*
the parts of a	*church*	include the	*nave*
the parts of a	*book*	include the	*pages*
the parts of a	*building*	include the	*rooms*
the parts of a	*hand*	include the	*fingers*
or			
the ***Xs***	**and the other parts of a**		***Y***
the *spokes*	and the other parts of a		*wheel*
the *nave*	and the other parts of a		*church*
the *pages*	and the other parts of a		*book*
the *rooms*	and the other parts of a		*building*
the *fingers*	and the other parts of a		*hand*

Fig 5.9 Formulae for meronyms

spoke	each of the bars or wire rods connecting the centre of a wheel to its outer edge
wheel	a circular object that revolves on an axle and is fixed below a vehicle or other object to enable it to move over the ground
nave	the central part of a church building, intended to accommodate most of the congregation
church	a building used for public Christian worship
page	one or both sides of a sheet of paper in a book, magazine, newspaper . . .
book	a written or printed work consisting of pages glued or sewn together along one side and bound in covers
room	a part or division of a building enclosed by walls, floor, and ceiling
building	a structure with a roof and walls, such as a house or factory
finger	each of the four slender jointed parts attached to either hand
hand	the end part of a person's arm beyond the wrist, including the palm, fingers, and thumb

Fig 5.10 Meronyms in *ODE-2* (2003) definitions

➔ Meronymy rule of thumb: X and the other parts of Y.

Quasi-meronymy Quasi-meronymy[5] reflects the relationship of the member to the group or class of people, or collection of objects. This is a rather loose relationship: because of this it's difficult to word a formula appropriate to all the varied LUs it should cover. The formula you can use to remember quasi-meronymy is shown in Figure 5.11.

Pairs of quasi-meronyms are less likely than meronyms to appear in their partners' definitions, but this does occur, as may be seen from the examples in Figure 5.12.

[5] Cruse's term (Cruse 1986): Figure 5.11 uses the words he chose to illustrate this concept.

An	*X*	belongs to/in a	*Y*
A	*tribesman*	belongs to a	*tribe*
A	*juror*	belongs to a	*jury*
A	*vicar*	belongs to the	*clergy*
A	*duchess*	belongs to the	*aristocracy*
A	*tree*	belongs in a	*forest*
A	*book*	belongs in a	*library*
A	*playing card*	belongs in a	*pack*
The	*workers*	belong to the	*proletariat*

Fig 5.11 Formula for quasi-meronyms

tribesman	a man who is a member of a tribe
tribe	a social group consisting of people of the same race who have the same beliefs, customs, language etc, and usually live in one particular area ruled by their leader
juror	a member of a jury
jury	a group of 12 ordinary people who listen to the details of a case in court and decide whether someone is guilty or not
duchess	a woman with the highest social rank outside the royal family, or the wife of a duke
aristocracy	the people in the highest social class, who traditionally have a lot of land, money, and power
vicar	a priest in the Church of England who is in charge of a church in a particular area
clergy	the official leaders of religious activities in organized religions, such as priests, rabbis, and mullahs
tree	a very tall plant that has branches and leaves, and lives for many years
forest	a large area of land that is covered with trees
book	a set of printed pages that are held together in a cover so that you can read them
library	a room or building containing books that can be looked at or borrowed
playing card	a small piece of thick stiff paper with numbers and signs or pictures on one side. There are 52 cards in a set
pack	a complete set of playing cards
workers	the members of the working class
proletariat	the class of workers who own no property and work for wages, especially in factories, building things etc.

Fig 5.12 Quasi-meronyms in *LDOCE-4* (2003) definitions

CONTAINER	*Put the can in the recycling bin.*
CONTENTS	*Did he eat the whole can?*
CONTAINER	*He dropped the glass and it broke.*
CONTENTS	*She drank six glasses that evening.*

Fig 5.13 Examples of CONTAINER–CONTENTS regular polysemy

➔ Quasi-meronymy rule of thumb: X belongs to / in a Y.

5.2.4 Regular polysemy

Some polysemous words have a particular relationship with others in their 'lexical set',[6] in that several of their meanings seem to parallel each other. Certain specific semantic components result in sets of words behaving lexicographically in a very similar way. This is known as 'regular polysemy'.[7] One of the best-known examples of regular polysemy ('container→ contents') is illustrated in Figure 5.13. You could slot other 'container' words into the two sentences with equal success, for instance *cup*, *bowl*, *packet*, *bottle*, etc.

In the examples in Figure 5.13, the semantic component CONTAINER in *can* and *glass* results in each of these words having parallel sets of LUs:

- the object itself, and
- its contents.

Such inter-word relationships are of immediate interest to lexicographers: once you've worked out the entry for *can* then when you come to *glass* you can use the shared LUs in the *can* entry as a model.[8] When you're planning the editorial work in a dictionary project, it's obviously a help to the team if you can list the major instances of regular polysemy, either by producing template entries or simply by issuing lists of headwords related in this way.

Apresjan (1973) described the semantic components which gave rise to this phenomenon in Russian words, and most of these components function

[6] This term denotes a group of words similar in meaning that belong to the same wordclass; for instance, the days of the week form a lexical set, as do names of liquids, motion verbs, colour terms, etc.

[7] Also called by different linguists *systematic polysemy*, *semantic transfer*, *regular meaning shift*, *semi-productive polysemy*, and *lexical implication rules*.

[8] That is the thinking behind the idea of 'template entries', discussed in §4.5.

equally well in English and other related languages.[9] However, English morphology encourages a wider range of regular polysemy than is found in languages that have specific forms for verbs and nouns, and so we include under this umbrella term of *regular polysemy* any combination of word-classes. Figure 5.14 shows a selection of instances of this phenomenon.[10]

(ANIMAL (etc.)	*nc*	There's a **squirrel**.
(ITS MEAT	*nu*	We don't eat **squirrel**.
(ANIMAL (etc.)	*nc*	There's a **mink** near the river.
(ITS SKIN OR FUR	*modif*	She wore a **mink** coat.
(CONTAINER	*nc*	He had his hands in his **pockets**.
(PUT INTO IT	*vt*	He **pocketed** the change and ran off.
(CREATOR	*n pr*	**Shakespeare** wrote plays.
(CREATED OBJECT	*nc*	It's in **Shakespeare** somewhere.
(MASS	*nu*	She doesn't drink **coffee**.
(UNIT	*nc*	Three **coffees** please.
(MATERIAL	*nu*	That looks like **silver**.
(MADE OF IT	*modif*	It's a **silver** bracelet.
(MUSICAL INSTRUMENT	*nc*	Do you play the **cello**?
(PERSON PLAYING IT	*nc*	The **cellos** came in late.
(OBJECT: TOKEN	*nc*	I like your **jacket**.
(OBJECT: TYPE	*nc*	They've got your **jacket** in the window.
(STATE'S CAPITAL	*n pr*	Have you been to **Rome**?
(STATE'S GOVERNMENT	*n pr*	**Rome** denied this.
(TREE	*nc*	She stood by a tall **pine**.
(ITS WOOD	*nu*	The desk was made of **pine**.
(UTENSIL	*nc*	I haven't got a **fork**.
(DO SOMETHING WITH IT	*vt*	He **forked** the peas into his mouth.
(WEAPON	*nc*	They all carry **knives**.
(ATTACK WITH IT	*vt*	He got **knifed** during a robbery.

Fig 5.14 Some classes of regular polysemy in English

Wordclass abbreviations as follows: *modif* modifier noun; *nc* noun countable; *nu* noun uncountable; *n pr* proper noun; *vi* verb intransitive; *vt* verb transitive.

Verb alternations The alternations recorded for various verb classes by Levin (1993) are – from the lexicographic point of view – very similar to Apresjan's classes of regular polysemy, in that they link specific verb behaviour to a specific semantic class (or component). This allows us to capitalize on work done on one entry when you come to the entry for the

[9] We don't know enough about the languages of the world to say whether regular polysemy is to be found in all the great language families.

[10] Our database currently stands at over 100 classes of regular polysemy.

next member of the class. A few of Levin's many alternations are exemplified in Figure 5.15.

alternation		example 1	example 2
'SPRAY-LOAD'	(X on Y	spray paint on the truck	load wood on the truck
	(Y with X	spray the truck with paint	load the truck with wood
'SEARCH'	(X for Y	search the woods for him	fished the lake for trout
	(for Y in X	search for him in the woods	fished for trout in the lake
'DATIVE'	(X to Y	gave the book to Peter	offer the wine to Peter
	(YX	gave Peter the book	offer Peter the wine
'BENEFACTIVE'	(X for Y	sew a dress for the child	cook a meal for him
	(YX	sew the child a dress	cook him a meal

Fig 5.15 Some verb alternations in English

5.3 Sense relationships: differences

This section summarizes relationships between LUs that are in some way opposite in meaning.[11] Synonymy is simple (if rare); antonymy is more complex. The first three types outlined here (complementary, polar, and directional antonymy) are what most people think of as 'opposites', but they are subtly different. However, they all function well in dictionaries as antonyms of the headword, and are useful in definitions. Complementary and polar antonyms are especially useful in definitions of adjectives (as may be seen from the entries in Figure 5.18). Just as hyponymy holds more often between nouns, so antonymy 'belongs' more to adjectives.

5.3.1 Complementary antonymy

This relationship is sometimes called 'contradiction'. The rather small group of adjectives with complementary antonyms have no comparative or superlative forms, since the state they denote is not relative: you can't be *slightly alive* or *rather dead.* The formula defining complementarity is given in Figure 5.16.

➔ Complementary antonym rule of thumb: If it isn't X then it must be Y, and vice versa.

[11] The terminology and most of the examples here come from Cruse (1986).

If	***X***	**then not**	***Y***	**and if**	***Y***	**then not**	***X***
If	*alive*	then not	*dead*	and if	*dead*	then not	*alive*
If	*blind*	then not	*sighted*	and if	*sighted*	then not	*blind*
If	*hit*	then not	*miss*	and if	*miss*	then not	*hit*

Fig 5.16 The relationship of complementarity

5.3.2 Polar antonymy

This relationship is similar to, but more complex than, complementarity. *If X then not Y and if Y then not X* holds good here as well, but is not enough. There is a gradient between X and Y in polar antonymy. X and Y are at the poles of this gradient, but in between there is an indeterminate area, where *more X* and *less Y* are found. Something is not necessarily *good* because it is not *bad*, a surface can be *smoother* or *rougher* than another surface, and so on. The formulae defining polar antonymy are given in Figure 5.17.

If	***X***	**then not**	***Y***	**and if**	***Y***	**then not**	***X***
If	*good*	then not	*bad*	and if	*bad*	then not	*good*
If	*light*	then not	*dark*	and if	*dark*	then not	*light*
If	*poor*	then not	*rich*	and if	*rich*	then not	*poor*
If	*smooth*	then not	*rough*	and if	*rough*	then not	*smooth*

and

If more	***X***	**then less**	***Y***	**and if more**	***Y***	**then less**	***X***
If more	*good*	then less	*bad*	and if more	*bad*	then less	*good*
If more	*light*	then less	*dark*	and if more	*dark*	then less	*light*
If more	*poor*	then less	*rich*	and if more	*rich*	then less	*poor*
If more	*smooth*	then less	*rough*	and if more	*rough*	then less	*smooth*

Fig 5.17 The relationship of polar antonymy

➔ Polar antonymy rule of thumb: If it's X then it can't be Y, and vice versa, but it can be somewhere in between.

Complementary antonyms and the more common polar antonyms are useful in definitions of adjectives and adverbs, as may be seen from the entry shown in Figure 5.18.

5.3.3 Directional antonymy

Directional antonyms include various subtypes: some denote contrary movement or position, for instance, pairs of words representing opposing 'poles' along a shared axis; in other cases, the shared axis is one of the many Lakoffian extensions of spatial concepts (see Lakoff and Johnson 1980, esp. 14–21, 56–60). Cruse (1986) offers a much finer-grained analysis of this

dead ▶ **adjective 1** no longer alive: *a dead body* | [as complement] *he was shot dead by terrorists* | [as plural noun] (**the dead**) *there was no time to bury the dead with decency*. ■ (of a part of the body) having lost sensation; numb, lacking emotion, sympathy, or sensitivity: *a cold, dead voice*. ■ no longer current, relevant, or important: *pollution had become a dead issue*. ■ devoid of living things: *a dead planet*. ■ (of a place or time) characterized by a lack of activity or excitement: *Brussels isn't dead after dark, if you know where to look*. ■ (of money) not financially productive. ■ (of sound) without resonance; dull. ■ (of a colour) not glossy or bright. ■ (of a piece of equipment) no longer functioning: *the phone had gone dead*. ■ (of an electric circuit or conductor) carrying or transmitting no current: *the batteries are dead*. ■ no longer alight: *the fire had been dead for some days*. […]

ODE-2 (2003)

Fig 5.18 Complementary and polar antonyms used in definitions

type of antonymy and in Figure 5.19 the pairs of directional antonyms are grouped according to his subclassification ('directions', 'counterparts', 'antipodals', and 'reversives'); the names are not important but it is clear from the sets of words illustrating the groups that the relationships are valid.

DIRECTIONS			COUNTERPARTS		
north	≠	*south*	*male*	≠	*female*
up	≠	*down*	*convex*	≠	*concave*
forward	≠	*backward*	*yin*	≠	*yang*
ANTIPODALS			**REVERSIVES**		
top	≠	*bottom*	*appear*	≠	*disappear*
zenith	≠	*nadir*	*tie*	≠	*untie*
start	≠	*finish*	*pack*	≠	*unpack*
cradle	≠	*grave*	*widen*	≠	*narrow*
attic	≠	*cellar*	*heat*	≠	*cool*

Fig 5.19 Directional antonyms

Directional antonyms are solid material for lexicographers who need to add antonyms to a dictionary entry: here are nouns, verbs, adjectives, and adverbs in abundance, their number greatly increased by the negating prefixes *un-*, *dis-*, and so on.

5.3.4 Converseness

Converseness holds between pairs of words which have a certain semantic symmetry, so that although not antonyms one of the pair is felt in some

way to be linked by 'oppositeness' to the other. The formula which defines converseness is shown in Figure 5.20, where examples are given of converse pairs of nouns, verbs, and prepositions. There is little direct application of converse pairs in dictionaries, but if a word is difficult to define a look at its converse's definition can be helpful. And we believe it's useful for lexicographers to understand this relationship, ever since the day we happened upon a draft entry for *butler* in which *cook* was offered as its antonym.

NOUNS

If	**A**	**is B's**	***X***	**then**	**B**	**is A's**	***Y***
If	A	is B's	*husband*	then	B	is A's	*wife*
If	A	is B's	*teacher*	then	B	is A's	*student*
If	A	is B's	*doctor*	then	B	is A's	*patient*
If	A	is B's	*child*	then	B	is A's	*parent*

VERBS

If	**A**	***Xs***	**to B**	**then B**	***Ys***	**from**	**A**
If	A	*gives*	to B	then B	*receives*	from	A
If	A	*sells*	to B	then B	*buys*	from	A

PREPOSITIONS

If	**A**	**is**	***X***	**B then B**	**is**	***Y***	**A**
If	A	is	*below*	B then B	is	*above*	A
If	A	is	*behind*	B then B	is	*in front of*	A
If	A	is	*before*	B then B	is	*after*	A

Fig 5.20 Converse pairs of nouns, verbs, and prepositions

5.4 Frame semantics

> The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described.
>
> (Fillmore 1995)

This section offers a brief introduction to frame semantics: the application of this theory to practical lexicography results in the approach to lexicographic relevance discussed in §5.5, which helps lexicographers to identify useful facts in corpus texts.

Frame semantics is essentially the brainchild of Charles J. Fillmore, a linguist from the University of California and the International Computer Science Institute (ICSI), Berkeley, California. This complex theory, summarized in Fillmore (2005), describes words, their various meanings, and how these are combined with others to form the utterances and sentences

of a language. In this section we introduce only the absolute basics, the 'currency units' of a frame semantics analysis of text. Information about other aspects of the theory is to be found on the website of the FrameNet project (http://framenet.icsi.berkeley.edu/). A comprehensive summary of this work is to be found in *International Journal of Lexicography*, 16.3 (2003).

This project, of considerable importance to professional lexicographers, is based in ICSI and led by Charles Fillmore. Its aim is to analyse and record, for each sense of a word or phrase, the full range of its semantic and syntactic relations. To do this, they have devised a suite of codes denoting semantic roles ('frame elements') and grammatical relationships, which allow them to document in detail the corpus contexts in which a word is found. The work is computer-assisted, in that the annotation of example sentences is done semi-automatically, and the resultant database is automatically displayed in a number of different ways for human scrutiny, and is computer-searchable. At the time of writing, the database – freely available – continues to grow, and is now in use by hundreds of researchers, teachers, and students around the world. Similar analyses of German, Spanish, and Japanese are in progress, closely linked to the work in Berkeley. In our experience, the frame semantics approach to word behaviour is the most helpful and appropriate approach to corpus data, ensuring as it does that the analysis is correctly carried out, and no vital fact is overlooked.

5.4.1 What are frames and frame elements?

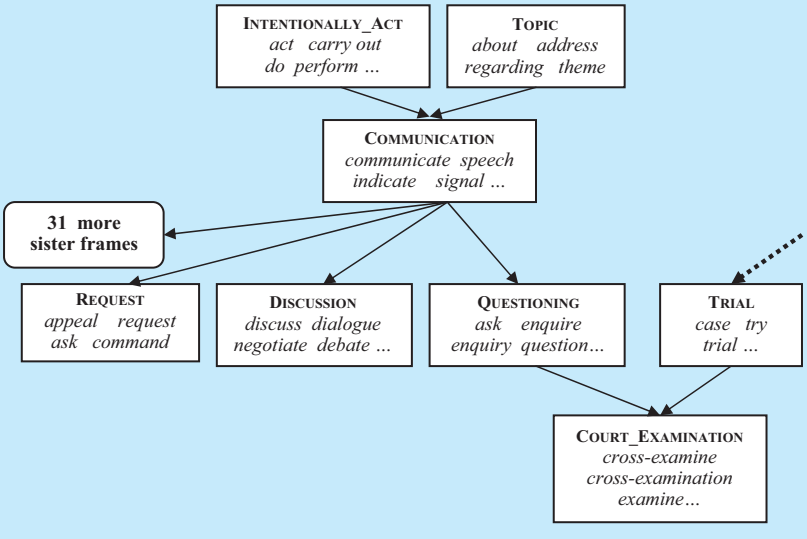
Frame semantics describes the meanings of words and phrases (lexical units, or LUs) in terms of the *frame* to which they belong (or, in frame semantics terminology, which they evoke) and the *contexts* in which these LUs are found.

- A semantic frame is a schematic representation of a situation type (e.g. speaking, eating, judging, moving, comparing, etc. – activities and situations which make up our everyday life) together with a list of the typical participants, props, and concepts that are to be found in such a situation; these are the semantic roles, or 'frame elements' (FEs).
- The context, in a frame semantics analysis, is normally the phrase or clause, and maximally the sentence, in which the target word appears in corpus data.

Box 5.1 How are frames related to each other?

Frame semantics envisages a network of interrelated frames (a 'frame net') which accounts for word meaning. Frames are related one to another by principles of inheritance, the 'child' frame (for instance, REQUEST) being more specific than the 'parent' frame (for instance, COMMUNICATION). Frame inheritance is illustrated in the diagram below, where the frame names are in capitals, and some of the lemmas which evoke each frame are shown in italics.

A frame can have no parents, or one, or several. Similarly, a frame can have no children, or one, or several. Thus, in the figure below, the REQUEST frame is the child of COMMUNICATION, which itself is the child of INTENTIONALLY_ACT and TOPIC. REQUEST has (so far) no children, but its sibling QUESTIONING has a child COURT_EXAMINATION, whose other parent is TRIAL. (Clearly, verbs like *cross-examine* evoke simultaneously the QUESTIONING and the TRIAL frames.) Frame inheritance also involves the inter-mapping of the elements of the child and parent frames. More details are to be found in the 'FrameGrapher' on the FrameNet website.



Partial diagram of frame inheritance relationships

We can communicate in language because the words and phrases we use evoke their frames in our minds, so that we share an interpretation of what is said or written. For instance, a friend says to you, *Jo asked her brother to help her.* In our own personal experience the situation in which someone makes a request normally contains certain elements. The vocabulary and syntax of its context lets you identify the LU *ask* in that sentence

as belonging to the REQUEST frame. At this point you instantly – and of course subconsciously – expect to find mentioned both the person who is doing the requesting (*Jo*) and the person being asked (*her brother*), as well as what that person is asked to do (*to help her*). Your knowledge of English leads you to interpret the subject of the verb as the 'requester', its object as the person being asked, and its infinitival complement as the requested action. The frame elements in our example sentence are:

- the **Speaker** (*Jo*)
- the **Addressee** (*her brother*)
- the **Message** (*to help her*).

These elements of the REQUEST frame are used to describe the behaviour of the other words in that frame: for instance, verbs such as *order*, *appeal*, *command*, *suggest*, *beg*, and nouns such as *order*, *appeal*, *command*, *suggestion*, and of course *request* itself.

The lexical analysis being carried out by the FrameNet project is far from complete, and the network of frames and their relationships changes as the research progresses. The way in which frames are related to each other is not of such immediate relevance to the work of lexicographers, fixated as we are on the words of the language. Box 5.1 summarizes the inter-frame relationships. Of more immediate interest to working lexicographers is the corpus analysis aspect of the project.

5.4.2 How is the analysis done?

There are several distinct steps in the analysis process.

- First, the frame is defined, and its 'core' elements[12] named and described.
- Next, a list is made of as many words as can be found which in one of their senses evoke that frame.
- Then, for each sense, or LU, a set of corpus sentences is extracted, in which the word is used in the particular sense.
- Each sentence is annotated by marking off any section which instantiates an FE, and by recording, for each FE thus identified:

[12] These are the FEs essential to the frame itself. Language learners must know how these FEs are expressed grammatically, or they cannot use the word correctly. FrameNet also recognizes 'peripheral' FEs (those common to whole sets of frames, such as LOCATION, DURATION, FREQUENCY, etc.); these are not covered in this summary.

– its 'phrase type' (noun phrase, verb phrase, adjective phrase, and so on);
– its 'grammatical function' (subject, object, complement, etc.) within the clause containing the target word.

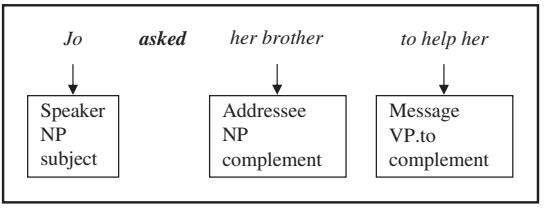


Fig 5.21 Annotation of a sentence where *ask* evokes the REQUEST frame

The annotation of our example sentence is set out in Figure 5.21, using the elements of the REQUEST frame. The keyword is of course *ask* in the REQUEST frame. Each set of threefold information (frame element, its phrase type, and its grammatical function) is called a 'valency group'; there are three of these in Figure 5.21. In the first, the FE 'Speaker' is realized by the noun phrase ('NP') *Jo*, functioning as the 'subject' of the target word *ask*. The complete set of valency groups drawn from a single sentence is called a 'valency pattern'; the valency pattern of the *ask* sentence is shown in Figure 5.22.

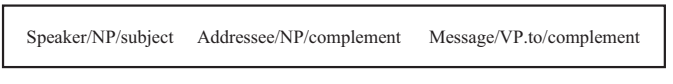


Fig 5.22 One valency pattern of *ask* in REQUEST frame

The same frame element can be expressed in different ways. The sentence *Jo asked him if he would help her* contains the same three FEs, in a different valency pattern, shown in Figure 5.23, where the Message is expressed by the *if*-clause. The complete set of valency patterns found in the corpus for an LU is the 'valency description' of that LU, and formalizes, for human and computer use, the syntactic and combinatory properties of that LU.

Fig 5.23 Another valency pattern of *ask* in REQUEST frame

5.4.3 Why is this useful for lexicographers?

The facts in the valency description are the most important facts that the lexicographer needs to be aware of when writing the dictionary entry, as is clearly seen from the *MED* entry in Figure 5.24. (Pedagogical dictionaries are careful to set out in detail the constructions that language-learners need if they are to use the headword correctly, and the list in the *ask* entry is quite comprehensive.) The ways in which the frame elements are expressed are what language-users need to know. The words used to express them are the important collocates of the keyword of the entry being written.

ask /…/ verb ***
1 [I/T] to speak or write to someone in order to get information from them: *I wondered who had given her the ring but was afraid to ask.* […] ♦ **ask (sb) why/how/whether** etc: *She asked me how I knew about it.* ♦ **ask (someone) about something**: *Did you ask about the money?* […]
2 [I/T] to speak or write to someone because you want them to give you something: *If you need any help, just ask.* ♦ **ask (sb) for sth**: *The children were asking for drinks.* ♦ **ask sb's permission/advice/opinion** etc: *I think we'd better ask your mum's opinion first.* […]
3 [I/T] to expect someone to do something or give you something: ♦ **ask sth (for sth):** *It's a nice house, but they're asking over half a million pounds.* […] ♦ **ask sb (not) to do sth**: *We ask guests not to smoke in the hotel.* […]
4 [I/T] to say that you want something to happen, or that you want someone else to do something: ♦ **ask sb (not) to do sth**: *Then the computer will ask you to restart it. He asked us to join him.* ♦ **ask to do something**: *I asked to see the manager.* ♦ **ask (not) to be**: *The writer has asked not to be named.* ♦ **ask that sb (should) do sth**: *The committee has asked that this scheme be stopped for now.*
5 [T] to invite someone to do something or go somewhere with you: ♦ **ask sb to sth**: *How many people have you asked to the party?* ♦ **ask sb for sth**: *We should ask them for a meal sometime.* […] ♦ **ask sb to do sth**: *They asked me to stay the night.*

MED-2 (2007)

Fig 5.24 Entry for *ask* with expressions of frame elements (the essential constructions)

The frame semantics approach, grounded in a coherent theory, offers the possibility of a more systematic, less subjective way of analysing corpus data, and gives us confidence that all relevant features are being captured. How this approach is translated into corpus analysis for practical dictionary-writing is set out in §5.5.

5.5 Lexicographic relevance

When you first look at the wealth of concordances the corpus offers for a word, you think 'How can I decide what of all this I should record in the database?' Another way of putting that is 'What is lexicographically relevant?' Lexicographic relevance is at the heart of all good lexicography, whether mono- or bilingual. For many excellent lexicographers this underlying theory is never made explicit: their intuition tells them what's worth saying about the headword, once they've scrutinized the corpus evidence. However, personal intuition is difficult to transmit to an apprentice, and notoriously unreliable. 'Watch what I do and see if you can get the hang of it' is a teaching method that is liable to destroy learners' self-confidence before they turn into good lexicographers (or go mad, or switch careers, whichever comes first). The practical application of frame semantics to lexicography is the focus of Atkins (1995, 1996), Fillmore and Atkins (1994, 2000), Atkins, Fillmore, and Johnson (2003), and Atkins, Rundell, and Sato (2003).

In what follows, we consider lexicographic relevance from the standpoint of Fillmore's frame semantics. There is, however, one other approach that is equally thorough and valid, that of the linguist Igor Mel'čuk: his theory of 'lexical functions' is summarized in Box 5.2 and deserves serious study by lexicographers working on corpus data. Note that Mel'čuk's use of the word *collocation* is slightly different from the way we use it in the rest of this volume.

Three types of information are relevant to making a lexicographic record of a word:

(1) what we know, as native speakers, about the headword (its inherent properties)
(2) what we learn from its use in corpora and elsewhere (its contextual features)
(3) what we know about where the citations came from (the properties of the source texts).

(1) and (3) above are fairly obvious; the second can create a lot of problems. We consider each of these in the sections that follow, using the verb *argue* as a case study.

➔ It's important to remember that 'lexicographic relevance' relates to what is relevant to an LU, and not to a lemma, i.e. the focus is the headword in one of its senses, not the whole word.

Box 5.2 Describing collocations: Mel'čuk's lexical functions

Collocation has been a central preoccupation of the linguist Igor Mel'čuk since the mid-1960s. His theoretical insights, and the typology of collocations which they have given rise to, have direct and practical applications to the description of language in dictionaries. For Mel'čuk, *phrasemes* – or (semi-)prefabricated word combinations of various types – represent 'the numerically predominant lexical unit' in any language. And *collocations* form the largest subset of 'the phraseme inventory' (Mel'čuk 1998: 24). Mel'čuk exhaustively catalogues every type of word combination, but the aspect of his work of most interest to lexicographers is his system of *lexical functions*.

A lexical function characterizes a specific type of word combination. Consider, for example, the sentence *Wilson had committed a serious crime*, which illustrates two common lexical functions:

- If you want to say that someone 'does' or 'performs' a *crime*, which verb or verbs usually instantiate this semantic relationship? (*commit* occurs most frequently in this slot, with *perpetrate* an occasional alternative)
- If you want to indicate that a crime is 'big' or 'major', which adjectives would you usually use? (*serious* is by far the most common choice here)

In Mel'čuk's typology, these two combinations (*commit* + *crime*, *serious* + *crime*) exemplify lexical functions which he labels, respectively, 'Oper' and 'Magn'. The 'Oper' function broadly describes 'operating verbs': the verbs you use in order to indicate performing an action. Thus we say '*carry out* a search', '*create* a diversion', '*conduct* a survey', and so on. The 'Magn' function deals with 'intensification', and is realized in combinations like '*highly* improbable', 'as skinny *as a rake*', and '*deep* commitment'. These are just two of dozens of lexical functions, which collectively describe every conceivable type of combination. For Mel'čuk, the lexical functions are an essential component of what he calls an 'Explanatory Combinatorial Dictionary' (ECD) – a theoretical lexicon which aims to catalogue, in a formalized way, the semantic and combinatorial features of every lexical unit, so that 'a lexical entry includes whatever a native speaker knows about the LU in question' (ibid.: 50). Mel'čuk and his co-workers have produced ECDs for Russian and French. For us mere mortals, with our more modest goal of producing dictionaries for everyday use, this complex and ambitious system may seem too daunting to be of practical use. But there is much to learn here, and it is quite feasible to distil – from Mel'čuk's comprehensive inventory – a subset of the most common lexical functions. These will form an invaluable checklist for lexicographers, ensuring that important collocational patterns are recorded in the database. The table below illustrates this approach in terms of some common types of combination involving *nouns*.

(*Cont.*)

Box 5.2 (*Continued*)

Combination	Lexical function	Examples
verb + noun object	doing, making	*commit* a crime, *do* your homework, *perform* an operation
	making something start	*launch* an inquiry, *acquire* a habit, *impose* sanctions
	making something end	*lift* sanctions, *stamp out* abuse, *break* a habit
noun subject + verb	what the noun typically does	rumours *circulate*, storms *rage*, heart *beats*
	how something starts	war *breaks out*, an impression *forms*
	how something ends	storms *abate/die down*, sounds *die away*, a meeting *closes*
adjective + noun	a big or major example	a *serious* accident, an *unmitigated* disaster, *intense/fierce* competition
	a small or minor example	*gentle* exercise, a *minor* injury, a *modest* improvement

A checklist of lexical functions for noun headwords

➔ It's a good idea to identify the most common forms of combination for each of the main wordclasses and produce checklists like this. This kind of information is very valuable, especially for pedagogical dictionaries and for Style Guide development.

5.5.1 Inherent properties of the headword

This is the knowledge of our language that we all bring to analysing the corpus data and writing the dictionary entry. The properties of the headword that principally concern us can be summarized very briefly:

- its wordclass: *argue* is a verb, and most of the other properties depend on this classification
- its wordforms: its inflections are *argue, argues, argued, arguing*
- its grammatical behaviour: constructions like *argue with someone about something*; the reciprocal alternation (*A argues with B* / *A and B argue*, etc.), and so on
- its semantics: it is polysemous (we can identify several meanings), but essentially it is a verb of communication.

This sort of knowledge, while valid, is not under our control, nor is it very clear-cut, and it is certainly not objective – you only have to listen to two native speakers arguing about how a word is used to realize that. In corpus lexicography we use our inherent knowledge of the headword rather to help us discover the really useful facts in the corpus, and make sure the entry is comprehensive and the examples are pleasing to our native speakers' ears. Our knowledge of the headword's inherent properties serves as quality control during our work on corpus data, as we discover and record its contextual features in each of its LUs.

5.5.2 Contextual features of the headword

An understanding of lexicographic relevance helps you identify in a corpus sentence all the *essential* components of the headword's context, all the facts that you need to take into account when writing any entry for that word.

1	The congestion on our roads	argues	that a serious vehicle tax should be levied.
2	'You'll stop	arguing	and do as you're damned well told!'
3	We spent most of our time in cafes,	arguing	and holding hands.
4	He was penalised for joking and	arguing	disruptively yesterday.
5	These features	argue	for a local origin.
6	Margaret Mead	argued	for a nurture perspective on behaviour.
7	There was a lot of	arguing	going on between Mum and Dad.
8	This can be seen, they	argue,	in many forms of state intervention.
9	The teachers and medics were	arguing	about who has which square inch of my time.
10	Dr Wilson	argues	that if ants were to disappear, most of the . . .
11	Richard Dawkins has	argued	that it is their genes that survive.
12	Like Pareto, Burnham	argued	that Marxism was a self-serving ideology.
13	This lack is a key factor	arguing	against the existence of such a relationship.
14	Don't try to	argue	him out of it now – it's too late.
15	The platoon commander was	arguing	with a gang of Christian Phalangists.

Fig 5.25 KWIC concordances for *argue*

5.5.2.1 *Case study: argue* Scanning the concordances shown in Figure 5.25, you begin to feel your way around the word.[13] You argue *about* something (one sense here – 'quarrel') but you can also argue *for* and *against* something – a second sense, surely, meaning something like 'make a case, maintain'. Then we find *the congestion on our roads argues that a serious tax should be levied*, and add a third sense to our armoury, that of 'be evidence of, indicate'; finally *don't try to argue him out of it now* reveals a fourth sense,

[13] See Chapter 8 (esp. §8.4, §8.5) for a detailed discussion about finding word senses.

'persuade'. So we may say that the headword *argue* contains four LUs; each represents a distinct *sense* of the headword, and we've used our knowledge of its inherent properties to identify these senses, or – in frame semantics terms – to identify the specific frame which each LU evokes. The LUs are:

LU-1 the sense of 'quarrel, dispute' (*don't argue with her*), i.e. *argue* in the COMMUNICATION frame (cf. lines 2, 3, 4, 7, 9, and 15 in Figure 5.25)

LU-2 the sense of 'maintain, make a case for' (*he argued for a change in tactics*), in the REASONING frame (cf. lines 6, 8, 10, 11, and 12 in Figure 5.25)

LU-3 the sense of 'indicate' (*this argues a lack of support*), in the EVIDENCE frame (cf. lines 1, 5, and 13 in Figure 5.25)

LU-4 the sense of 'persuade' (*she argued them out of going*), in the PERSUASION frame (cf. line 14 in Figure 5.25).

The lexicographically relevant components in each *argue* sentence are those which express the various frame elements. They will differ for each LU, since the frame elements depend on the frame the LU belongs to. In the outline of frame semantics in §5.4, the examples were drawn from the COMMUNICATION frame (and in that instance the key verb was *ask*). We'll stay with that frame, and focus now on the LU-1, '*argue*-quarrel', as in *Sam was arguing with his brother about the money*.

In an argument there are three principal frame elements or semantic roles:

- Participant-1, i.e. one of the people arguing (*Sam* in our example above)
- Participant-2, the other arguer (*his brother*)
- Topic, what they are arguing about (*the money*).

Starting from our understanding of what it means to be in an argument, and noting the principal semantic roles involved, we look for any or all of these in the context of the verb *argue*. On this basis, the sentence can be analysed as shown in Figure 5.26.

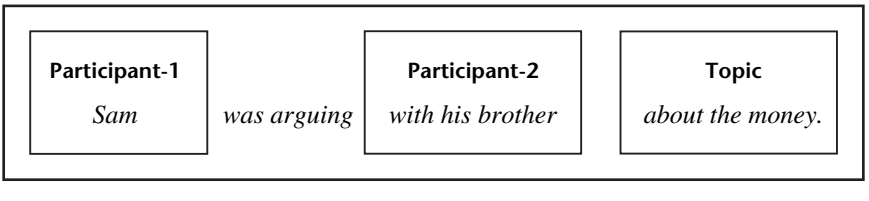


Fig 5.26 Frame elements in this context of *argue*

These three sentence components linked to the frame elements flag up the information you need to record from this particular context for your headword *argue*. First, however, you have to add the grammar to this semantic analysis. Two types of grammatical information are noted for each of the frame elements:

- its phrase type, i.e. the type of phrase that expresses it
- the role which that phrase plays within its clause (its grammatical function).

These are shown in Figure 5.27.

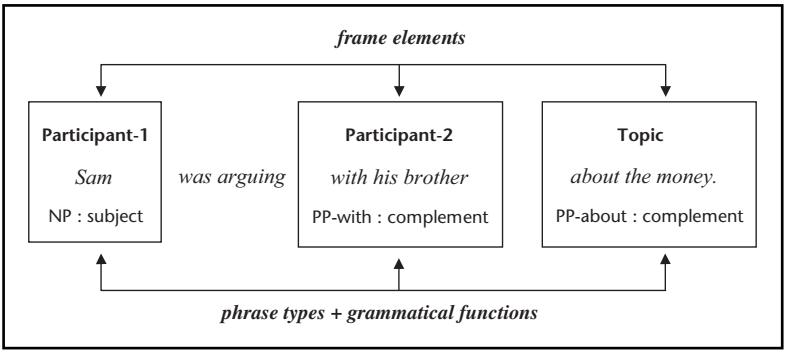


Fig 5.27 Threefold description of the relevant components in this context of *argue*

The phrase type information ('NP', 'PP-with', etc.) allows you to mark off in the sentence the actual section(s) relevant to your description. The information about grammatical function ('subject', 'complement') lets you assess the importance of the component for your database. From the point of view of its dictionary entry, a verb's complements are more important than its subject. Dictionaries rarely say much about the subject of a verb headword, although for automatic information retrieval and other computerized processes the fact that the subject noun is singular and semantically +HUMAN is of interest.

The set of threefold descriptions of each component (frame element, phrase type, and grammatical function) in the example sentence constitutes the valency pattern to be found there, and contains most of the information you need to extract from this sentence for your database. The total of all possible complements of the verb (including direct objects) in one LU is the valency of the headword for that LU. Knowing its valency is essential if people are to use the headword correctly. What we have learned in this

analysis of one sentence containing the headword *argue* is a piece of contextual information, namely that the verb *argue* in the 'quarrel' sense can be used with the preposition *with* when you want to mention the person being argued with, and with the preposition *about* when you want to mention the reason for the argument.

This type of information is usually explicitly set out in good dictionaries for learners, as may be seen both from the *ask* entry in Figure 5.24 and from the *argue* entries in Figure 5.28. The prepositional complementation is spelled out clearly in all the entries, and they all indicate that the verb is intransitive.

argue /…/ *v* **1** [I] to disagree with someone in words, often in an angry way: *We could hear the neighbours arguing* | **[+ with]** *Gallacher continued to argue with the referee throughout the game* | **[+ about]** *They were arguing about how to spend the money* | **[+ over]** *The children were arguing over which TV programme to watch*. […]

LDOCE-2003

argue /…/ **1** vi **a** (*dispute*, *quarrel*) se disputer (**with sb** avec qn) (**about sth** au sujet *or* à propos de qch). **they are always arguing** ils se disputent tout le temps; **don't argue!** pas de discussion! […]

CRFD-2006

Fig 5.28 Showing headword complements in learners' dictionaries

These complements are however not the only grammatical possibility for this sense of *argue*. A different context will yield different information; for instance, in the sentence *they are always arguing* we see that the two arguers can be conflated into a plural noun as subject, and the headword can be used without any complementation. The language-learner (as well as a computer handling language) needs to know this as well. And so you go on, studying the context details, until you're sure you've got all the essential information about its complementation from your set of corpus lines. At that point you can move on to the other facts about the headword that need to be recorded in the database. (Naturally, you don't go through this explicit, detailed, step-by-step analysis for each of your corpus lines. You soon learn to recognize the useful information in each.)

But before moving from one LU to the next, there is one more contextual feature that needs to be recorded: it's important to note not only the grammar of the various contexts but also the information that they hold about the collocates of the headword (cf. §9.2.7). What actual words co-occur with

argue in the citations for this sense? In particular, what words co-occur in a statistically significant way? Although the Word Sketches in the Sketch Engine (described in §4.3.1.5) give collocate statistics for the *lemma* and not the *LU*, they can be quite revealing about the collocational contexts of particular LUs. In the case of *argue*, we find that if you want your dictionary example to sound typical it should be about arguing with the *umpire* or the *referee* or your *father* – something of a social commentary, perhaps. The register and style of the collocates (cf. §6.4.1.4, §7.2.8.3–4) are also very clear from the Word Sketch. The collocates of *argue that*... in LU-2 comprise the following splendid set of formal and/or administrative words (given in order of statistical significance): *law*, *ban*, *Microsoft*, *Freud*, *treaty*, *constitution*, *Britain*, *restriction*, *client*, *merger*, *approach*, *court*, *system*, *change*, *Marx*, *government*, *feminist*, *distinction*, *sociology*, *recount*, *measure*, *factor*, *bureaucracy*, and *legislation.*

➔ All this is worth recording, even if you are analysing the corpus with a particular design of dictionary in mind, and you're pretty sure it won't be used immediately. Especially it's worth recording every last thing if the database is to feed into any kind of bilingual dictionary, because of the anisomorphic relationship between the source language and the target language. You can't be sure that something which looks like a 'normal' use in the SL doesn't have a single-word equivalent in the TL. For instance, *ask for*, as in *ask for a glass of water*, has a transitive equivalent in French, *demander.*

5.5.3 Properties of the source texts

Finally, when we are analysing corpus data in an attempt to collect the facts about a word for our dictionary entry, it's important to be able to discover from the concordances the actual source of each citation. This information is stored in the 'document headers' of each text in the corpus (cf. §3.6.2), of which a sample is shown in Figure 5.29.

doc.docid	eb000j1c
doc.author	
doc.title	[Leeds United e-mail list]
doc.genre	inf
doc.genre2	leisure

Fig 5.29 Part of the header on a corpus document

Using this information the computer can tell you whether a particular citation comes mainly from spoken or written language, or political documents, or feminist publications, and so on. Checking the document headers isn't something you do the whole time, but if you have any doubts about the way a word is used, it's useful to be able to check up on where the citation came from. How many different speakers and writers actually use it? What kind of texts is it found in? If there is only one instance of a doubtful construction from (say) the text identified as 'eb000j1c', or if there are only a couple and both from the same source, you may decide not to note it. This is probably a wise decision, since the header in Figure 5.29 indicates that the text is an email message, and as yet email messages are not prime sources of dictionary material.

To summarize lexicographic relevance: the wordclass of the word is central to what is relevant to record, and there are lists of the principal co-constituents of a clause that are relevant for each of the four major wordclasses (cf. §9.2.5). However, lexicographically relevant information is of course more than simply grammatical facts, and includes multiword expressions and other types of collocation in which the headword participates, its significant collocates in the corpus, and a judicious choice of example sentences. The individual lexicographically relevant items to be recorded for any single lexical unit are discussed in detail in Chapter 9.

Reading

Recommended reading

Theory: Cruse 1986, 2004; Lyons 1981 (esp. 136–151); Apresjan 2002.
Regular polysemy: Apresjan 1973.
Application of theory: Atkins 1993; Hanks 1993; Atkins, Rundell, and Sato 2003.
Frame semantics: Atkins, Fillmore, and Johnson 2003.

Further reading on related topics

Theory: Apresjan 1992; Clark and Clark 1979; Cruse 1990; Fillmore 1992, 1995, 1997, 2002, 2005; Lakoff 1987; Lakoff and Johnson 1980; Lehrer 1990; Levin 1993; Lyons 1969 (esp. 400–435), 1977 (esp. 270–301); Mel'čuk 1996, 1998; Mel'čuk, Clas, and Polguère 1995; Murphy 2003; Rosch 1973; Ruppenhofer, Baker, and Fillmore 2002; Shcherba 1995; Taylor 1995; Vandeloise 1990; Zaenen 2002.

Regular polysemy: Copestake and Briscoe 1995; Nunberg and Zaenen 1992; Ostler and Atkins 1992.

Application of theory: Atkins 1995; Atkins, Kegl, and Levin 1988; Atkins and Levin 1995; Atkins, Levin, and Song 1997; Atkins and Grundy 2006; Fillmore and Atkins 1994, 2000; Fontenelle 2000.

WordNet: Miller et al. 1990; *EuroWordNet* Vossen 2004.

Websites

FrameNet: http://framenet.icsi.berkeley.edu/

WordNet: http://wordnet.princeton.edu/ ; http://www.globalwordnet.org

Discussion list: the LINGUIST list is both a discussion list and an invaluable information source for linguists: http://www.linguistlist.org/

6

Planning the dictionary

6.1 Preliminaries

In this chapter, we discuss the major decisions that have to be made about what the dictionary will contain. At this point, we have already determined the type of dictionary we plan to produce, and we have clear ideas about who will use it and what they will use it for (Chapter 2). We have collected the linguistic data that will form our raw material (Chapter 3). And the other resources we will need – software, Style Guide, template entries – are in place (Chapter 4). When we talk about the *content* of a dictionary (as opposed to the resources which underpin its creation), the terms 'macrostructure' and 'microstructure' are often used. Deciding on the types of entry the dictionary will include, and organizing the headword list, are *macrostructure* decisions, and these are the issues we address in this chapter. Planning the entries in the dictionary and deciding on their structure and components are *microstructure* decisions, and these form the focus of Chapter 7.

Part of our objective here is to introduce the vocabulary you will need in order to talk about the structure of dictionaries. In particular we aim:

- in §6.1 to clarify some of the basic terms and concepts
- in §6.2 to name and describe the different kinds of words and phrases which you have to be able to recognize in corpus data and which can be dealt with as headwords in a dictionary
- in §6.3 to set out the various large components of a print dictionary (the actual dictionary text and other material)
- in §6.4 to explain features you have to consider when deciding the words to include in the dictionary
- in §6.5 to look at the other main decisions that have to be made about the headword list
- in §6.6 to describe the principal types of entry that are to be found in most current dictionaries.

Figure 6.1 gives an outline of the chapter and the issues it covers.

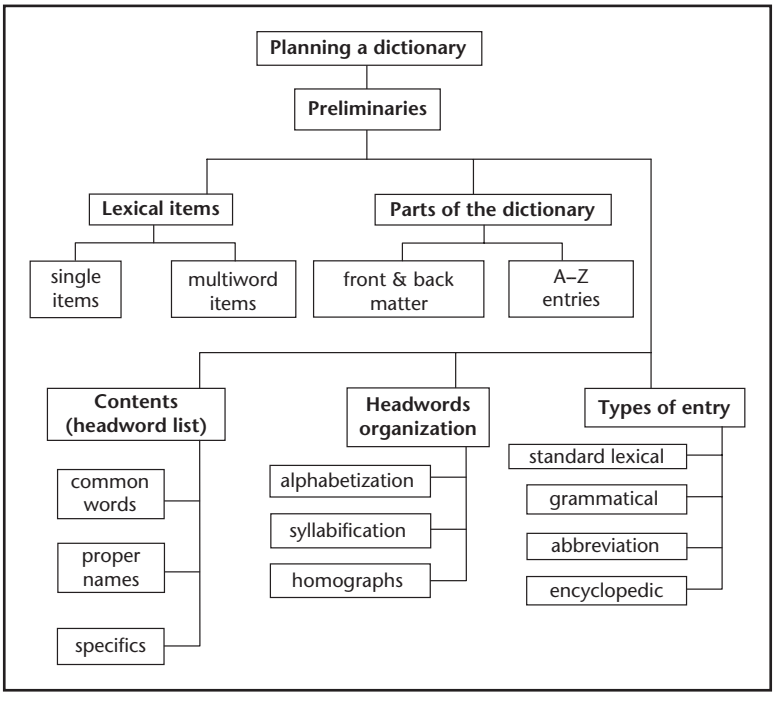


Fig 6.1 Contents of this chapter

6.1.1 Talking about words

The headword list is a list of the words that are the headwords of entries in the dictionary.

How many 'words' does this sentence contain? The answer depends on what you call a word. If it is simply a string of letters bounded by spaces, then there are eighteen. If however the four instances of *the* are not counted separately, and similarly the two instances of *list*, then there are fourteen. And if *headword* and *headwords* are taken as two forms of the same word, then this brings the total down to thirteen. So for the sake of clarity (especially when discussing corpora and their contents), we can say that the sentence contains eighteen *tokens*, fourteen *types*, and thirteen *lemmas*.

When we say that this sentence contains thirteen lemmas, we are using the word in its lexical and morphological sense: this is the way 'lemma' is used when people are talking about words and word forms. Thus used, the lemma *play* is made up of the forms *play*, *plays*, *played*, and *playing*.

But in the context of dictionaries and dictionary planning, we look at 'words' from a different perspective. The word *lemma* can be used to mean the headword in all its forms, but it has an extended sense in discussions of meaning and grammar, where it's often used to denote a word belonging to a particular wordclass, as for instance the two lemmas *play*:

- *play* (noun): formed by the noun forms *play* (singular) and *plays* (plural)
- *play* (verb): formed by the verb forms *play* (various persons), *plays* (3rd person singular), *played* (past tense and past participle), and *playing* (present participle).

A single-word lemma can have various senses, which we call *lexical units* (explained in the next paragraph). Some lemmas exist in multiword form, and these can also have more than one sense: for instance the phrasal verb *set off* has several meanings, including (1) begin a journey, and (2) detonate (a bomb, etc.). Some types of multiword lemma, such as compounds (*ice cream*) and phrasal verbs (*set off*), regularly appear as headwords in dictionaries. For clarity in this chapter we will use the term *headword* to denote a lemma when it is the headword of an entry in a dictionary, and keep *lemma* for discussions of meaning and grammar.

A headword in one of its senses is a *lexical unit* (or LU), and in this book we use the term to denote one sense (either during the analysis process or

within a dictionary entry[1]). LUs are the core building blocks of dictionary entries. In the entry in Figure 6.2 for the lemma *absolute*, there are five LUs, numbered in bold, each with its own definition and example(s).

absolute / ˈæbsəluːt/*adj* **1** complete or total: *I have absolute confidence in her.* | *We don't know with absolute certainty that the project will succeed.*
2 [*only before noun*] *especially BrE informal* used to emphasize your opinion about something or someone: *Some of the stuff on TV is absolute rubbish.*| *How did you do that? You're an absolute genius.* | *That meal last night cost an absolute fortune.*
3 definite and not likely to change: *We need absolute proof that he took the money.*
4 not restricted or limited: *an absolute monarch* | *Parents used to have absolute power over their children.*
5 true, correct, and not changing in any situation: *You have an absolute right to refuse medical treatment.*
LDOCE-4 (2003)

Fig 6.2 Partial entry showing five lexical units

Every piece of information within one numbered dictionary sense is valid only for that LU. Thus there is a restriction on the way the second LU of *absolute* (but none of the others) is used – only before the noun. Similarly, that LU is the only one that is informal, and that use is especially in British English (as opposed to American, etc.).

6.2 Types of lexical item

A 'lexical item' is any word, abbreviation, partial word, or phrase which can figure in a dictionary (often as the headword of an entry) as the 'target' of some form of lexicographic description, most commonly a definition or a translation. It's important to be aware of the various kinds of lexical item, as there are important differences in the way each is handled in the dictionary. What constitutes a lexical item is to a certain extent language-specific, but the principal types in English are summarized in Figure 6.3 and described in this section. In §6.4.1.3 we consider these types in the context of deciding what should be given headword status in the dictionary. In this section we simply describe them and exemplify them.

[1] In this case it can be called a *dictionary sense*.

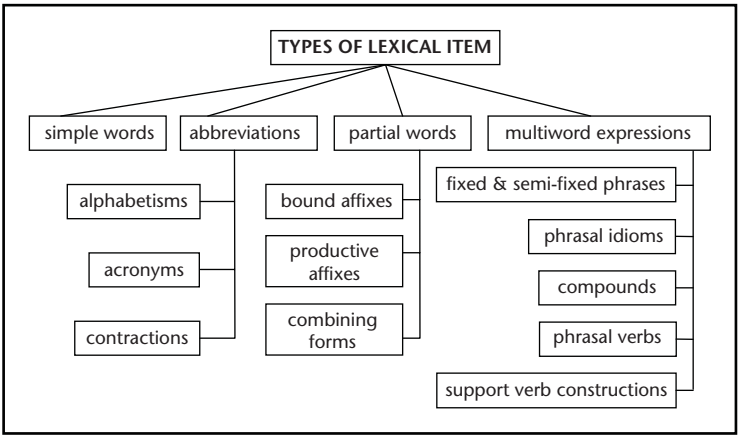


Fig 6.3 The various types of lexical item

6.2.1 Single items

There are essentially three types of single item: simple words, various kinds of abbreviation, and partial words.

6.2.1.1 *Simple words* These are the common words of the language (e.g. *be*, *like*, *head*, *possible*, *remember*, and thousands of others). This type includes all the wordclasses and may be subclassified into two types, lexical and grammatical words.

Lexical words These consist of nouns, adjectives, verbs, adverbs and interjections: they often have several meanings; they make up the bulk of the words of the language and hence of the words in the dictionary.

Grammatical words As distinct from lexical words, the principal role of grammatical words (also known as 'function words' or 'closed-category items') is to perform a function in the sentence. This usually involves either linking parts of the sentence (e.g. *and*, *or*, *because*, *when*), referring to something mentioned already or about to be mentioned (*she*, *our*, *yours*, *who*), or specifying (*the*, *many*, *every*). From the dictionary-making perspective there are at least five different types of grammatical word. They are:

- prepositions: e.g. *to*, *from*, *with*, *up*, and so on, including uses both transitive (*put it in the box*) and intransitive (*put it in*)[2]

[2] We call such transitive and intransitive prepositions *particles* when they are phrasal verb elements.

- conjunctions: e.g. *because*, *when*, *and*, *or* . . .
- pronouns: including personal, reflexive, possessive, demonstrative, relative, interrogative, negative, indefinite, etc. pronouns, e.g. in that order *we*, *yourself*, *his*, *that*, *which*, *what*, *none*, *someone*
- auxiliary verbs: e.g. *be*, *do*, and *have*, and the modals *may*, *could*, etc.
- determiners: including the definite and indefinite articles, demonstratives, possessives, numerals, negatives, quantifiers, and predeterminers, e.g. in that order *the*, *a*, *this*, *his*, *three*, *no*, *some*, *all*.

6.2.1.2 *Abbreviations and contractions* There are three subclasses here, and most dictionaries will give all three headword status:

- Alphabetisms
  the initial letters of a group of words, pronounced as series of letters, e.g. *BBC*
- Acronyms
  the initial letters of a group of words, pronounced as a word, e.g. *NATO*
- Contractions
  two or more words fused with some letters omitted, e.g. *don't*, *wouldn't*.

6.2.1.3 *Partial words* These include:

- Bound affixes
  e.g. *im-* (*impossible*), *-ment* (*attainment*). These are rarely given headword status in modern dictionaries.
- Productive affixes[3]
  e.g. the prefix *ex-* attached to nouns denoting a person having some specific status, as in *ex-wife*, *ex-mayor*; also, the suffix *-gate* attached mainly to proper names and indicating a scandal, as in *Monicagate* referring to the Clinton–Lewinsky affair, also called *Zippergate* in the press. Productive affixes are constantly used (in specific environments) to create new complex word forms, and they must be explained in a dictionary. Productive prefixes (*un-*, *de-*, *anti-*, etc.), of which there are a good number in English, usually appear as headwords, so it's important to recognize the productive examples when you come across them in the corpus. There are fewer productive suffixes, and it's difficult

[3] Some dictionaries call these *combining forms*, but we give this term a more specific definition.

to believe that users, having failed to find *Zippergate* or *Italianness*, would look up *-gate* or *-ness*. For that reason, some dictionaries decide to omit productive suffixes from the headword list.

- Combining forms
  These are essentially headwords or their inflected forms which occur as first or second elements of hyphenated compounds. The meaning each carries remains constant throughout the compounding process, but the process itself is open-ended, with the result that many of the instances found in the corpus are one-offs (*hapax legomena*). Corpus frequency does not help in such cases, and we need somehow to make it possible for the dictionary user (especially the language-learner) to understand these compound words when they find them. The most common word-classes found in initial position in such compounds are numerals (*one-legged*), nouns (*vinyl-covered*), and adjectives (*flat-leafed*). Numerals pose no problem of understanding. As for nouns and adjectives, it is possible to make these hyphenated forms into stand-alone headwords (although their entries may show that *vinyl*, *flat*, etc. frequently form the first element in hyphenated compounds). It is more difficult to decide what to do with the second element of such compounds – forms like *-covered* and *-leafed*. In such cases the choice lies between making them into headwords, or handling them within the entry for *cover* and *leaf* (see Figure 7.37 for an example of such treatment).

6.2.2 Multiword expressions

Of the four principal classes in Figure 6.3, only multiword expressions pose real problems of identification. The term covers all the different types of phrases that have some degree of idiomatic meaning or behaviour. Many groups of words, such as *she put it in the* or *immediately below the*, co-occur frequently in corpus text but are of no real interest to lexicography.[4] Our problem is to sort the wheat from the chaff. Which multiword items should be treated as 'multiword expressions' (MWEs) in our dictionaries?

There is a large body of work by theoretical linguists on the classification of MWEs, but no clear set of criteria emerges for the various subclasses proposed. Many dictionaries give specific treatment to compounds and phrasal

[4] Some theorists call such fragments 'collocations', but for us the term has a more precise meaning (cf. §7.2.7.1).

verbs, but it is not usual for dictionaries to distinguish many subclasses of phrases. This is because the boundaries are so fluid that it has proved impossible to establish watertight criteria for lexicographers to apply in dealing with multiword items. For those who want to delve further into this topic, we recommend the work of A. P. Cowie and Rosamund Moon, particularly Cowie 1994, 1998, and 1999a, and Moon 1998, Mel'čuk 1998 presents an interesting and thorough theoretical approach to MWEs.

MWEs are a central part of the vocabulary of most languages, and need to be accounted for in the dictionary. They are particularly important for learners' dictionaries, both monolingual and bilingual, since language learners may not recognize them as significant units of meaning, cannot usually compose them, and will often have problems understanding them. Some may be easy to spot (such as *kith and kin*, *kick the bucket*, or *birds of a feather*), but many are less idiomatically salient. There are several helpful tests which you can apply to a phrase you are doubtful about. The lexicographer's rule of thumb is 'its meaning is more than the sum of its parts'.

In this section we give a very pragmatic lexicographers'-eye-view of this difficult area of language, distinguishing some types of MWE to help you recognize them in corpus data, and deal with them as specified in your dictionary's Style Guide.

6.2.2.1 *Fixed and semi-fixed phrases* All fixed and semi-fixed phrases are important, and worth recording during the analysis process of dictionary-writing. It is useful to be able to recognize the following types at least:

- Transparent collocations: i.e. phrases which are salient in corpus citations yet seem to have no idiomatic meaning[5] e.g. *to risk one's life*.
- Fixed phrases: e.g. *ham and eggs*; *knives, forks and spoons*; *kith and kin*. Some fixed phrases function as compounds (see §6.2.2.3).
- Similes: e.g. *white as snow*; *pale as death*; *drunk as a lord*.
- Catch phrases: e.g. *if you can't beat 'em*, *join 'em*; *horses for courses*.
- Proverbs: e.g. *too many cooks (spoil the broth)*.
- Quotations: e.g. *to be or not to be*; *an eye for an eye*.

[5] In that there is an open paradigm at each of the lexical nodes, cf. *to save / value one's life*, and *to risk one's fortune / future*. In such cases substituting a different word for one of the nodes does not affect the meaning of the other, and vice versa. The meaning of the whole collocation is simply the sum of the parts.

- Greetings: e.g. *good morning*; *how do you do?*
- Phatic phrases: e.g. *have a nice day*; *take care of yourself.*

6.2.2.2 *Other phrasal idioms*[6] These are the most difficult MWEs to handle in lexicography. In the absence of hard and fast criteria, it is well nigh impossible to be wholly consistent. If the phrase passes the 'meaning is more than the sum of the parts' test, then check to see if it has one or more of the properties listed below, remembering that:

- Every idiom has at least one.
- Some have several.
- No idiom has them all.

Some of the properties are lexical, relating to the actual words which make up the idiom; some are morphological, relating to inflections which the constituent parts may undergo; some are syntactic; some are semantic; and some relate to more than one of these aspects of language.

(1) The wording is never entirely fixed: some common variations are given below.
   - Alternative words may be substituted without changing the meaning:
     e.g. *to throw in the sponge / towel*; *hit and / or miss*; *hop, skip / step and jump*.
   - There are parallel idioms with opposite meanings: e.g. *to have a heart of gold* and *to have a heart of stone*; *to be in someone's good books* and *to be in someone's bad books.*
   - There is no fixed canonical form[7]: e.g. the variants on *chicken and egg*, as in *which came first, the chicken or the egg?*, *it's a chicken-and-egg situation*, *it's another case of the chicken and the egg*, and so on.
   - There is no complete canonical form, but there are semantic restrictions on what can fill the open slot: e.g. *it was a . . . 's dream* or *it*

[6] This term is intended to include all types of phrases routinely recognized as idiomatic, with the exception of: the fixed and semi-fixed phrases discussed in §6.2.2.1, and the compounds, phrasal verbs, and support verb constructions discussed in §§6.2.2.3–6.2.2.5, which – for English, at least – it is more convenient to consider separately.

[7] This is the most basic form of a word or phrase, the one used when it is entered in dictionaries; cf. §9.2.6.2 for a fuller explanation.

was a . . . 's dream come true (here the possessive noun must introduce the idea of some typical activity), as in the corpus lines in Figure 6.4.

the audience were	a comic's dream	
the airline will be	a commuter's dream,	the expressway to the sky
it's more than	a consumer's dream,	an experiment in capitalism
undiscovered evidence like	a convict's dream	come true
it was	a couch-potato's dream	– bed-size settees and chairs
I thought "This is	a decorator's dream"	– it was in such disrepair,"
[his] customers are	a demographer's dream,	with incomes of over $35,000
the 1990s began as	a disarmer's dream	come true
so much live football . . . it is	a fan's dream	come true to watch it
romance between them	a father's dream	

Fig 6.4 KWIC lines for *a . . . 's dream*

(2) There are syntactic restrictions upon the idiom's behaviour, in that it undergoes only limited grammatical transformations: e.g. *it was a football manager's dream* but not **the dream of a football manager*; *it's raining cats and dogs* but not **cats and dogs are being rained.*

(3) The idiom shows morpho-syntactic flexibility, allowing inflections, agreement of possessives, and so on: e.g. *to get too big for one's boots*, as in:

> Joe *is getting* too big for *his* boots.
> She *had got* too big for *her* boots.
> People who *are* too big for *their* boots.

6.2.2.3 *Compounds* Compounds of interest to lexicographers belong mainly to three wordclasses: nouns (the most frequent case, e.g. *lame duck*, *civil servant*), adjectives (e.g. *sky blue*, *stone deaf*), and verbs (of which by far the most common are the 'phrasal verbs', cf. §6.2.2.4). There are also compounds in other wordclasses, for instance *in spite of* is a compound preposition, but these are not so difficult to recognize and we shall not consider them in this section.

As in the case of other MWEs, there are idiomatic and non-idiomatic types of compounds. Compounding is a function of the language, and non-idiomatic compounds (e.g. *table leg*) are spontaneously produced and found in their thousands in corpus data. Semantically transparent, they pose few problems to lexicographer or dictionary user; conforming to the

grammatical rules of the language, they pose no problem to the language-learner. They will not be further considered here.

However, in the course of corpus analysis, it can be difficult sometimes to distinguish certain idiomatic compounds from non-idiomatic. Here again, there are no watertight criteria for identifying idiomatic compounds (which we shall now call simply 'compounds') in corpus data, but there are a few properties shared by many of them, in particular:

- The compound is fixed in form. It can take inflections (e.g. *civil servants*, *courts martial*), but words can't be added to it or removed from it.
- It participates in semantic relationships (synonymy, antonymy, hyponymy: cf. §§5.2–3) with single words: e.g. *civil servant*, like *teacher* and *doctor*, is a hyponym of *professional*; *sky blue* and *bottle green*, like *red* and *purple*, are cohyponyms of *colour*.
- Like other MWEs, its meaning is more than the sum of its parts. This semantic test (however unscientific) may be further expanded to give three types of compound worth including in a dictionary.

These three types of compound[8] are listed below: each compound-type in the three-part distinction carries a rule of thumb based on meaning to help you identify them. Most dictionaries do not distinguish these three types in their entries, normally treating them in the same way, but the lexicographer must learn to recognize them in the corpus.

[8] This analysis may not be couched in linguistic terminology but it comes with a guarantee. The first time one of us met Igor Mel'čuk, already a renowned linguist known for his direct manner, the exchange went like this:

IM How do you handle compounds in your dictionary?
SA (timorously explained this three-part analysis)
(pause)
IM You are right.
(pause)
SA How do you know I'm right?
IM Because you agree with me.
(pause)
SA How do you know you're right?
IM God told me.

Later SA told this story to Mel'čuk's long-time friend and colleague, Juri Apresjan, who instantly recognized the authentic Mel'čuk voice. 'God was right,' said Apresjan.

- Figurative compounds
  'An XY is not a Y that is X: it is not necessarily a Y at all.'
  A *lame duck*, for instance, is not a duck that is lame, nor is a *civil servant* a servant that is civil (often, indeed, far from it). In terms of the semantic hierarchies (cf. §5.2) the second element is not a cohyponym of the whole unit.
- Semi-figurative compounds
  'An XY is a Y but it is not a Y that is X.'
  Thus a *high school* is a school but not a school that is high; if you're *blind drunk* you are drunk but not necessarily blind. Here the second element is the superordinate (hypernym) of the whole unit.
- Functional compounds
  'An XY is a Y that has to do with Xs, but also more than that.'
  It is a specific type of thing or person, and may well have a specific translation – often a single word – in another language, e.g. *house agent, police dog, can opener*. Not everyone who sells a house is a *house agent*. No one hearing the expression *police dog* will see in their mind's eye a spaniel or a poodle. Not everything you use to open a can is a *can opener*, as many unprepared picnickers will agree.

The first and second (figurative and semi-figurative) are fairly easily recognized in corpus data, but the third (functional) frequently escapes notice, as it is often confused with open, productive, non-idiomatic compounds like *house size* or *police pensions*. Boundaries in language, and especially in linguistic classification, are notoriously fuzzy, and there are many compound nouns which some lexicographers would classify as simply productive uses of the two words (as in *table leg*) and which others would wish to treat as functional compounds (as in *table football*). The skill in dictionary writing is to be as systematic as possible across the language: regular feedback is needed if a team of lexicographers is to achieve a harmonious whole.

6.2.2.4 *Phrasal verbs* In this section, we are concerned simply with recognizing in corpus data phrasal verbs of interest to lexicography. A phrasal verb is a multiword expression consisting of a verb plus one or more particle(s). The particle may function either as an adverb (*away*, *out*) or as a preposition (*with*, *to*), or both (*in*, *through*). In the context of dictionary-writing, where phrasal verbs must be classified in the Style Guide so that they can be handled systematically in the dictionary, it is useful to

look at the kind of meaning they can carry (their semantics), and how they interact with the rest of the language (their syntax).

Semantics A phrasal verb unit may have a 'literal' meaning and one or more 'figurative' or 'metaphorical' meanings. A good example (see Figure 6.5) is the intransitive *run out*, where the literal sense of *run out* ('go outside at a run') in lines 1–5 contrasts with the figurative meaning ('become depleted') in lines 6–10.

1	I looked to see if Jason would come	running out.	
2	The rat came	running out,	eyes left and right..
3	A youth with blood on his face suddenly came	running out.	
4	Carine	runs out,	her mother after her.
5	As the guard	ran out,	the gunman shot him.
6	My money is	running out.	
7	Time was	running out.	
8	The bananas have	run out.	
9	Our patience finally	ran out.	
10	The world's supply of oil had nearly	run out.	

Fig 6.5 KWIC lines for *run out*

In the analysis stage of lexicography, all verb + particle uses should ideally be recorded in the database, because it may be necessary to include even the literal uses in a bilingual dictionary, since there may be a one-word equivalent in the target language. However, when analysing corpus data for a monolingual dictionary, it is often possible to omit literal verb + particle uses in order to concentrate on the figurative ones, which usually need to be defined. As with other MWEs, there are no watertight criteria for recognizing dictionary-worthy phrasal verbs, but some of the indications of phrasal-verbhood are given below.

- It is a fixed MWE with syntactic rules regarding (where these apply) pronoun objects, which must be embedded between verb and particle in the case of some two-part phrasal verbs, e.g. *pass it over*, *take him away*, and must always follow three-parters *come up with it*.
- It has a discrete unitary meaning, and (like *civil servant* and other compounds) may participate with single words in semantic relationships (synonymy, antonymy, hyponymy: cf. §§5.2–3): thus the phrasal verbs *put off* and *put up with* have the single-word synonyms (respectively) *postpone* and *tolerate*.

- It often has a single-word translation in another language, for instance, *come up with* is often *trouver* in French, *hand over* is *passer*, and so on.

Syntax Most Style Guides distinguish the following syntactic forms of phrasal verbs:

(a) verb + adverbial particle operating as an intransitive unit
e.g. *get up early*, *he passed away last year*
(b) verb + adverbial particle, a transitive unit
e.g. *hold over the decision until then* | *hold it over until then*
(c) verb + prepositional particle, a transitive unit
e.g. *see through someone's evil plans* | *see through them at once*
(d) verb + adverbial particle + prepositional particle, a transitive unit
e.g. *look forward to a party*

Of these, (a) – the only intransitive unit – does not usually present any problems of identification. In the case of transitive constructions, it is sometimes difficult to decide between types (b) and (c), i.e. whether the particle is adverbial or prepositional. The rule of thumb is: if a pronoun object can be placed between the verb and particle, the particle is an adverb, as in type (b), e.g. *hold it over*; if the pronoun object must be placed after the particle, the particle is a preposition, as in type (c), e.g. *see through them*.

Problem areas Phrasal verbs constitute a difficult area for lexicographers to handle: some are easy to identify and treat according to the Style Guide, but others are not. Three of the commonest problems are briefly discussed below: the Style Guide must tell the lexicographers how to deal with them.

(1) Two- or three-part phrasal verbs?
Some phrasal verb units are easy to identify as type (d) in the classification given in 'Syntax' above: *come up with* is clearly type (d), since in response to *he came up with a good idea*, you can't say **he's always coming up*. However, in the case of other three-part phrasal verbs, there can be confusion between types (a) and (d). Some intransitive units are commonly found with a prepositional complement, e.g. *get away* is very often followed by the preposition *from*, as in the corpus lines in Figure 6.6. The phrasal verb *get away* is very common in the meaning of 'escape', and will be recorded in the database. The problem is to decide whether *get away from* (lines 5–10) should be recorded separately, as a three-part phrasal verb. There are no clear-cut linguistic criteria to help here, but it is essential that the Style

Guide gives clear guidance on such items, in order to maintain a consistent approach across a team of lexicographers.

1	He wants to let the fox	get away.	
2	They	got away	on time for once.
3	The raiders	got away	in another truck.
4	I'm not going to let you	get away.	
5	You said you wanted to	get away	from it all.
6	I go to the pub to	get away	from the wife.
7	It will be good for her to	get away	from here.
8	She left as soon as she could	get away	from lunch with them.
9	I want to	get away	from party politics.
10	I wanted to	get away	from winter in Norway.

Fig 6.6 KWIC lines for *get away*

(2) Motion verbs + directional particle

Many motion verbs regularly combine with a particle (prepositional or adverbial), resulting in a phrasal verb which always has a literal meaning and often one or more figurative meanings as well. This is clear from the corpus examples of *run up* in Figure 6.7.

1	My sister	runs up	to speak to us.
2	They came	running up	to the laboratory.
3	He	ran up	the stairs into the Long Room.
4	Keep a check on how your bill is	running up	
5	The market has	run up	quite a bit.
6	I saw a path	running up	to the bank on my left
7	The average student is expected to	run up	an overdraft.
8	Two of the volunteers	ran up	the white flag.
9	She bought cotton material and	ran up	simple shift-like dresses.

Fig 6.7 KWIC lines for *run up*

In lines 1–3 inclusive the particle is directional and the meaning of the phrasal verb unit is simply the sum of its parts. However, in lines 4–9 the phrasal verb is used, both intransitively and transitively, in some of its many figurative meanings. It is best if all directional particles used routinely with motion verbs are recorded in the analysis database. Although few monolingual dictionaries have space for the literal uses, in many bilinguals they are essential for translation purposes. Line 2, for instance, would be translated into French as *ils sont arrivés en hâte au laboratoire*.

(3) Semantically related, syntactically distinct
Dictionaries which organize phrasal verbs exclusively on the basis of their syntax face the problem of physically separating on the printed page very similar usages. This is particularly awkward for bilingual dictionaries, where the intransitive and transitive units may have the same translation, as for instance *get across* in the corpus lines in Figure 6.8, which in French has the equivalent *traverser* (intransitive and transitive). It is always better if the dictionary design allows these related intransitive and transitive usages to be handled together, or at least to be explicitly linked.

1	Jenny could	get across	by the stepping stones.
2	His mare would never	get across,	but he might.
3	How did they	get across	the river?
4	They greased the ropes so the rats couldn't	get across	them.
5	It's tough, but they	get across	the lake.

Fig 6.8 KWIC lines for *get across*

6.2.2.5 *Support verb constructions* The term 'support verb' is used differently by different people. Here we restrict it to the so-called 'light verbs'[9] which carry less meaning in such constructions than in many other contexts. Of these verbs, the most frequent are *make*, *take*, *have*, *give*, and *do*, as in *to make a complaint*, *to take a decision*, *to have a rest*, *to give a lecture*, *to do a dance*. Compare the semantic content of *take* in the first group of sentences with that in the second group below:

He took from his pocket a blue handkerchief.
He took a slice of bread from the wooden board.
She took some money from a cash-point machine.
Chris took the trolley from her.

A few of us took a walk through the village.
We must influence those who will take the decision.
Shall we take a vote on it?
I always take a shower in the morning.

In the first group, the verb *take* carries its full lexical meaning of 'remove'. In the second group the verb turns the noun into a predicate, and the

[9] In support verb constructions they may also be called 'delexical verbs' or even 'empty verbs'.

construction is the semantic equivalent of the verb cognate with the noun, as may be seen from these equivalences:

A few of us took a walk through the village → walked through the village.
We must influence those who will take the decision → who will decide.
Shall we take a vote on it? → vote on it?
I always take a shower in the morning. → always shower in the morning.

Support verb constructions are an essential part of the vocabulary, and must be recorded when found in corpus data. The Style Guide must therefore indicate how they should be handled in database and dictionary.

6.3 The constituent parts of a dictionary

People talk of 'the dictionary', but every dictionary is unique. A good dictionary reflects the type of people who will be using it and what they will be using it for (cf. §2.3). Knowing these facts helps us decide what goes into the dictionary and how the material should be structured (though in most projects, commercial constraints have a bearing on these decisions too). Despite wide variations in content, most dictionaries have two major components: the A–Z entries (or their equivalent in languages which don't use the Roman alphabet), and all the other 'non-linear' material which we can broadly categorize as 'front matter' and 'back matter'. We briefly describe these components here, before looking at headword-selection principles in the next section.

6.3.1 Front and back matter

Print dictionaries traditionally include material of various types as 'front matter' (whatever precedes the A–Z text), and 'back matter' (whatever follows it). These 'locational' terms are of course irrelevant in the case of electronic dictionaries, but the same material (and often a great deal more besides) will be accessible in an electronic environment too. The content of these sections varies a great deal depending on the perceived needs of users. Pick up any two dictionaries and you will find that they have quite different material in their front and back matter. The front matter typically contains a foreword and acknowledgements, some kind of introduction to the dictionary, and an explanation of abbreviations, labels, and codes used in the text.

But it may also offer mini-essays on certain aspects of language ('the history of the language' or 'English throughout the world', for example), depending on the type of market it aims at. The back matter (sometimes also called the 'end matter') often includes lists such as verb tables, numbers, weights and measures, chemical elements, Roman numerals, the books of the Bible, etc., but it may also provide maps, diagrams, and other material geared to the needs of the target user. In pedagogical dictionaries (whether bilingual or monolingual), you will often find additional information in a centre section (the 'mid-matter'). This may deal with language issues (such as grammar, collocation, word formation, and regional varieties), or provide useful study aids such as guidance on writing essays, reports, and CVs, as well as model letters and emails. Bilingual dictionaries may also include lists of *faux amis* and practical guides to various aspects of living in the countries where the two languages are spoken.

One thing that most types of English dictionary have in common is a front-matter section called something like *How to use the dictionary*, which introduces the reader to the conventions of the dictionary layout. Figure 6.9 gives an example of such material, drawn from the *Concise Oxford Dictionary* 9th edition (1995). (Compared with what is on offer in some contemporary dictionaries, this example is rather minimalist.)

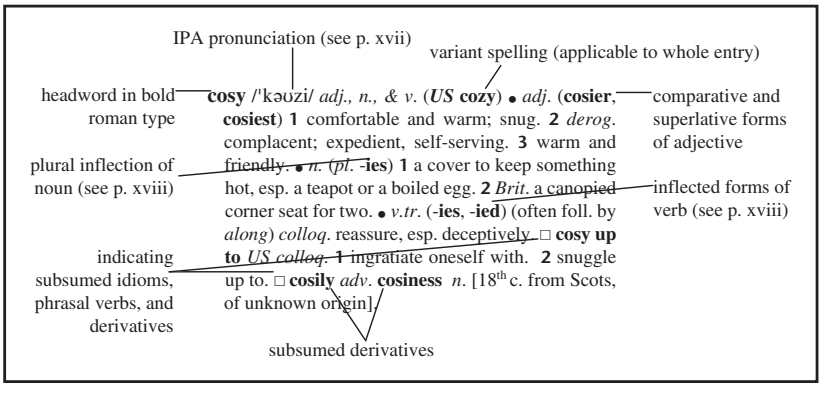


Fig 6.9 Explicated entry from *Concise Oxford Dictionary* 9th edn (1995)

6.3.2 The A–Z entries

The core of the dictionary is of course the great body of entries holding details of the meaning, grammar, and usage conventions associated with

each headword. Every dictionary is subtly different from every other in the principles applied during the headword selection, and in the design and content of the various types of entries used to present the information. As always, decisions on these matters are driven by the user profile, the target market of the dictionary, its competitors in that market, and consequently its costing and budget.

In the next section, we look at the various properties of words that you need to take account of when making decisions about what to include in your dictionary. It's important to remember that in this chapter we will be discussing words only from the point of view of *headword status*. In §6.5 we outline some decisions to be made regarding the way the headword is set out in the dictionary, and in §6.6 we briefly consider some classic entry types. Then in Chapter 7 we look in detail at the entry itself: the type of information it can hold, and the various components that can be used to present this information in the most user-friendly way.

6.4 Building the headword list

Dictionary users have high expectations. As Samuel Johnson noted drily: 'They that take a dictionary into their hands, have been accustomed to expect from it a solution of almost every difficulty'. We all know how annoying it is to find that the word you're looking up isn't in the dictionary. No dictionary (not even the electronic *Oxford English Dictionary*) can include everything everyone might want, so it follows that decisions about what to include in a dictionary (and what to exclude) are critical.

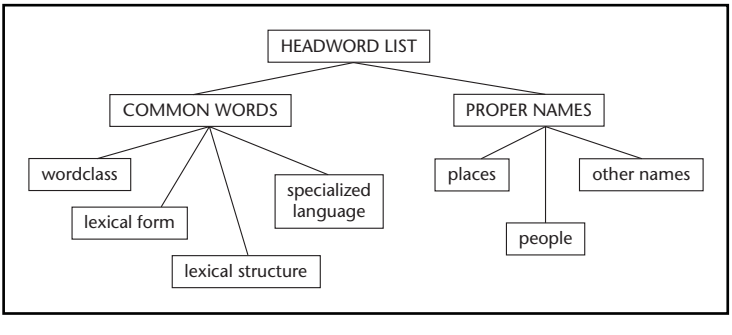


Fig 6.10 Factors to consider when deciding on headwords

Figure 6.10 summarizes the properties of words to be taken into account when deciding the types of headword that would best meet users' needs. For

each property briefly considered in this section, the question to ask is 'Shall we include as headwords in the dictionary words which have this property?' In the first few categories discussed, the answer is obviously 'Yes' – yes, we should include words which are nouns (or verbs, or adjectives, etc.); yes, we should include words whose lexical form is a single word; and so on. However, things soon become less clear-cut: 'Should we include as headwords lexical items which are only partial words?' (well, probably . . .); 'Should we include MWEs like *lame duck* as headwords?' (perhaps . . . it depends); 'MWEs like *to kick the bucket* – should that idiom be a headword in the dictionary?' (probably not, in many cases). The answers to such questions depend of course on the users; but they also depend on the size of the dictionary, and the associated costs of increasing the material, thus raising its price perhaps above that of its competitors. (There's no such thing as a free lunch.)

One final decision to be made: should the headword list contain all the headwords? Or should some be siphoned off into the back matter? It used to be commonplace to find proper names (for instance) excluded from the headword list and consigned to the end of the book. From a theoretical point of view this meant a 'purer' headword list, but from the point of view of users (who don't normally care about such things) it was simply another obscure idiosyncrasy of dictionary editors. Current practice is to include all headwords in one single list.

6.4.1 Common words

These make up the bulk of the headwords in the dictionary: the others are principally names and related words, and are considered in §6.4.2.

6.4.1.1 *Wordclass* The headword list will include all the major wordclasses, traditionally nouns, verbs, adjectives, adverbs, conjunctions, prepositions, determiners, and interjections, e.g. in that order *table*, *give*, *splendid*, *badly*, *because*, *in*, *the*, *ouch!* However, in the past few decades scholars have amended the classical set of parts of speech in order to reflect more accurately the grammar of English and other European languages, and so it is useful at the planning stage of a dictionary to decide which wordclasses (e.g. nouns, verbs) and subclasses (e.g. count and mass nouns, transitive and intransitive verbs) you need to identify in corpus analysis and record for the dictionary, and in particular which if any of these are to be headwords.

6.4.1.2 *Lexical form* There are four types of variant to consider here, and normally any item included must be corpus-attested.

- Variant forms
  e.g. *aluminium* (which is British English) and *aluminum* (the American form of the word).
- Variant spellings
  e.g. *ageing* or *aging*, and such British–American variants as *harbour* and *harbor*, *analogue* and *analog*.
- Inflections
  e.g. irregular plurals of nouns (*oxen*, *children*); irregular comparatives and superlatives of adjectives (*better*, *best*); verb inflections (*speaks*, *speaking*, *spoken*). Note that this refers only to inflections to be treated as *headwords*. The separate question of which inflections will be shown *within* a dictionary entry is a microstructure decision taken when the contents of entries are considered.
- Derived forms
  i.e. words related by derivation to other headwords, e.g. *highhandedness*, *blissful*, *nakedly*. You have to decide if they should all be headwords (which is very space-consuming) or be included within the entry of the root word (which makes them more difficult for users to find).

6.4.1.3 *Lexical structure* There are five types of lexical structure to be distinguished when you are deciding the kind of items to include as headwords. They have already been discussed in detail in §6.2; we give some examples here in parentheses simply to situate them for you.

Simple words i.e. any complete word written between two spaces (*be*, *like*, *head*, *possible*, *now*, *in*). Most headwords have this form.

Abbreviations and contractions By this we mean initials standing for the component words in a phrase (e.g. *BBC*, *EU*, *i.e.*), or two words contracted and written as one (*hasn't*, *o'clock*).

Partial words This term covers prefixes and suffixes, both bound[10] (*distaste*, *civility*) and productive[11] (*ex-wife*, *ex-mayor*; *grandparenthood*, *servanthood*); and combining forms which occur in hyphenated compounds either as first element (*flat-topped*) or second element (*broad-leafed*).

[10] i.e. part of standard words.

[11] i.e. able to be used creatively to form 'new' words.

Multiword expressions[12] Of the various types of multiword expression discussed in §§6.2.2.1–4 above, it is unusual to give headword status in a general dictionary to the following:

- Transparent collocations
  e.g. *to risk one's life.* These make ideal examples of normal usage within an entry in monolingual learners' dictionaries (but they have less claim on space in a monolingual dictionary for native speakers), and it is important to record them in the database for that reason.[13] More importantly, in bilingual dictionaries they may be included because they require a specific translation, sometimes an idiomatic collocation in the target language, sometimes a single lexical item.
- Fixed and semi-fixed phrases
  e.g. *by and large*. These are normally handled within the entry of one of the lexical words in the phrase.
- Other phrasal idioms
  e.g. *raining cats and dogs*. The same is true of these. Chapter 7 deals with this question in more detail, with regard to the various ways in which they can be incorporated into the entry.

The other two types of MWE are more commonly afforded headword status in current English dictionaries:

- Compounds
  e.g. *civil servant*, *high school*, *police dog*. (Solid compounds – single word, non-hyphenated – appear as headwords in current dictionaries.) Of the types of compound discussed in §6.2.2.3, figurative compounds are the most likely to be accepted as headwords in present-day English dictionaries, with semi-figurative less likely, and functional in third place. The current tendency however is to make compounds into headwords when space permits. This is logical, given the 'unitary' status of compounds, but many users fail to find them as full headwords because they expect them to be tucked away in the entry for the first element. (Dictionary skills training is not yet given a high priority in educational establishments.)

[12] In this chapter we consider these only from the point of view of their headword status; how they are presented in the dictionary is discussed in Chapter 7.

[13] They are also useful for automatic word sense disambiguation in computational text handling, but that falls outside our remit here.

- Phrasal verbs
  e.g. *set about*, *come in for*, *look forward to*. Dictionaries for learners of English rely on their users knowing what phrasal verbs are. In such dictionaries, phrasal verbs may be given full headword status, but more commonly appear as secondary headwords, appended to the entry for the verb itself. Some modern monolingual dictionaries for native speakers now handle phrasal verbs in the same way. However, native speakers of English – even educated ones – have often no idea what a phrasal verb is, and fire off petulant letters to dictionary editors complaining of the absence of *set about* ('begin') or *come in for* ('receive') in their new dictionary, only to discover these lurking at the foot of the entry for *set* or *come*. (They rarely write again to apologize.)

6.4.1.4 *Vocabulary types* In designing a headword list for a particular dictionary, you need to make conscious choices about lexical items that do not form part of the 'unmarked' basic general vocabulary and may not even be known to educated native speakers. It helps to consider these expressions as belonging to various types of specialized vocabulary. Some dictionaries will include them all, some will be selective, some (pocket dictionaries, perhaps) may exclude all of them. The final decision will depend – as always – on the market, the user profile, and the cost of production. Figure 6.11 gives an overview of the choices.

Once you have decided to include any of these vocabulary types, the problem for each type is to produce a list of items to be included in the dictionary. This question is addressed in §6.4.3 below. When an indication of language type is given in a print dictionary, this is normally in the form of a *linguistic label*. (The use of such labels is discussed in §7.2.8).

Domains Few will dispute that there is a 'common core' of vocabulary that most adult native speakers know, at least well enough to understand the core words in context. This may be contrasted with an infinite number of sublanguages – the vocabulary of plumbers, brain surgeons, corpus linguists, bridge enthusiasts, and many hundreds of other such groups. Each of these subjects has its own 'in' vocabulary. Examples of such specialist vocabulary items are *tibia* (medicine), *fractal* (mathematics), *lien* (the law), *byte* (computing), *quark* (physics), and *gouache* (art).

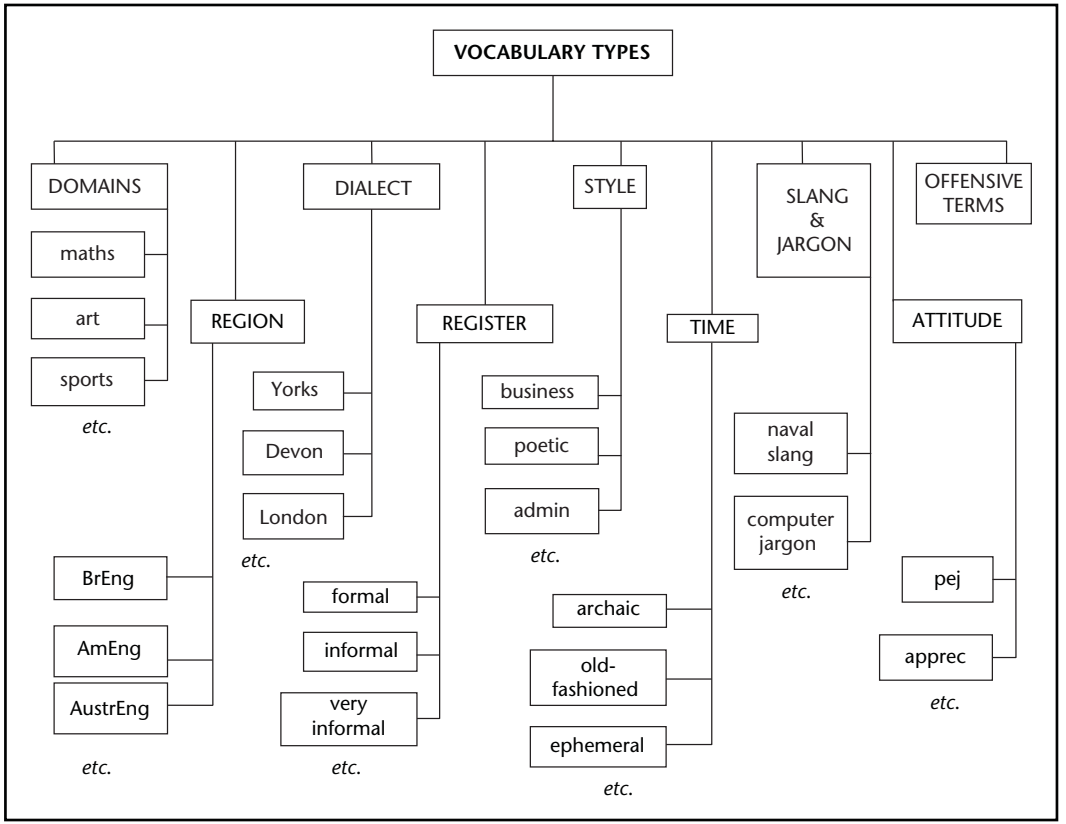


Fig 6.11 Vocabulary types and some of their realizations

It is normal practice in planning a dictionary to decide on which of the literally hundreds of domains should be given preference in the headword list. (Often there will be a target number of expressions from each of the favoured domains.) Planners of a schools dictionary might decide to include all the vocabulary of school subjects – from physics through sports to religious education and civics – likely to be encountered in secondary school. Planners of a bilingual dictionary with users from two linguistic communities will take care to include for example the vocabulary of local and national government and the military, of both cultures. An unabridged scholarly monolingual dictionary on historical principles, like the *OED*, may attempt to cover all known domains (although that is rare). But whatever the dictionary type, you need at this stage in the planning to draw up a list of possible domains, before considering what each individually will contribute to the headword list. It is possible to conceive of a totally 'flat' (non-hierarchical) list of domains: one that would include such items as medicine, administration, public relations, aerospace, education, fashion, and so on. However, it is more practicable to try to build a domain list with a certain hierarchical structure, so that instead of 'physics', 'chemistry', etc., you have 'science: physics', 'science: chemistry', and so on, as shown in Figure 6.12.

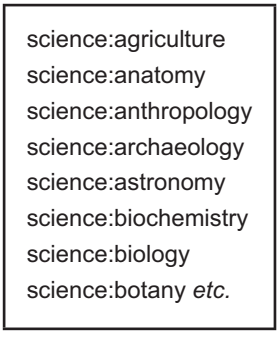


Fig 6.12 Partial listing of domains

This has two advantages:

- It makes it easier, when drawing up your domain list, to ensure that there are no glaring omissions. For instance, in this format it is simple to group all the sciences together and make sure that none is missing.
- It allows you to mark vocabulary items more accurately. Those which are common to several domains can receive the 'higher-level' domain

marker, so that items like *test tube* and *laboratory* may be labelled 'science' rather than 'physics, chemistry, biology' and so on, while those specifically belonging to lower-level domains carry the more specific labels, e.g. *metabolism* (biology), or *skeleton* (anatomy).

Region This refers to the varieties of a language found in countries where it is spoken as an official language, e.g. British English (*postcode*) and American English (*zipcode*). Nowadays an English dictionary is normally designed to be sold on both sides of the Atlantic, as well as probably in both hemispheres, so you would expect to include in the headword list a good number of items from American, Canadian, Australian, New Zealand, South African, Indian, etc. English. French dictionaries will be expected to cover not only metropolitan French, but Belgian, Swiss, Canadian, etc. French as well. Here again, the market (or user profile) drives the decisions.

Dialect This refers to non-standard words used in local areas and not outside them: e.g. for English, dialects include Yorkshire (*beck* 'stream') and Scots (*peely-wally* 'ill-looking'). Here again, much will depend on the user profile. The Chambers dictionary range, produced by the respected and long-established Edinburgh publishing house, is renowned for the high proportion of Scots words in its headword list, which makes it attractive to crossword and Scrabble addicts.

Register This refers to current expressions which are either more formal than the norm (*deeply indebted*), or more informal (*dead chuffed*). It is normal to have at least three 'levels' of formality: usually one above the 'unmarked' (perhaps 'formal', even sometimes 'official' or 'correct'), and two below it (some variation on words like 'informal', 'familiar', 'casual', 'relaxed', etc.). There is no absolute scale of formality, and maintaining consistency throughout the dictionary presents quite a challenge. What's more, levels of formality vary between regions in which the language is spoken.

Style This refers to expressions that are literary (*revels*), bureaucratic (*incentivization*), journalese (*romp*, *fashionista*), and so on. Here again, there are no absolute values on this scale, and each dictionary will choose to mark what is most useful for its users.

Time This refers to words which are not time-neutral: they may be archaic (*greensward*) or old-fashioned (*jolly* in the sense of 'very'), or ephemeral (*cool* in the sense of 'excellent'), which is much more difficult to detect. Here

again, there are no absolute values, and what you call the points in the time scale depends on your user profile. What is 'archaic' for one dictionary is 'obsolete' for another; what is 'old-fashioned' for one is 'obsolescent' for another, and so on.

Slang and jargon This refers to non-standard expressions used within specific groups of people (for instance *avast* in naval slang) or sharing the same interest (e.g. *plug and play* in computer jargon). Here again, there are no absolutes, and dictionaries differ widely in their approach to slang- and jargon-marking. Slang is further down the informality scale than jargon, which is often used among technical experts on quite formal business occasions. Both slang and jargon normally need to be accompanied by another label indicating which group of people uses the term.

Attitude This group indicates the attitude of the speaker or writer towards what is being discussed. Typical attitude labels are *pejorative* or *derogatory* (indicating disapproval) and *appreciative* (indicating approval). If the meaning of the word (such as *miserly* or *cruel*) is clearly derogatory, this is made clear in the definition of the word and a label is redundant; however if the word has an unmarked sense, but can be used to indicate disapproval (as *conventional* in *He's very conventional*) then this usage is normally labelled pejorative. The label *ironic* is helpful when the intention of the speaker has to be clarified.

Offensive terms This group covers racist terms (*mick*, *jock*) and others including swear words which may give offence and/or are taboo. Normal practice is not to specify *why* an item is offensive, simply that this is so. This is a particularly dangerous area of language and care must be taken when deciding which (if any) offensive terms should be included as headwords.

6.4.2 Proper names

At a very early stage in dictionary planning, you have to decide whether or not to include proper names ('encyclopedic material') in the headword list. Here are some points to be aware of when considering these items as potential headwords:

- Formerly, proper names were usually excluded from the headword list, and sometimes corralled into a list at the end of the book.

- Nowadays most reasonably sized English dictionaries include them as headwords.
- Even dictionaries which exclude encyclopedic entries will make honourable exceptions for proper names with metonymic force (*White House*, *Downing Street*) and cultural entities (*Big Brother*, *Father Christmas*).
- There are difficult boundary issues: compare the fully encyclopedic *Alice in Wonderland* (the book by Lewis Carroll) with the adjectival use of *Alice-in-Wonderland* (*the Alice-in-Wonderland world of European agriculture*). Here it is no longer simply an encyclopedic item, for it has wholly lexical functions (the comparative and superlative are possible, with *more* and *most*) – indeed it behaves exactly like its synonym, the standard lexical adjective *topsy-turvy*.
- Proper names come in two kinds: closed sets (such as the twelve apostles or the planets of the solar system) where the all-or-nothing approach applies, and – much more common – open sets which impose difficult choices.
- The actual decision about what to include and what to exclude will depend on how important the various classes of proper name are for the dictionary's intended market. (Here again, the user profile comes into play.)

6.4.2.1 *Place names* Names of places may be conveniently divided into some main types, although the list below is not exhaustive by any means and for a specific dictionary other classes may be identified.

- 'Basic' names: the oceans, continents, countries, states, provinces, counties, other administrative divisions familiar to the dictionary's intended market.
- Capital and non-capital cities: e.g. *London*, *Glasgow*, *Washington*, *New York*.
- Major geographic features: seas, lakes, rivers, mountains, regions, islands, and others.
- Metonyms: names of places used to denote the people who work there, e.g. *Whitehall*, *the Pentagon*.
- Famous places and buildings: major battlefields, important buildings, major airports, sites of religious significance, e.g. in that order *Waterloo*, *the Tower of London*, *Heathrow*, *Mecca.*

- Extra-terrestrial objects: planets, stars, constellations, galaxies, moons and satellites, comets, etc.
- Imaginary, biblical, or mythological places: e.g. *the Garden of Eden*, *Lilliput*, *Hades*, *Armageddon*.
- Nicknames for places: e.g. *the Big Apple* (New York), *the Square Mile* (the City of London).

6.4.2.2 *Personal names* These are normally subdivided into generic and specific names, together with the related adjectives of the latter group.

- Generic names
  These include both first names (*John*, *Mary*) and surnames (*Smith*, *McGregor*), although in practice few dictionaries include surnames, and first names normally figure as headwords only in bilingual dictionaries.
- People's names
  These include those who figure here simply on grounds of renown (*Beethoven*, *King Lear*), and those whose name carries certain connotations (as in *he's a real little Hitler*, or *he's the office Casanova*). The latter type is obviously more important for dictionary purposes. This class subdivides into:
  – real people: including famous people alive today and historical figures such as writers (*Samuel Johnson*), artists (*Michelangelo*), musicians (*Mozart*), military and political figures (*William the Conqueror*, *Alexander the Great*, *Abraham Lincoln*), and so on.
  – others: including religious (*Christ*, *Muhammad*, *Buddha*), biblical (*Solomon*, *Mary Magdalen*), mythological (*Jupiter*, *Gaia*), semi-historical (*Robin Hood*, *Lady Macbeth*), and purely fictional characters (*Othello*, *Jane Eyre*).
  – related adjectives: such as *Dickensian, Christian*.
- nationalities
  – and also names for natives of cities, counties, regions, etc. (nouns and adjectives), e.g. *French*, *American*, *Chinese*, *Mancunian*, *New Yorker*.
- names of ethnic groups etc.
  – both nouns and adjectives, members of ethnic groups (*African-American*, *Arab*), Native American peoples (*Apache*), ancient peoples (*Minoan*, *Celt*), and so on.

6.4.2.3 *Other names* The wealth of other proper names that could be included in the headword list may usefully be divided into a few broad classes. Specific dictionaries may require others at the headword planning stage.

- Festivals, ceremonies: e.g. *Christmas*, *Ramadan*, *Thanksgiving*, *Fourth of July*, *Bar-mitzvah.*
- Organizations: such as political parties (*New Labour*, *Republican*), institutions (*Congress*, *Parliament*, *Appeal Court*, *European Central Bank*), government departments (*Department of Trade and Industry*, *Defense Department*), other official or semi-official agencies (*NASA*, *National Guard*, *UNESCO*), and clubs and other social groupings (*Freemasons*, *Ivy League*).
- Languages: national and major regional languages, major language groups/families, e.g. *Dutch*, *Mandarin*, *Flemish*, *Arabic*, *Hindi*, *Sanskrit*, *Basque*, *Indo-European*.
- Trademarks: for products and services, e.g. *Band Aid*, *BlackBerry*, *Frisbee*, *iPhone*, *Yellow Pages*.
- Beliefs and religions, and their adherents: (nouns and adjectives) e.g. *Church of England*, *Muslim*, *Judaism*, *Taoism*, *Baptist*, *Marxism*, *Freudian*, *Jain*, *Scientology*, *Zen.*
- Miscellaneous: These would be included on the basis of frequency and local high profile, and would cover such disparate items as *Holy Grail*, *Nikkei Index*, *Academy Awards*, *Holocaust*, *Olympic Games.*

6.4.3 Deciding the specifics

In §6.4, the focus has been on deciding the *types* of expression to be given headword status. Once these decisions have been made, the headword list is drawn up and the editing starts. However, the initial headword list is never set in stone, and 'in or out' decisions always have to be made along the way. For instance, *corruptibility* is a 'common core' word (not slang, hardly even formal, not British only, not domain-specific, etc.) but should it go into the dictionary or not? What guidelines can we use when faced with such a dilemma? It is good to take the following factors into account:

- the item's corpus frequency
  The rarer it is in the corpus (if your corpus is fairly representative), then the less likely your users are to be looking it up.

- its 'profile' or salience
  How familiar it is to the dictionary users, and how widely known: for instance, learners' dictionaries despite their relatively small headword lists will include a lot of linguistic terms (e.g. *affricate*, *agglutination*, *alveolar*), which would clearly fail any frequency test, and indeed most British adults wouldn't know them, but they happen to be very relevant to the specific user-group with its high proportion of teachers of English.
- its possible translation (for bilingual dictionaries)
  Gerunds are a good case in point here: words like *running* (as in *the running of the company proved too much for him*) are so easily understood from their verb root that they are very often omitted from the headword list of monolingual dictionaries; however these are often difficult to translate, and so appear more frequently as headwords in bilingual dictionaries. The same is true for specific proper names: their inclusion may depend on whether or not there is a target-language name for the place, person, etc. (The *White House* is *la Maison Blanche* in French.)
- its additional meanings or connotations
  This factor is particularly relevant for proper names: for instance, *Parliament* has a meaning beyond the physical building, and *Orwellian* means more than simply 'relating to Orwell'. Such items demand inclusion in a dictionary of any type.

6.5 Organizing the headword list

Of the three matters touched upon in this section, only the third, homograph headwords, is of any real importance to lexicographers.

6.5.1 Alphabetization[14]

Deciding on the alphabetical order of the headword list is a quagmire, but one which poses few real problems for editors of current English dictionaries. This is principally because every publishing house has its own policy, enshrined in dictionaries already in print. For that reason we will not spend

[14] The discussion here relates only to print dictionaries; alphabetization holds no fears for editors or users of electronic dictionaries.

much time on this, particularly since the Style Guide will give explicit guidance on what goes where in the entries you are writing, and the software will see to the ordering. However, here are some factors for dictionary planners to take into account when devising an alphabetization policy:

- If all the headwords are single words, there is no alphabetization problem.
- If the headword list contains multiword items, then problems arise. These are discussed more fully in §6.2.2 above.
- Essentially, there are two options: to alphabetize word by word, or letter by letter.
- In a word-by-word list, the space between words takes precedence, hyphens normally come next, and letters come last. The result of this is that *set piece* will come before *set-up* and they both precede *setback.*
- In a letter-by-letter list, spaces and hyphens are disregarded, and the words would appear in this order: *setback*, *set piece*, *set-up*.
- Dictionaries therefore tend to alphabetize letter by letter, ignoring capitalization (whereas for instance British telephone directories place capitals before lowercase letters, so that *BBC* will come before *Barnet* in their listing).

6.5.2 Syllabification

Syllabification is the marking of syllables within the headwords in the dictionary, by means of a centred period, or a vertical line, or other similar device, thus:

bread·fruit *or* **bread|fruit**.

Dictionary designers need to decide whether or not to include this feature, which is becoming ever rarer in English dictionaries. Not too many people need it nowadays, since word processing programs know where to insert word breaks (which is the main point of syllabified headwords). Also, syllabification marks are distracting for readers in their attempt to find the word they're looking up.

6.5.3 Homographs

Homograph headwords are a common feature of dictionaries, and before work starts a decision has to be made on whether or not to allow these,

and, if so, on the principles that must be applied by the editorial team. Homograph headwords consist of two or more identically written words, each given its unique number and treated as a discrete entity in its own right. Only the presence of the superior number (see Figure 6.13) indicates that there is more than one entry for that word form.

bear[1] (bɛə) *vb.* **bears, bearing, bore, borne.** (*mainly tr.*) **1.** to support or hold up; sustain. **2.** to bring or convey *to bear gifts.* **3.** to take, accept or assume the responsibility of: *to bear an expense.* **4.** *(past participle* **born** in passive use) …

bear[2] (bɛə) *n. pl.* **bears** *or* **bear. 1.** any platigrade mammal of the family *Ursidae:* order *Carnivora* (carnivores). Bears are typically massive omnivorous animals with a large head, a long shaggy coat, and strong claws. …

Fig 6.13 Homograph headwords for *bear* in *CED-5* (2000)

There are homographs of various types: the Style Guide of a dictionary with homograph headwords must be clear on what is involved. According to the classical definition, the term 'homograph' is vague, denoting a word with identical spelling to another word, but different meaning, etymology, and/or pronunciation. Below are some of the criteria currently used in lexicography to decide whether there should be one entry or more.

(1) Same spelling; different meaning and etymology
e.g. **tear**[1] (tɪə) (from weeping) and **tear**[2] (tɛə) (in paper, cloth)
bear[1] (bɛə) 'animal' and **bear**[2] (bɛə) 'carry, tolerate, support'
Pronunciation is irrelevant here. Etymology is an easy rule for lexicographers to apply. Historical and scholarly dictionaries, and dictionaries developed from them, normally follow this approach, but nowadays the needs and abilities of the dictionary user are given much more weight. Since very few native speakers (far less language-learners) know the origin of words, for many modern dictionaries the etymology of homographs is not a consideration.

(2) Same spelling; different meaning and pronunciation
e.g. **tear**[1] (tɪə) (from weeping) and **tear**[2] (tɛə) (in paper, cloth)
In most current dictionaries, the simple difference in sound will automatically generate homograph headwords, so that the appropriate pronunciation may be given for each.

(3) Same spelling and pronunciation; different meaning and capitalization

e.g. **may**[1] (meɪ) (modal verb) and **May**[2] (meɪ) (month)

pole (pəʊl) 'long stick' and **Pole** (pəʊl) 'native of Poland'

Most modern dictionaries will consider these as separate headwords.

(4) Same spelling and pronunciation; different meaning

e.g. **bank**[1] (bæŋk) 'edge of river' and **bank**[2] (bæŋk) 'financial institution'

bear[1] (bɛə) 'animal' and **bear**[2] (bɛə) 'carry, tolerate, support'

Words like these are often given homograph status, but difference in meaning is a grey area, and there are no clear criteria for lexicographers to apply (and of course the user looking up a word often does not know its meaning). For that reason, a number of dictionaries have no homograph headwords at all and put all the senses under the same headword.

(5) Same word (spelling, meaning and pronunciation); different word-class

e.g. **hit**[1] (hɪt) verb 'to strike' and **hit**[1] (hɪt) noun 'a blow'

Most monolingual dictionaries for learners follow the homograph path here, on the basis that the user may very well be able to identify the wordclass of an otherwise unknown word; this is a simple rule for lexicographers to apply consistently. Other dictionaries vary, according to the users they expect, or their own publishing conventions.

6.6 Types of entry

Once you've decided what kinds of word are to be headwords in your dictionary, you have to consider the varieties of entry structure needed if the information about these words is to be presented clearly. What you want to say about verbs like *settle* and *decrease* will be more easily understood if the information is presented differently from the structures of entries for prepositions like *in* and *with*, or abbreviations like *EU* and *WMD*. The microstructure of the entry (its components and their content) is discussed in Chapter 7. Here it is enough to outline four principal entry types, used respectively for lexical words, abbreviations, grammatical words, and encyclopedic words.

6.6.1 Standard lexical entry

Lexical words carry a full definable meaning and their contribution to a sentence is principally to add meaning (although of course they make a

necessary contribution to the syntax as well). Most of the lexicographical work we discuss focuses on this type of entry, so at this point it is enough to say that it holds the bulk of the headwords in the dictionary, including:

- **nouns** *head*, *saucepan*, *capitalism* . . .
- **verbs** *fascinate*, *meet*, *see*, *come* . . .
- **adjectives** *happy*, *interested*, *good* . . .
- **adverbs** *now*, *well*, *illogically* . . .
- **interjections** *oh! ouch!* . . .

Figure 6.14 shows the treatment given to the word *paramilitary* in three different standard-sized dictionaries: a monolingual English dictionary for native speakers, the *Chambers Dictionary* (*CD*), published in 1993; a one-volume learners' dictionary (*MED-2*, 2007); and a bilingual English-French dictionary (*CRFD-5*, 1998). Each of the entries shown here, though very short, may be viewed as typical of lexical entries in their particular type of dictionary.

paramilitary *par-ə-mil'i-tər-i*, *adj* on military lines and intended to supplement the strictly military; organized as a military force, *esp* when engaged in active rebellion. – Also *n.* **[para-**[1] and **military]**

Chambers Dictionary

paramilitary[1] /ˌpærəˈmɪlɪt(ə)rɪ/ adj **1** organized and operating like an army: *an outlawed paramilitary group* **2** providing support for an army or other military organization: *a paramilitary police force*

paramilitary[2] /ˌpærəˈmɪlɪt(ə)rɪ/ noun [C usually plural] a member of a paramilitary organization

MED-2 (2007)

paramilitary/ ˌpærəˈmɪlɪtərɪ/ **1** ADJ [*organization, group, operation*] paramilitaire. **2** N (= *member*) membre *m* d'une force paramilitaire. **the paramilitaries** les forces *fpl* paramilitaires. **3 the paramilitary** NPL (= *organizations*) les forces *fpl* paramilitaires.

CRFD-5 (1998)

Fig 6.14 Three different lexical entries for *paramilitary*

These three lexical entries have a good deal in common, but they differ according to the needs and expectations of their respective target users. Here are some points of similarity and difference:

- All three dictionaries aim to give their users much the same basic information, in as small a space as possible.
- The 'basic' information consists of:
  - the headword *paramilitary*
  - its pronunciation

  - the fact that this word is an adjective
  - an explanation of meaning of the adjective LU
  - the fact that this word is also a noun
  - an explanation of meaning of the noun LU.
- The way the dictionaries differ includes the following:
  - In showing the pronunciation: in the International Phonetic Alphabet (IPA) in the case of *MED* and *CRFD*, whose readers are probably language students; in Chambers' own re-spelling system for *CD*, whose readers probably don't know IPA.
  - In explaining the meaning of the adjective LU: definitions in the case of the two monolingual dictionaries, translation in the bilingual one.
  - The native speakers' dictionary, *CD*, has no definition of the noun LU; their readers are expected to understand this use in context and not to use it incorrectly; *MED*, whose readers are not native speakers, is at pains to explain that *a paramilitary* is a person, not an organization, but gives no examples of the noun use; *CRFD*, catering for both English and French native speakers gives both the French translation of the headword as a noun, and its gender.
  - The monolingual dictionary for language learners, *MED*, gives two examples of the word's adjectival use; the bilingual dictionary relies on the 'sense indicators' *organization*, *group*, and *operation* to tell the English speaker the type of noun the French adjective can modify; the monolingual dictionary for native speakers rightly does not think it worth expanding the entry for this word by including unnecessary examples of use.
  - *CD* and *MED* show two senses for the adjective, while this is unnecessary in a bilingual dictionary where the same French word *paramilitaire* translates both.
  - *CRFD* also includes the noun use of the headword in its singular form as a collective plural *the paramilitary*.
  - *MED* describes the noun as a count noun ('C') and notes that it is usually found in the plural.

In addition to lexical words, some very common abbreviations are treated more like lexical entries than abbreviation entries when they constitute the most usual way of referring to the concept in question. (An example of this is the entry for *UNESCO* in Figure 6.15.)

UNESCO (juːˈnɛskəʊ) *n acronym for* United Nations Educational, Scientific, and Cultural Organization: an agency of the United Nations that sponsors programmes to promote education, communication, the arts, etc.

i.e. *abbrev. for* id est. [Latin: that is (to say); in other words]

EU *abbrev. for* European Union.

European Union *n* an economic and political grouping that was formed (1993) to extend the European Community by adding common foreign and security policies to the single market. […] Abbrev.: **EU**

Fig 6.15 Abbreviations and their full forms in *CED-5* (2000)

6.6.2 Abbreviation entry

Entries for abbreviations, for instance *EU* and *i.e.*, do not carry much information, for they have rarely more than one sense and rarely belong to more than one wordclass. In this, they resemble proper-name entries. However, abbreviation entries must explicitly cross-reference the full form. For reasons of space, the information is normally given only once, either at the abbreviation entry or the full form entry.

Figure 6.15 shows various ways in which one dictionary handles this type of word. *UNESCO* is never referred to by its full form, and is pronounced as a word. The dictionary consequently gives it a full lexical entry (headword, pronunciation, wordclass, definition) plus the information relating to its full form. The other abbreviations, *i.e.* and *EU*, are pronounced as a group of letters and do not – in this dictionary – merit pronunciation information. Note the different treatment given to these two abbreviations: no one would ever look up *id est* in an English dictionary, and so the meaning of *i.e.* is shown as an explicated translation of the Latin words, while *European Union* is frequently heard and for that reason the entry for its abbreviation *EU* is a simple cross-reference.

6.6.3 Grammatical word entry

As distinct from lexical words, which carry meaning, the principal role of grammatical words (introduced in §6.2.1.1) is to perform a *function* in the sentence. Devising the most useful way of presenting such information forms part of the work of the dictionary pilot study, and must take account not only of the type of dictionary but of the skills and needs of the expected user. The Style Guide must be specific on the words to be considered ‘grammatical words’, and of course on how to handle them.

Because each wordclass (prepositions, conjunctions, determiners, etc.) and indeed subclass perform different functions, there is no set structure for grammatical entries. The entry in Figure 6.16 shows what can be done to describe function rather than meaning.

because (…) *conj.* For the reason that; since [ME.]
Usage *Because* is the most direct of the conjunctions used to express cause or reason. It is used to state an immediate and explicit cause: *He stayed behind because he was ill. Since, as,* and *for* are all less direct than *because;* they often express the speaker's or writer's view of the causal relation between circumstances or events. The clause introduced by *since* most frequently comes first in the sentence: *Since he stayed behind, he must have been ill* […]

Fig 6.16 Part of grammatical word entry for *because*

This example is drawn from the *AHD-2* (1985), a collegiate dictionary for native speakers. After a brief 'substitutable' definition ('for the reason that, since'), the entry consists of a very long usage note contrasting the way *because* is used with the use of other conjunctions such as *since* and *as*. That is to say, the *function* of the headword is discussed and exemplified in this entry, and since that function may also be carried out by other conjunctions information about these words is included in the entry too.

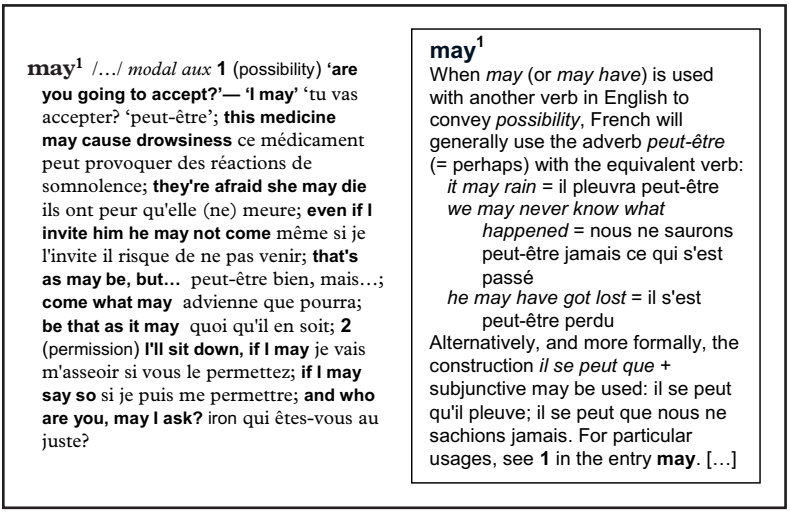
may[1] /…/ *modal v* (*neg* **may not**; *rare short form* **mayn't** /ˈmeɪənt/; *pt* **might** /maɪt/; *neg* **might not**, *rare short form* **mightn't** /ˈmaɪtnt/ **1** (*rather fml*) (indicating permission) *You may come if you wish.* ○ *May I come in?* ○ *That was a delicious meal if I may say so.* ○ *Passengers may cross by the footbridge.* → note. **2(a)** (indicating possibility): *This coat may be Peter's.* ○ *That may or may not be true.* […]

NOTE Using **may** (negative **may not**) is a polite and fairly formal way of asking for, giving or refusing permission: *May I borrow your newspaper?* […] Children often use **may** when speaking to adults: '*Please may I leave the table?*' […] **Can** and **cannot** (or **can't**) are used to give and refuse permission: *You can come with us if you want to.* ○ *You can't leave your bike there.* **Could** is a neutral and polite word, used mostly in requests […]

Fig 6.17 Part of the entry for *may* and part of its usage note in *OALD-5* (1995)

Even more than native-speaker dictionaries, learners' dictionaries (monolingual and bilingual) rely heavily on usage notes when dealing with grammatical words. Figure 6.17 shows how the modal verb *may* is handled in a monolingual learners' dictionary and Figure 6.18 illustrates its treatment

in a large one-volume English-French bilingual dictionary. In both types of learners' dictionaries, the *function* which the modal verb *may* carries out in the language is explained at considerable length: it can express permission, possibility, and so on. Both dictionaries concentrate on the various nuanced ways in which permission and possibility may be expressed: the monolingual dictionary focuses on English for its user-learners of any nationality, while the bilingual focuses on French for its English-speaking user-learners.



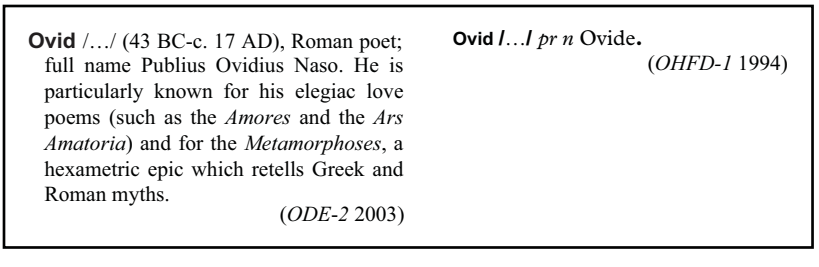
may[1] /…/ *modal aux* **1** (possibility) **'are you going to accept?'— 'I may'** 'tu vas accepter? 'peut-être'; **this medicine may cause drowsiness** ce médicament peut provoquer des réactions de somnolence; **they're afraid she may die** ils ont peur qu'elle (ne) meure; **even if I invite him he may not come** même si je l'invite il risque de ne pas venir; **that's as may be, but…** peut-être bien, mais…; **come what may** advienne que pourra; **be that as it may** quoi qu'il en soit; **2** (permission) **I'll sit down, if I may** je vais m'asseoir si vous le permettez; **if I may say so** si je puis me permettre; **and who are you, may I ask?** iron qui êtes-vous au juste?

may[1]
When *may* (or *may have*) is used with another verb in English to convey *possibility*, French will generally use the adverb *peut-être* (= perhaps) with the equivalent verb:
it may rain = il pleuvra peut-être
we may never know what happened = nous ne saurons peut-être jamais ce qui s'est passé
he may have got lost = il s'est peut-être perdu
Alternatively, and more formally, the construction *il se peut que* + subjunctive may be used: il se peut qu'il pleuve; il se peut que nous ne sachions jamais. For particular usages, see **1** in the entry **may**. […]

Fig 6.18 Entry for *may* and part of its usage note in *OHFD-1* (1994)

6.6.4 Encyclopedic entry

Entry types for proper names are necessarily slimmer than lexical and grammatical entries. Monolingual dictionaries vary considerably in the amount of information they give (particularly when the name has cultural connotations, as *Napoleon* or *Hitler*), while often for bilingual dictionaries a simple translation is enough, as Figure 6.19 shows.



Ovid /…/ (43 BC-c. 17 AD), Roman poet; full name Publius Ovidius Naso. He is particularly known for his elegiac love poems (such as the *Amores* and the *Ars Amatoria*) and for the *Metamorphoses*, a hexametric epic which retells Greek and Roman myths.
(*ODE-2* 2003)

Ovid */…/ pr n* Ovide.
(*OHFD-1* 1994)

Fig 6.19 *Ovid* entry in monolingual and bilingual dictionaries

Exercise

Choose a dictionary you are familiar with, one that includes proper names in its headword list. Then:

- Take 30 pages and list the proper names in it.
- Classify them along the lines shown in §6.4.2.
- Describe the editorial policy on proper names that lies behind this list.
- Given the type of dictionary and the people who are likely to use it, is this policy a sensible one? Can you improve it?

Reading

Recommended reading

Atkins 1993; Sinclair 1991 (esp. chapters 5 and 8); Cowie 1994, 1998.

Further reading on related topics

Aitchison 2003; Algeo 1993; Atkins and Grundy 2006; Kilgarriff 1994; McArthur 1986.

How words work with other words: Benson 1990; Čermak 2006; Coffey 2006; Cowie 1981, 1999a; Cowie and Howarth 1996; Fontenelle 1992, 1996; Grossmann and Tutin (eds.) 2003; Hanks 2004b; Hanks, Urbschat, and Gehweiler 2006; Hausmann 1989, 1991; Heid 1994, 1998; Kilgarriff 2006b; Mel'čuk 1988; Moon 1988, 1992, 1996, 1998; Siepmann 2005, 2006; van der Meer 1998.

Websites

Phrases in English: http://pie.usna.edu/ : allows users to perform a range of searches for multiword items, using a powerful search engine linked to the BNC.

7

Planning the entry

7.1 Preliminaries

A dictionary entry is designed to present facts as clearly as possible. It must take account of the needs and skills of the kind of people expected to be using the dictionary, based on the user profile (§2.3.1). A consistent approach to the structure and content of the entries is essential, or users will simply give up.

Figure 7.1 gives an outline of this chapter, which surveys both the form and the content of the dictionary entry. We start by introducing in turn each of the principal components of a standard entry, then discuss some of the alternative ways in which these may be assembled into an entry.

Consistency of approach is ensured by the Style Guide (§4.4), which must give clear instructions not only on the type of information to include in the particular dictionary, but also on how to set out that information, including *inter alia* guidance on the following:

- which of the many possible entry components are to be used in the dictionary, the types of information that each may hold, and how it should be presented: this forms the focus of §7.2 below
- the various decisions that you will have to make when compiling the entry; these are discussed in §7.3, and, among many other things, will cover the following:

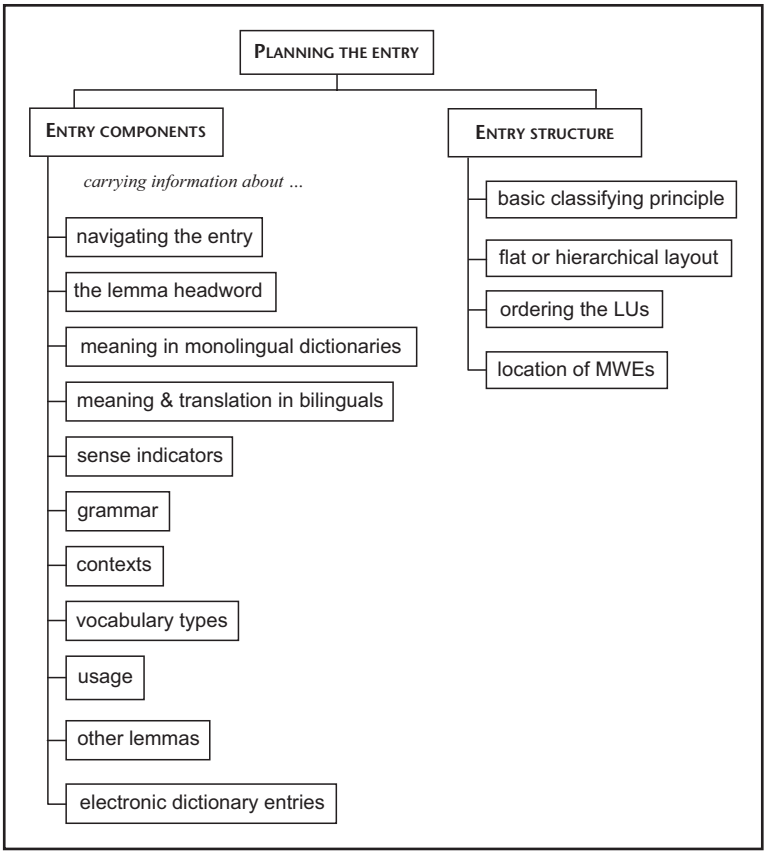


Fig 7.1 Contents of this chapter

- the basis (meaning or wordclass) on which the whole entry is divided into manageable sections
- how to order the various sections within the entry
- whether or not these sections should be hierarchical (i.e. include subsections, etc.)
- how to handle the various types of multiword expressions (MWEs)[1]
- whether or not to give curtailed information in the form of secondary headwords, run-ons, etc. (these terms are explained in §7.2.10).

The time is past when the lexicographer, using only her own judgment and knowledge of the Style Guide, could decide what goes where. Many

[1] MWEs described in §6.2.2, and briefly discussed in §7.2.7.1.

dictionary publishers now use a *dictionary writing system* or *DWS* (§ 4.3.2), a very complex piece of software that takes the dictionary text all the way through from the editors' computers to the printed book and/or electronic dictionary. One of its functions is to maintain consistency over many lexicographers throughout a lengthy editorial period, by ensuring that the components are ordered in a legitimate way and that the contents of many of the components are valid.[2]

7.2 Information in the various entry components

Decisions in designing the microstructure relate to the separate pieces which go to make up the dictionary entry, and their relationship one to another. The purpose of this section is simply to introduce the principal entry components and illustrate them. How you use them to build up an entry is discussed in Chapter 10 (monolingual dictionaries) and Chapter 12 (bilingual). The components introduced here are to be found in most print and electronic dictionaries.[3] Moreover, new entry components are appearing in each generation of electronic dictionaries, since the space constraints of print do not apply there. Some of these new components are introduced in section §7.2.11.

Look at any dictionary entry and you will find many of the components we describe in this section. They won't all be there, since the way information is presented depends not only on dictionary type (§2.2), but also on the properties of the language under review. For instance, if it's a monolingual dictionary, it will have no translations, and dictionaries of Italian, Finnish, and other languages written phonetically rarely include regular pronunciations. You may also find in your entry some component that is not mentioned here.

Before we look at the various components in detail, one further point: the components chosen for a dictionary during the design phase cannot

[2] The components whose contents are open to software control usually include: *section/subsection marker*, *pronunciation*, *frequency marker*, *wordclass marker*, *valency*, and the vocabulary type labels. For each of these components there is a pre-determined list of valid material which the software will accept. These lists are drawn up during the design stage of the dictionary, included in the Style Guide, and often available to the lexicographers in the form of a pull-down menu in the text input screen.

[3] The list is not exhaustive, however, and in particular we do not attempt to cover everything found in a large scholarly work such as the *Oxford English Dictionary*.

be assembled in any order that may take an editor's fancy. Every dictionary entry has its own 'syntax' which controls where the various components may be inserted (there are usually several valid locations for each component). Only linguistic labels (§7.2.8) may appear almost anywhere in an entry, and even then their scope and the way they combine with other labels is subject to the entry syntax. Moreover, the various types of entries described in §6.6 (lexical, grammatical, etc.) all require a different configuration of components.

We shall now briefly introduce the commonest of these basic building blocks of any dictionary entry, focusing on the type of information that each holds. We first illustrate the components under review, then discuss them briefly. The legitimate content of some of these components is very restricted, e.g. those holding grammatical information (§7.2.6) and the linguistic labels (§7.2.8), and will be prescribed in the Style Guide. The content of most of them, however, will depend on the skill of the lexicographers writing the entry. Chapters 10 and 12 deal with the way in which the entry components discussed here are used to build a complete entry.

7.2.1 Navigating the entry

The role of the components introduced in this section (and illustrated in Figure 7.2) is to structure the entry and to help users find their way around

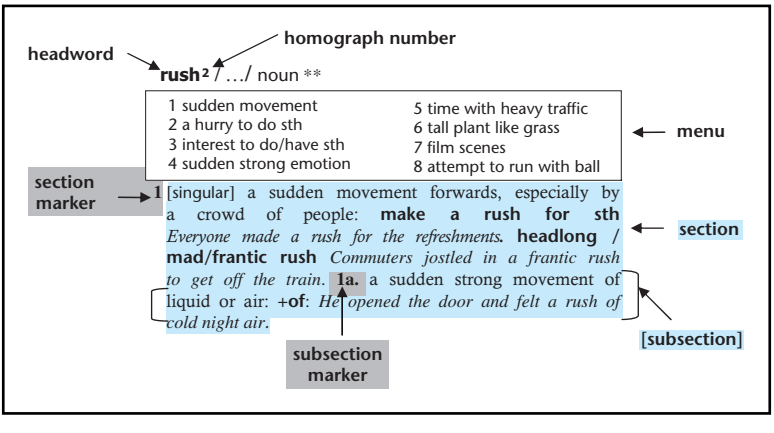


Fig 7.2 Navigational components in *MED-2* (2007)

all the information it contains.[4] You rarely have a free hand in deciding how to use this small set of components, as this is normally strictly controlled by the software or the Style Guide, and you will be offered a list of items to choose from. Only a 'menu' leaves the lexicographer with a choice of content, since what you put there depends on the senses of the headword.

7.2.1.1 *Headword* This component holds the lexical form of the headword, showing how it is written, whether in a single word, a hyphenated word, or in several words (the various options are explained in §6.4). Many dictionaries also show wordbreaks, by means of a centred dot or other marker indicating where the headword may be split at the end of a line. These are becoming rarer because of automatic spellcheckers.

7.2.1.2 *Homograph number* The presence of this component (usually in the form of a superscript number) indicates that the headword is one of two or more homographs, and that the same word appears as a headword again in an adjacent entry. The options here are fully explained in §6.5.3.

7.2.1.3 *Menu* Lexical units (LUs) are the numbered divisions of a dictionary entry, commonly known as the headword's 'senses'. The 'menu' (a brief set of mnemonics, appearing at the top of an entry, for the LUs in the entry) is a late-comer to the list of components, though dictionaries produced in Japan and Korea (among others) have used this device for some time. It currently appears mainly in dictionaries for non-native speakers, and is designed to streamline the difficult task of locating the 'right' part of a complex entry. (The same function is also catered for, in a slightly different way, by 'signposts', which we describe later: §7.2.5.2.) The *MED* menu shown in Figure 7.2 is a good example: the 'definitions' are kept as brief as is consistent with intelligibility. In many cases, they take the form of a telegraphic version of the main definition, but they can also work on the basis of contextual or collocational 'hints': so for example, the *MED* menu for *service* includes one item that simply reads 'in tennis etc', while the sense of the verb *pitch* that describes the movement of planes or ships is indicated by a menu item saying 'about ship/aircraft'.

[4] Another group of components also function as navigation aids: these are the 'sense indicators', and are introduced in section §7.2.5.

➔ For menu items, remember to choose simple words which the user is likely to understand.

7.2.1.4 *Section/subsection* A section (or subsection) holds the facts relating to one LU. Whether a dictionary entry consists simply of a number of sections, or of sections and subsections and even subsubsections, depends on the entry structure prescribed for that dictionary. The options here are discussed in §7.3.2.

7.2.1.5 *Section/subsection marker* Most commonly, numbers and/or letters indicate the start of a new section or subsection, as in Figure 7.2. The less common symbol is illustrated in the *ODE* entry in Figure 7.3, where ▶ is used to mark a new wordclass section.

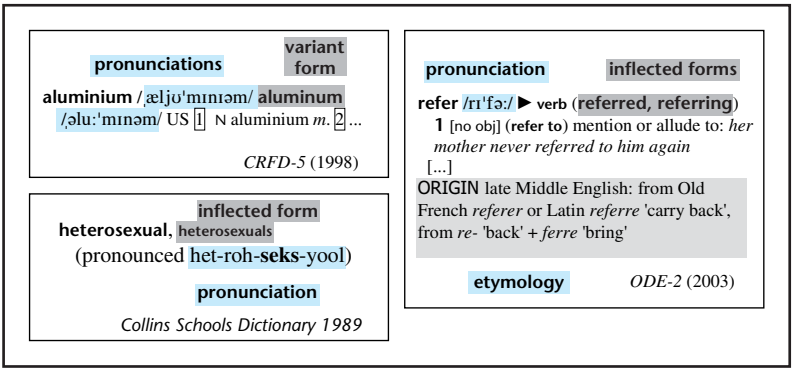


Fig 7.3 Headword-related components in three dictionaries

7.2.2 The lemma headword

A polysemous word is a *lemma* containing several *LUs*; it can belong to one or more than one wordclass. A monosemous word belonging to a single wordclass is both a *lemma* and an *LU*. The components introduced in this section are principally used to carry information about the headword (lemma), although they can also appear within an LU, with information that refers only to that particular LU. (The bulk of the components carry information about the LU, and are discussed in §§7.2.3–8 inclusive.) These headword-oriented components are *pronunciation*, *variant form*, *frequency marker*, *inflected form*, and *etymology*. Some or all of these may be inserted in a semi-automated process distinct from the entry-compiling.

Box 7.1

The International Phonetic Alphabet is internationally recognized, and an IPA transcription allows the dictionary to include:

- the sounds of the language e.g. ˌəluːˈmɪnəm
- vowel length: indicated by the colon following the 'u' in ˌəluːˈmɪnəm
- stress: in ˌəluːˈmɪnəm the subscript and superscript dashes indicate secondary and primary stress respectively
- other language-specific features, e.g. tones in tone languages.

7.2.2.1 *Pronunciation* The most common way of showing how a word is pronounced[5] is to use the *International Phonetic Alphabet* (*IPA*), as shown in Figure 7.3, in the entries from *CRFD* and *ODE*. Where the user cannot be expected to know IPA, as for instance in dictionaries for school students, a re-writing system may indicate how the headword is pronounced. This is illustrated in the entry from *Collins Schools Dictionary* in Figure 7.3. Often only difficult words are treated in this way. The disadvantage is that this system suits speakers of educated southern English more than those whose pronunciation is Scottish or Texan or Australian, etc., although it does give everyone an approximate idea of the pronunciation.

7.2.2.2 *Variant form* This component shows an alternative spelling or slight variation in the form of this word.
➔ Remember that the variant form may need a label, like *(US)* and *(Brit)* shown in the *CRFD* entry in Figure 7.3.

7.2.2.3 *Frequency marker* This is a relatively new component, which depends on access to a large corpus. It reflects the frequency of the headword in the corpus (usually calculated as so many occurrences per million words) relative to the other words of the language. The frequency marker, expressed in numbers, symbols, and/or abbreviations, is used mainly in learners' dictionaries to give students and teachers an idea of a word's relative importance (and how far it is 'worth learning').

[5] This refers of course to print dictionaries only – electronic dictionaries routinely offer the actual pronounced sound of a word, and sometimes a phrase or sentence, often with the option of regional accents such as British and North American English.

Box 7.2

Dictionaries vary in the way they show frequency. The two stars (**) after 'noun' in Figure 7.2 indicate that – in the system used in *MED* – the noun *rush* is a high-frequency word: it is more frequent than words with one star or none, but not as common as words in the highest frequency band, which have three stars. In Figure 7.39 below, we see from the lozenges (♦◊◊) in the *COBUILD* margin that *operator* is in the third frequency band. Other dictionaries (e.g. *LDOCE*) rate frequency differently according to whether the word is used in written or spoken language. The part-of-speech tagging in the corpus allows frequency calculations to be made for the various word-classes separately, while the information about the various text-types in the corpus allows the written/spoken distinction. Unfortunately, word sense disambiguation is not yet far enough advanced to permit the frequency marking of LUs.

7.2.2.4 *Inflected form* This component indicates the various inflections of the headword. Two types of inflected form are shown in Figure 7.3. This information is rarely given for every headword: the usual method is to settle on defaults – the regular forms – and specify the others. (The *Collins Schools Dictionary* bucks the trend here.) Once again, learners' dictionaries give more inflectional information than those for adult native speakers, which – like *ODE* – restrict themselves to helping the user with words likely to prove problematic in some way.

Paradigmatic grammar information is also included, of course. For languages where this is relevant, dictionaries will show the conjugation to which a verb headword belongs, often by a numbered cross-reference to a table at the end of the book where all the inflected forms are shown; similarly, for nouns the declension will be shown, here again usually by a numbered cross-reference. This type of information is normally shown for the headword only; it is not repeated for every TL noun and verb, although a unidirectional bilingual dictionary (defined in §2.2.1) often contains lists of TL noun and verb forms in the end matter.

➔ Remember that varieties of one language sometimes differ in their spelling of inflected forms and you may need to label the variants: for instance, the verb *travel* in British English inflects as *travelling*, *travelled*, while in American English the forms are *traveling*, *traveled*.

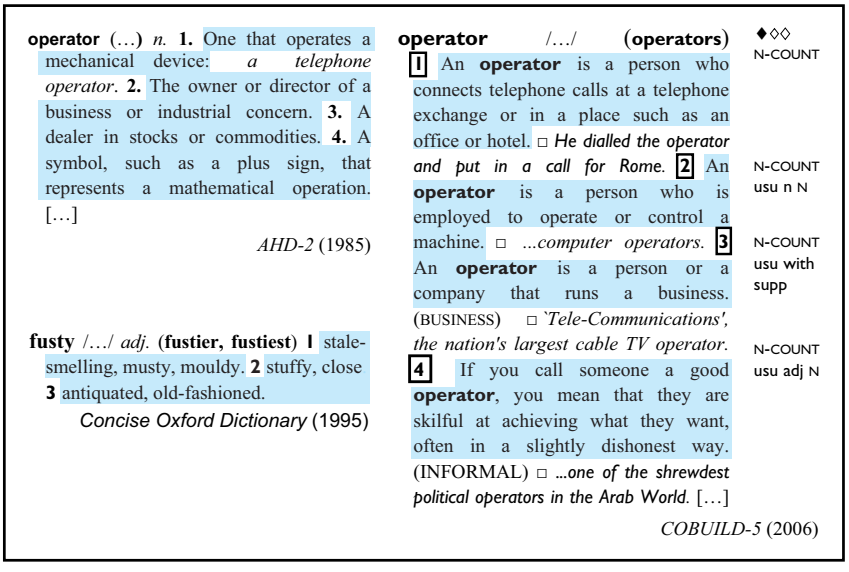
7.2.2.5 *Etymology* This component, illustrated in the *ODE* entry in Figure 7.3, shows the origin of the word and how it developed through time. Etymology is normally included in standard-sized monolingual dictionaries, but is rarely found in bilingual and monolingual learners' dictionaries, although it has begun to appear in some electronic dictionaries for learners, cf. §7.2.11.1.

➔ When you are writing etymologies, remember to word them so that they can be understood by the people you expect to use them. A plethora of abbreviations and typographical conventions is often lost on an unsophisticated user.

7.2.3 Meaning in monolingual dictionaries

From this point onwards, the components described all hold information about one specific LU, rather than about the whole headword. In monolingual dictionaries, the obvious way of transmitting the meaning of the headword is by means of the definition. In bilingual dictionaries, definitions are very rare, though they do occur in *bilingualized* dictionaries, i.e. monolingual learners' dictionaries which have been partially explicated in another language for a particular linguistic market, usually by the translation of all or parts of each entry. Standard bilingual dictionaries use the translation component as the principal way of telling the user what the headword means. Meaning-transmitting components of bilingual dictionaries are treated separately in §7.2.4.

7.2.3.1 *Definition* The definition explains the meaning of the headword in one particular sense. Three types of definition are shown in Figure 7.4. In the first, the *AHD* definitions are traditional, standard descriptions of the various meanings of the word *operator*, where clarity is not sacrificed to brevity, and a careful selection of facts makes the entry informative and intelligible. The use of *one that* . . . instead of the more usual *someone who* . . . gives a slightly formal flavour to the definition. (In contrast, the *LDOCE* entry for *know* in Figure 7.5 follows current British practice in its more informal approach, for instance in the use of *you* rather than *one* in both the senses here.) The second type of definition, seen in the *COBUILD* entry in Figure 7.4, illustrates an informal defining style designed to answer in a more 'natural-sounding' way the question 'What does this word mean?'



operator (…) *n.* **1.** One that operates a mechanical device: *a telephone operator*. **2.** The owner or director of a business or industrial concern. **3.** A dealer in stocks or commodities. **4.** A symbol, such as a plus sign, that represents a mathematical operation. […]

AHD-2 (1985)

fusty /…/ *adj.* (**fustier, fustiest**) **1** stale-smelling, musty, mouldy. **2** stuffy, close **3** antiquated, old-fashioned.

Concise Oxford Dictionary (1995)

operator /…/ (**operators**) **1** An **operator** is a person who connects telephone calls at a telephone exchange or in a place such as an office or hotel. □ *He dialled the operator and put in a call for Rome.* **2** An **operator** is a person who is employed to operate or control a machine. □ *...computer operators.* **3** An **operator** is a person or a company that runs a business. (BUSINESS) □ *'Tele-Communications', the nation's largest cable TV operator.* **4** If you call someone a good **operator**, you mean that they are skilful at achieving what they want, often in a slightly dishonest way. (INFORMAL) □ *...one of the shrewdest political operators in the Arab World.* […]

COBUILD-5 (2006)

♦◇◇ N-COUNT

N-COUNT usu n N

N-COUNT usu with supp

N-COUNT usu adj N

Fig 7.4 Definition components (unshaded) in three monolingual dictionaries

The third type, in the *Concise Oxford Dictionary* entry, illustrates the still common practice in dictionaries for adult native speakers of relying on a number of semi-synonyms to transmit the headword's meaning. This is convincing if you know what the word means already, but at best can only be complementary to a paraphrase definition. At worst it makes it impossible for anyone to learn from such entries the difference between these partial synonyms.

No definition can cover all the uses of a word, and all but the smallest dictionaries rely on examples (phrases or sentences) to fill some of the more obvious gaps: cf. §7.2.7.2. Nonetheless, the definition lies at the heart of the monolingual entry and is its most important component.

➔ Defining techniques can be learned: see Chapter 10 for a full discussion.

7.2.3.2 *Gloss* This component, in parentheses in both left-hand entries in Figure 7.5 and introduced by the equals sign (=) in *LDOCE* and 'ie' in *OALD*, allows a more informal explanation of the meaning of a multiword expression or example (or even part of one) in the entry, and is chiefly used in monolingual dictionaries for learners, to help understanding. Glosses are rare in monolingual dictionaries for adult native speakers. Another type of gloss, in the target language of a bilingual dictionary, is useful when a direct translation cannot be found; cf. TL gloss (§7.2.4.3).

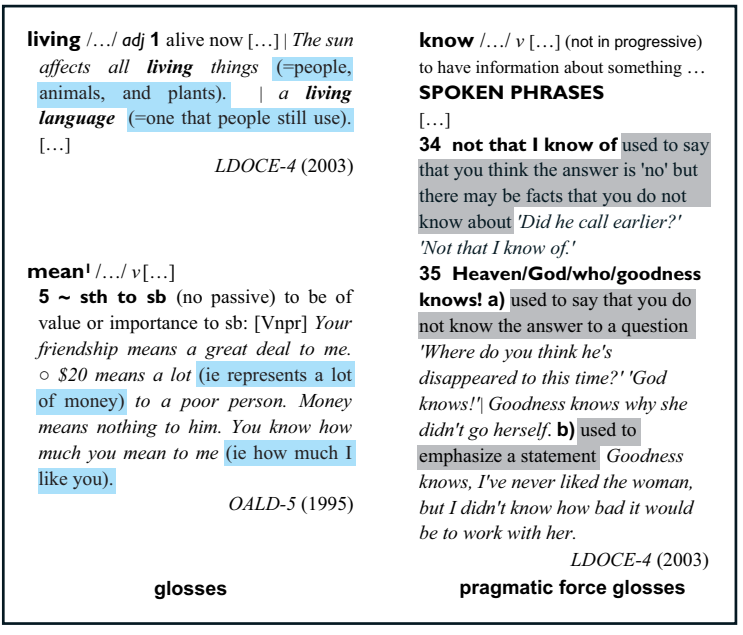


Fig 7.5 Two types of glosses in monolingual learners' dictionaries

➔ When inserting a gloss, make sure that the user can see exactly which part of its context the gloss refers to.

7.2.3.3 *Pragmatic force gloss* The pragmatic force gloss is a particular kind of gloss; its purpose is to explain the pragmatic message carried by a word or phrase. An example of this is seen in the two 'spoken phrases', introduced by 'used to . . .', from the *LDOCE* entry for *know* in Figure 7.5. This type of gloss is a very useful component in learners' dictionaries and can carry many different types of information.

➔ When inserting a pragmatic force gloss you must make it clear from the wording that it is not a simple explanation of meaning (i.e. a gloss proper) but an explanation of how the phrase is used to convey much more than its surface meaning.

7.2.3.4 *Graphic illustration* This component includes photographs, drawings, diagrams, etc. which appear in the text in order to clarify the meaning of a headword. It is used particularly in dictionaries for learners, and as well as a shortcut to meaning explanation – see for instance the pictures at *chimney*, *brush*, *broken*, etc. in *LDOCE-4* (2003) – it can serve to group

together vocabulary sets (as in the detailed illustration at *bicycle* in the same dictionary, where each part is named – 'saddle', 'handlebar', etc.). Some illustrations include information about grammar as well as vocabulary: for instance, the notes on countability in the two-page spread entitled *Fruit and vegetables* in the end matter of *OALD-7* (2005).

Bilingual dictionaries may also contain illustrations, and some exploit the potential of the *bicycle* type described above by naming the parts in both the source language (SL) and the target language (TL). However, illustrations are not used so much in bilinguals, simply because once you've given the translation of words like *chimney*, *brush*, *broken*, etc. there is really no need to add a picture.

7.2.4 Meaning and translation in bilingual dictionaries[6]

Anyone translating into or out of their own language uses entry components whose main function is to lead them to the best translation for their context. The bilingual dictionary is very flexible in this regard: four principal components serve this purpose. Two types of translation figure in entries – the direct translation, given without context (although often with sense indicators, cf. §7.2.5), and the contextual translation attached to an idiom or example phrase. In cases where no translation exists, you can use a near-equivalent or a TL gloss, or indeed both. These components are outlined here, and discussed more fully in §12.3.2. A lot of additional information can be given in the MWEs and example phrases with their translations, summarized in §7.2.7 below; this is amply illustrated by the *club* entry in Figure 7.6.

7.2.4.1 *Direct translation* The 'direct translation' is the component that holds the TL word or words offered as the most useful equivalent(s) to the SL headword. It must suit as many as possible. Direct translations in Figure 7.6 include 'club' (sense 1a), 'boîte de nuit' and 'boîte' (1b), 'massue', 'gourdin', 'matraque', and 'club' (1c), 'trèfle' (1d), 'frapper avec un gourdin', 'frapper avec une massue', and 'matraquer' (2) and 'du club' (4); in Figure 7.7 the only *direct translation* is 'A, a'.

[6] In this section we take into account both 'encoding' and 'decoding' use, by speakers of the source language and of the target language respectively, cf. §2.4.2.

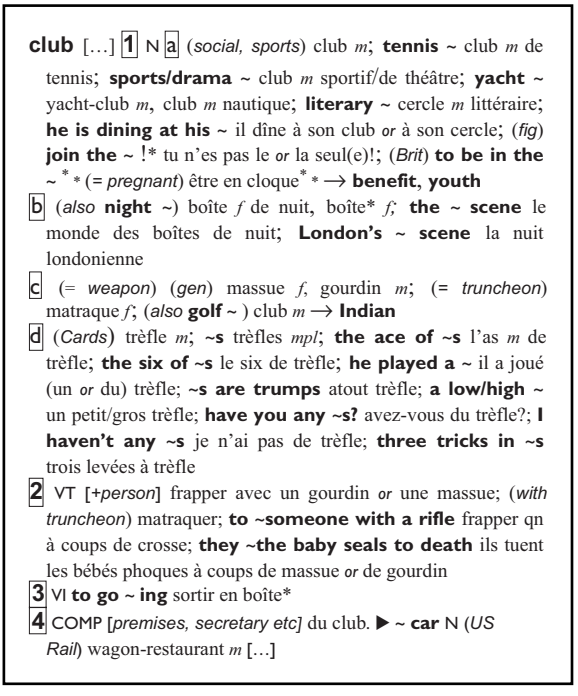

club […] 1 N a (*social, sports*) club *m*; **tennis ~** club *m* de tennis; **sports/drama ~** club *m* sportif/de théâtre; **yacht ~** yacht-club *m*, club *m* nautique; **literary ~** cercle *m* littéraire; **he is dining at his ~** il dîne à son club *or* à son cercle; (*fig*) **join the ~** !* tu n'es pas le *or* la seul(e)!; (*Brit*) **to be in the ~** * * (= *pregnant*) être en cloque* * → **benefit, youth**
b (*also* **night ~**) boîte *f* de nuit, boîte* *f;* **the ~ scene** le monde des boîtes de nuit; **London's ~ scene** la nuit londonienne
c (= *weapon*) (*gen*) massue *f*, gourdin *m*; (= *truncheon*) matraque *f*; (*also* **golf ~**) club *m* → **Indian**
d (*Cards*) trèfle *m*; **~s** trèfles *mpl*; **the ace of ~s** l'as *m* de trèfle; **the six of ~s** le six de trèfle; **he played a ~** il a joué (un *or* du) trèfle; **~s are trumps** atout trèfle; **a low/high ~** un petit/gros trèfle; **have you any ~s?** avez-vous du trèfle?; **I haven't any ~s** je n'ai pas de trèfle; **three tricks in ~s** trois levées à trèfle
2 VT [+*person*] frapper avec un gourdin *or* une massue; (*with truncheon*) matraquer; **to ~someone with a rifle** frapper qn à coups de crosse; **they ~the baby seals to death** ils tuent les bébés phoques à coups de massue *or* de gourdin
3 VI **to go ~ ing** sortir en boîte*
4 COMP [*premises, secretary etc*] du club. ▶ **~ car** N (*US Rail*) wagon-restaurant *m* […]

Fig 7.6 Translations and sense indicators in *CRFD-1998*

The *direct translation* lies at the heart of the bilingual entry and is perhaps its most important component. (The writer suspects that many students read no further.) Lexicographic translating skills can be learned: see Chapter 11 for a full discussion.

➔ Make sure that TL words given as direct translations are general enough to suit most contexts.

➔ If you have to give two direct translations, be sure to use sense indicators to highlight the difference between them. (This goes for contextual translations too.)

7.2.4.2 *Near-equivalent* This component may serve in place of a direct translation or a contextual translation, and is used when there is no real TL equivalent of the SL headword or phrase. In the entries in Figure 7.7 the near-equivalents are all introduced by the 'swung equals' (≈) sign. 'A comme André' doesn't *translate* 'A for Able', but is the equivalent phrase in the TL, used in exactly the same circumstances. Similarly, 'le baccalauréat' is not a translation of 'A levels' but is an equivalent exam

A,a […] 1 N a (= *letter*) A, a *m*; **A for Able** ≈ A comme André; **to know sth from A to Z** connaître qch de A à Z; **24a** (*in house numbers*) ≈ 24 bis; […] (*Brit Aut*) **on the A4** sur la (route) A4, ≈ sur la nationale 4 […] 2 COMP […] ▶ **A levels** NPL (*Brit Scol*) ≈ baccalauréat *m*; ▶ **to do an A level in geography** ≈ passer l'epreuve de géographie au baccalauréat […]

foreign […] ADJ […] (*Brit Pol*) **Foreign Secretary** ≈ ministre *m* des Affaires étrangères; […]

AA […] N a (*Brit*) (ABBR = **Automobile Association**) *société de dépannage* b […] c (*US Univ*) (ABBR = **Associate in Arts**) ≈ DEUG *m* de lettres.

Fig 7.7 Near-equivalents (≈) and TL glosses (in colour) in *CRFD-5* (1998)

in the French education system, and the accompanying example phrase is rendered into French by 'passer l'epreuve de géographie au baccalauréat'. In near-equivalents, the SL and TL items are often culturally equivalent. In the *foreign* entry in Figure 7.7, 'le ministre des Affaires étrangères' doesn't translate 'Foreign Secretary', but refers to that person's opposite number in the politics of francophone countries.

7.2.4.3 *TL gloss* When there is no direct translation and no near-equivalent, you have to fall back on the TL gloss: see the *AA* entry in Figure 7.7, where the motoring organization is glossed as 'société de dépannage' (literally, 'break-down recovery organization'). This explains the SL meaning to the TL user; it isn't much help to the SL user trying to find the French, but it would do at a pinch.

➔ If you have to compose a TL gloss, try to word it so that it will serve the encoding SL user too.

7.2.4.4 *Contextual translation* Like a definition, no direct translation can cover all the uses of a word, and all but the smallest dictionaries rely on translated examples (phrases or sentences: cf. §7.2.7.2 below) to fill some of the more obvious gaps. Thus the contextual translation, a twofold component consisting of an example phrase with its translation(s), plays an essential role in the bilingual entry; there are many instances of this type of translation in Figure 7.6: 'club de tennis' etc. in sense 1a, 'le monde des boîtes de nuit' etc. (1b), 'trèfles', 'le six de trèfle' etc. (1d), and so on. For more on the selection and translation of example material in the dictionary entry cf. §12.2.3.

➔ If there is no *direct translation*, *near-equivalent*, or *TL gloss*, then the only way to help people translate the headword is by means of carefully chosen *contextual translations* and if the worst comes to the worst this is what you have to do. This technique is shown in Figure 7.8, in an attempt to overcome the problem of the pronoun *next*: the translation of *the next* is entirely context-dependent.

next /.../ [...] A *pron* **after this train the ~ is at noon** le train suivant est à midi; **he's happy one minute, sad the ~** il passe facilement du rire aux larmes; **I hope my ~ will be a boy** j'espère que mon prochain enfant sera un garçon; [...]

OHFD-3 (2001)

Fig 7.8 Contextual translations in the absence of a direct translation

7.2.5 Sense indicators

A 'sense indicator' is a component designed to lead people as quickly as possible to the right part of the entry. (They are therefore a special kind of navigation aid.) Sense indicators are rare in monolingual dictionaries for native speakers, who can see from the definitions and examples the various senses of the headword. This is not the case, however, for learners of the language, and the sense indicator is an essential part of entries for learners. There are two main types of sense indicator: specifiers (in monolingual and bilingual dictionaries) and collocators (mainly in bilinguals). They are introduced here and discussed more fully in §12.3.4. You can also use other components such as domain labels for this purpose.

***club* entry**	**indicator**	**relationship with headword**
1 a (*social, sports*) club *m*;	specifier	modifiers (types of club)
b (*also* **night ~**) boîte *f* de nuit	specifier	compound synonym
c (*= weapon*)	specifier	superordinate
(*gen*) massue *f*, gourdin *m*;	specifier	'most often translated as …'
(*= truncheon*) matraque *f*;	specifier	'when the club is a truncheon'
(*also* **golf ~**) club *m*	specifier	another compound synonym
d (*Cards*) trèfle *m*;	domain label	'vocabulary of this subject matter'
2 [*+person*] frapper avec un gourdin	collocator	typical subject of verb
(*with truncheon*) matraquer;	specifier	'when hitting with a truncheon'
4 COMP [*premises, secretary etc*] du club.	collocators	typical nouns modified by headword

Fig 7.9 Various sense indicators in a bilingual entry

7.2.5.1 *Options in choosing how to indicate senses* Figure 7.6 shows the major part of the entry for *club* in *CRFD-5* (1998). A summary of the sense indicators in that entry is given in Figure 7.9, where they are isolated and named and their relationship with the headword is described. (Specifiers and collocators are explained in §7.2.5.2 and §7.2.5.3, and domain labels in §7.2.8.1.) As Figure 7.9 shows, there are several ways of indicating the specific sense of the headword. Take for instance the word *column*, which has a number of meanings. In bilingual dictionaries the TL may offer a different equivalent for each of these, and it's therefore important to tell the SL-speaking user which meaning is which.

sense	specifiers	labels
1. upright supporting building	'pillar' 'in building'	*Arch*
2. horizontal line of people, vehicles, etc.	'line' 'of people, cars'	(?)
3. vertical line of numbers on page	'of figures' 'in account book'	*Book-kpg*
4. vertical section of print on page	'of print' 'in newspaper'	*Press*
5. regular article in newspaper	'article' 'in newspaper' 'by journalist'	*Press*
6. vertical configuration of smoke etc.	'of smoke etc.'	*fig*

Fig 7.10 Different ways of indicating the senses of *column*

The table in Figure 7.10 shows some of the meanings of *column*, together with various ways in which these could be indicated, by using specifiers or labels. It is clear that, while domain labels such as *Press* (journalism) and *Arch* (architecture) are very space-saving, it's not always possible to find an appropriate domain not shared by other meanings, and specifiers are much easier to understand. The use of *fig* ('figuratively speaking') is common in bilingual dictionaries to indicate a metaphorical extension frequent enough to justify the status of sense, but so non-specific as to be difficult to pin down in a few words. (Our corpus offers columns of *air*, *ash*, *dust*, *eagles*(!), *mercury*, *rocks*, *smoke*, and much more.) Once again we're up against space vs. intelligibility . . .

➔ Domain labels may satisfy lexicographers but they're no good if the user can't make sense of them.

7.2.5.2 *Specifiers and signposts* As Figure 7.9 shows, specifiers can contain many different types of information, including superordinates, synonyms, cohyponyms, typical modifiers, paraphrases, and so on. Indeed, here almost anything goes.

➔ When you have to devise a specifier, think of the typical user. Try to fix on something that will conjure up just that one sense in the user's mind. (Easier said than done.)

rush[1] *n*

1 ▶FAST MOVEMENT◀[singular] a sudden fast movement of things or people : **rush of air/wind/water** *She felt a cold rush of air as she wound down her window.* **in a rush** *Her words came out in a rush.* | *At five past twelve there was a* ***mad rush*** *to the dinner hall.*

2 ▶HURRY◀[singular, U] a situation in which you need to hurry: *I knew there would be a* ***last-minute rush*** *to meet the deadline.* | *Don't worry,* ***there's no rush****. We don't have to be at the station until 10.* | **do sth in a rush** (=do something quickly because you need to hurry) *I had to do my homework in a rush because I was late.* | **be in rush** *I'm sorry, I can't talk now – I'm in a rush.*[…]

3 ▶BUSY PERIOD◀ **the rush** the time in the day, month, year etc. when a place or group of people is particularly busy → **peak** *The café is quiet until the lunchtime rush begins.* […]

4 ▶PEOPLE WANTING SOMETHING◀ [singular] a situation in which a lot of **people** suddenly try to do or get something **[+on]** *There's always a rush on swimsuits in the hot weather* […]

5 ▶FEELING◀ [singular] **a)** *informal* a sudden strong, usually pleasant feeling that you get from taking a drug or from doing something exciting → **high** *The feeling of power gave me such a rush.* | *an* ***adrenalin rush*** **b) rush of anger / excitement / gratitude etc.** a sudden very strong feeling of anger etc […]

6 ▶PLANT◀ [C usually plural] a type of tall grass that grows in water, often used for making baskets.

7 ▶FILM◀ **rushes** [plural] the first prints of a film before it has been edited […]

8 ▶AMERICAN STUDENTS◀ [U] AmE the time when students in American universities […]

Fig 7.11 Signposts in *LDOCE-4* (2003)

One particular type of specifier, generally known as a 'signpost', and illustrated in Figure 7.11, deserves separate mention because of its increasing use in monolingual learners' dictionaries. It is often realized by a synonym or paraphrase of the headword (senses 1–4 inclusive), but – as this *LDOCE* entry shows – may also offer a superordinate of the headword (senses 5 and 6) or an indication of the domain or subject matter (senses 7 and 8). The signposts have a similar function to the items shown in a 'menu' (as shown in the *MED* entry in Figure 7.2, cf. §7.2.1.3), but they are located beside the sense they apply to, and are typically even more telegraphic than menu items.

➔ For these signposts, remember to choose simple words which the user is likely to understand – that's more important than finding a very close synonym of the headword.

7.2.5.3 *Collocator* A collocator is a word chosen to represent a 'lexical set', i.e. a group of words belonging to the same wordclass and similar in meaning. Collocators exist to guide users towards the best translations. Collocators are therefore words from the language of the encoding user, i.e. the source language.[7]

clear /…/ […]
B *adj* 1 (transparent) [*glass, liquid*] transparent; [*blue*] limpide; [*lens, varnish*] incolore; 2 (distinct) [*image, outline, impression*] net/nette; [*writing*] lisible; [*sound, voice*] clair; […]
D *vtr* 1 (remove) abattre [*trees*]; arracher [*weeds*]; enlever [*debris, papers, mines*]; dégager [*snow*] **(from, off** de); […] 2 (free from obstruction) déboucher [*drains*]; dégager [*road*]; débarrasser [*table, surface*]; déblayer [*site*]; défricher [*land*]; […]
E *vi* 1 (become transparent, unclouded) [*liquid, sky*] s'éclaircir; 2 (disappear) [*smoke, fog, cloud*] se dissiper; 3 (become pure) [*air*] se purifier; 4 (go away) [*rash, pimples*] disparaître; [*skin*] devenir net/nette; 5 Fin [*cheque*] être compensé

Fig 7.12 Collocators in the entry for *clear* in *OHFD-3* (2001)

The grammatical relationship of collocator to headword depends on the wordclass of the LU. Collocators of adjectives are usually nouns typically modified by the headword, such as 'glass', 'blue', 'image', 'writing', etc. in B *adj* in Figure 7.12. The collocators for the transitive uses of *clear* in D *vtr* (the nouns 'trees', 'weeds', 'drains', 'road', etc.) are typical objects of the English verb in the two senses shown, while in E *vi* the nouns 'liquid', 'smoke', 'air', 'rash', 'cheque', etc. are typical subjects of the various senses of the intransitive verb. In this dictionary, typical subjects of headword verbs come before the appropriate translation, and typical objects afterwards. Collocators of noun headwords are normally either typical 'possessors' (as in A n 'of person' in Figure 7.13) or other nouns which

[7] Some bilingual dictionaries, instead of SL collocators, include collocators in the TL, as for example Hachette's *Dictionnaire Anglais-Français* (1934), where the intransitive uses of *clear* include '(temps) s'éclaircir, se dégager . . . (nuages, brume etc.) se dissiper . . . '. It is difficult to know who is likely to benefit from this approach – English speakers may very well not understand the TL collocators, while French speakers don't need them to help choose the correct equivalent in their own language.

are typically modified by the headword (as 'movement', 'muscle', etc. in B modif).

leg /…/
A *n* 1 (*of person*) jambe *f*; (*of animal*) gen patte *f*; (*of horse*) jambe *f*; […] 2 (*of furniture*) pied *m*; […]
B *modif* [*movement, muscle*] de la jambe; [*pain*] à la jambe; [*exercises*] pour les jambes.

Fig 7.13 Collocators in the entry for *leg* in *OHFD-3* (2001)

➔ When you are looking for collocators, see what words figure in the corpus data, group them semantically, and try to find more general words (such as superordinates) that can stand for them in the entry.

Box 7.3

Collocators are entry components and must not be confused with *collocates* (words with significant co-occurrence frequencies in corpora, cf. §9.2.7). Because collocators are thought up by the lexicographer as the words most likely to help the user choose a translation, the actual words themselves may not appear in the SL corpus at all. For instance, corpus data for the verb *develop* shows as subjects of the intransitive use the actual expressions *France*, *Indonesia*, *the west*, *the surrounding area*, and so on: these are summarized in the dictionary entry by the 'typical subject' collocator *region*.

7.2.6 Grammar

Every dictionary has its own underlying grammar schema, and the Style Guide will list the items (often abbreviations) you can use in the various grammar components, and explain how and when to use them. A simpler version of this information – for the benefit of the dictionary user – is usually also provided in the front matter. Learners' dictionaries, both monolingual and bilingual, tend to include more information about the grammar of the headword than do dictionaries for native speakers. In this section, we introduce the three principal components used to carry grammatical information, illustrated in Figure 7.14.

➔ If you can't show the headword grammar by means of these components, think of including an example to show how the headword is used (§7.2.7.2).

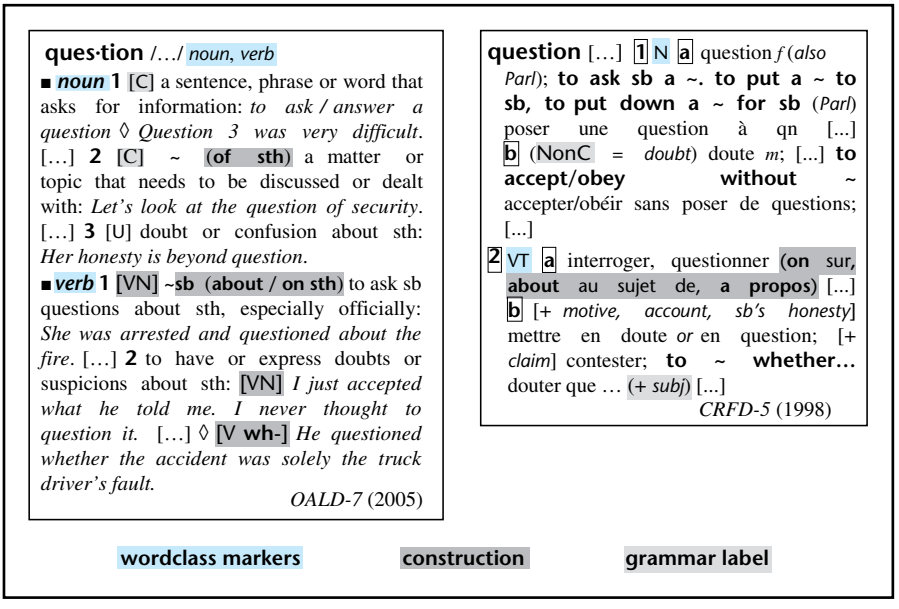


Fig 7.14 Three types of grammar components in learners' dictionaries

7.2.6.1 *Wordclass marker* Dictionaries don't differ much in the way they show the wordclass of the headword in its various uses (in English, the term would include at least noun, verb, adjective, adverb, pronoun, conjunction, preposition, article, and interjection). Most print dictionaries use abbreviations such as *n*, *v*, *adj*, and so on, but the grammatical terms are normally shown in full in electronic dictionaries.

Standard dictionary procedure is seen in the *OALD* entry in Figure 7.14, which uses wordclass markers ('noun' and 'verb') to introduce the two groups of LUs. Similarly, *CRFD* indicates the wordclass of the LU by 'N' and 'VT' (verb transitive). Less common is *OALD*'s opening summary of the two wordclasses of *question* ('noun, verb').

➔ Follow the Style Guide when you're inserting a wordclass marker: it's fairly straightforward.

7.2.6.2 *Construction* The construction[8] component is the 'second layer' of grammatical information and nowadays often reflects corpus evidence. The content of this component depends directly upon what is considered to be the headword's 'syntactic valency', i.e. all the constructions which a speaker

[8] Also called 'valency pattern', 'structure', or 'syntax pattern'.

of the language must know in order to use the word flexibly and fluently, and which ideally should be included in a learners' dictionary entry. Figure 7.15 shows some constructions which form part of the valency of the verb *watch*, together with codes[9] which might be used to record them. Every dictionary has its own view of what should be included, and the Style Guide contains codes and abbreviations to use in this component.

Contexts	Codes
She watched . . .	
the boat	NP
the car drive off	NP Vinf
the children playing	NP Ving
what they were doing	cl-wh
how they laughed and talked	cl-wh
how to tie the rope	wh-Vinf-to
through the telescope	PP-through
for the postman	PP-for
for the postman to appear	PP-for NP Vinf-to

Fig 7.15 Some constructions for the verb *watch*

Constructions need to be recorded for the four major wordclasses. A verb's constructions are of course an indication of its transitivity, and indeed much more. Dictionaries may indicate transitivity status specifically, labelling verbs as *vi* (intransitive) or *vt* (transitive), etc. Some label other subclasses of verbs, like reflexives (*v refl*), reporting verbs (*v rep*), modals (*v mod*), and other auxiliaries (*v aux*), as an alternative to spelling out their syntactic valency. There is a case for considering such labels as a distinct entry component (perhaps *subwordclass marker*), but because many dictionaries prefer to show these facts in the form of syntactic complements rather than subwordclasses, it is convenient to deal with them all as one single component: the 'construction'. Thus, in the *OALD* entry in Figure 7.14, the transitivity of *question* is shown by the construction 'VN' (verb + noun phrase, otherwise 'transitive verb').[10] More information about this verb is given further down the same entry by the construction 'V **wh**' (verb + wh-clause).

[9] A list of such codes for the major wordclasses is given in Atkins, Fillmore, and Johnson (2003).

[10] Note that *CRFD* gives 'transitive verb' ('VT') full wordclass status, on a par with 'noun' etc., as may be seen from the entry in Figure 7.14. These things are never cut and dried.

Just as the type of information in this component varies from dictionary to dictionary, so also does the way in which it is presented in the entry. Compare for instance the *question* entries in *OALD* and *CRFD* in Figure 7.14, where the same facts are coded as '~ sb (about / on sth)' and '(on sur, about au sujet de, à propos)' respectively (note the way in which the TL equivalent constructions for *interroger* and *questionner* are included in the bilingual dictionary). The constructions necessary to an adjective headword can be seen in the shaded components in Figure 7.16.

aware /. . . / *adj*
1 [not before noun] ~ (**of sth**) | ~ (**that** . . .) knowing or realizing sth: *I don't think people are really aware of just how much it costs.* [. . .]

equal / . . . / *adj., noun, verb*
■ ***adj.*** **1** ~ (**to sb/sth**) the same in size, quantity, value, etc. as sth else: *There is an equal number of boys and girls in the class* [. . .]

Fig 7.16 Some constructions in *OALD-7* (2005) adjective entries

7.2.6.3 *Grammar label* The third 'layer' of grammatical information is considerably less straightforward: it depends directly on the wordclass of the headword, and its contents reflect the amount and type of such information the editors believe will be useful for (and intelligible to) the user.

For nouns, countability is often shown, as in the entries from *OALD* ('C' countable and 'U' uncountable) and *CRFD* ('NonC' non-countable) in Figure 7.14. Proper nouns are sometimes marked as such (in *COBUILD* for instance). For verbs, information may be given about whether the headword is an activity, accomplishment, achievement, or stative verb, such as *LDOCE's* 'not in progressive' in the *know* entry in Figure 7.5, or other miscellaneous facts. *COBUILD* in its dedicated side columns (shown in Figure 7.4) indicates recurring contextual patterning as well as valency constructions. Adjective entries also need extra grammar information: in the *aware* entry in Figure 7.16 '[not before noun]' is a grammar label component. Others are shaded in Figure 7.17: the *MED* entry warns its users that the adjective *mere* is always attributive, never predicative; while *OHFD* warns its English-speaking users that the French equivalents *pur* and *simple* in this particular sense are also used only attributively.

Other grammatical information is often given in the metalanguage, but there's no practical point in classifying it further. One example is '(+ *subj*)' in the *CRFD* entry in Figure 7.14, where the subjunctive after *douter que* is specified for the benefit of SL speakers (French TL speakers know that already).

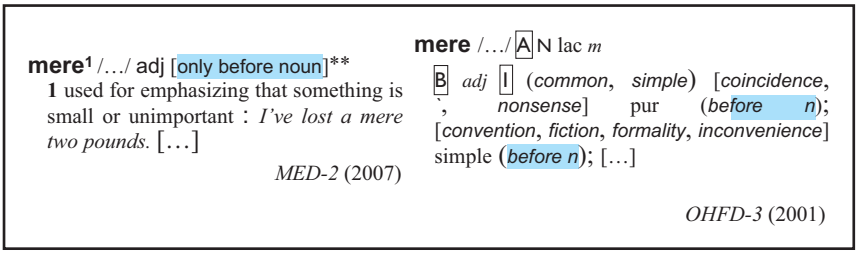


Fig 7.17 *Grammar label* components in learners' dictionaries

7.2.7 Contexts

All the entry components in this section hold facts about particular lexical contexts (words and phrases) in which the headword is found. Such contexts may consist of various types of multiword expression in which the headword occurs, or simply the headword's collocates, i.e. words with significant co-occurrence frequencies in the corpus. There are two main subdivisions:

- components relating to idiomatic material (outlined in §7.2.7.1 below)
- other illustrative sentences or phrases (outlined in §7.2.7.2).

Since it is language-learners who have most need of this kind of information, it is in learners' dictionaries that you find the richest context material. All of them appear in the *OALD* entry in Figure 7.18.

7.2.7.1 *Multiword expressions* As with grammar components, the specific MWE components selected for a particular dictionary depend upon the language being described. There are no absolute, clearly defined categories here (see the discussion in §6.2.2): as elsewhere in language, we are dealing with a gradient. Not surprisingly, different dictionaries do different things here. However, four types of MWE components are enough to hold most English contexts – idioms, collocations, phrasal verbs, and compounds[11] – and these are illustrated in the *OALD* entry in Figure 7.18.

MWE: idiom If your entry structure includes this component, the Style Guide will tell you which of the various types of MWE discussed in §6.2.2 it should hold – perhaps the easily recognizable phrasal idioms such as *beat about the bush* and *beat your breast* in the IDM section in the *OALD* entry in Figure 7.18, together with fixed phrases, catchphrases, proverbs,

[11] Note, however, that many learners' dictionaries, monolingual and bilingual, give phrasal verbs, compounds, and sometimes idioms the status of 'secondary headwords' (cf. §7.2.10.1).

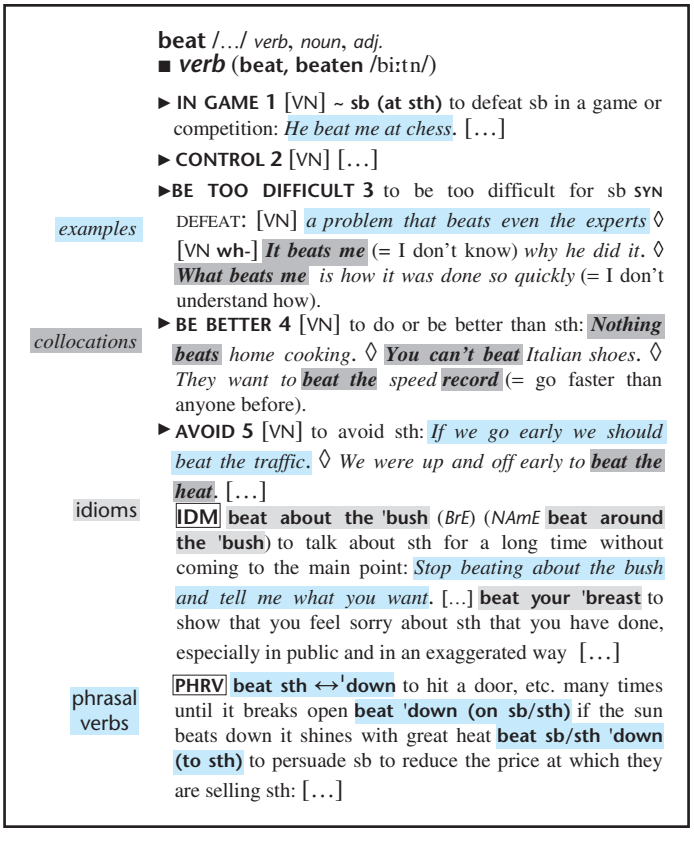


Fig 7.18 Context components in *OALD-7* (2005)

quotations, greetings, phatic phrases: essentially, any frequently occurring phrase whose meaning is more than the sum of its parts. There are no real objective criteria which distinguish idioms from collocations, as may be perceived from a comparison of the same material in two similar dictionaries. As the *club* entry in Figure 7.6 shows, the *CRFD* has no idiom or collocation component, all the phrasal idioms being included, together with non-idiomatic examples, within the example component.

MWE: collocation This component holds the kind of phrase called 'transparent collocation' in §6.2.2.1: a significantly frequent grouping of words whose meaning is quite transparent, such as *nothing beats…* and *beat the… record* within sense 4 of the main *OALD* sample entry in Figure 7.18, where these phrases are embedded amongst straightforward examples. Support verb constructions are often also treated in this way.

MWE: phrasal verb As its name implies, this component holds the phrasal verbs in which the headword figures. When this component is used, the term 'phrasal verb' must be defined and its treatment specified in the Style Guide (dictionaries vary on this point). In the *OALD* sample in Figure 7.18, phrasal verbs such as *beat down* have their own section of the entry, flagged by PHRV. In the *CRFD*, phrasal verbs like *club together* follow immediately upon the main entry, as *secondary headwords*, flagged by a solid triangle, as in Figure 7.19.

MWE: compound This component holds two-word or multiword compounds in which the headword appears as the first element.[12] How such compounds are treated varies from dictionary to dictionary. Here are two of the places in which they may appear, but there are many more options:

- within a dedicated section of the entry, e.g. *club car*, *club chair*, *club-footed*, etc. in the COMP (compound) section of the *CRFD* entry in Figure 7.19
- as headwords in their own right, e.g. *clubhouse*, *clubland*, etc. in the same dictionary.

club […] 1 N
[…]
4 COMP [*premises, secretary etc]* du club. ▶ **club car** N (*US Rail*) wagon-restaurant ▶ **club chair** N fauteuil *m* club […] ▶ **club-footed** ADJ pied bot *inv* […] ▶ **club subscription** N cotisation *f* (à un club)
▶**club together** VI (*esp Brit*) se cotiser; **to club together to buy sth** se cotiser pour acheter qch.
clubhouse ['klʌbhaʊs] N (*Sport*) pavillon *m*, club-house *m*
clubland ['klʌblænd] N […]
clubman ['klʌbmən] N […]

Fig 7.19 Phrasal verbs and compounds in *CRFD-5* (1998)

In the case of the *CRFD*, the location depends on the form of the compound. Two-word or hyphenated compounds are treated within the entry;

[12] As the *CRFD* entry in Figure 7.6 shows, compounds in which the headword is the second element (e.g. *tennis club*, *sports club*, etc.) sometimes appear within the entry as examples chosen to show how the headword is translated when modified by another noun.

solid words are headwords. Monolingual learners' dictionaries tend to make all compounds, whatever their form, into headwords, which makes sense, given the uncertain status of hyphenated forms. All these questions must be clarified in the Style Guide.

7.2.7.2 *Example* Every dictionary has its own detailed policy on the selection or production of examples. The example component may hold two types of illustrative sentence or phrase:

- one that simply illustrates facts already given elsewhere in the entry (for instance, in the grammar codes)
- one that adds information to the entry, either by telling the user something that (for instance) can't be coded into grammar components, or – in the case of bilingual dictionaries – by giving a translation for the headword in a particular context.

Examples are usually expected to pull their weight in the entry; wholly illustrative examples are rare, because of space constraints. For obvious reasons, learners' dictionaries, monolingual and bilingual, make more use of examples than do native-speaker dictionaries. In bilingual dictionaries, where each full example is normally translated, the example together with its translation(s) form a twofold unit described as a contextual translation (cf. §7.2.4.4).

In form, an example can be:

- a complete sentence, or
- a partial sentence.

There are several options with regard to the content of an example; the Style Guide must give guidance here. Examples may be:

- exactly as they are found in the corpus
- abridged from a corpus sentence, but otherwise unadulterated
- adapted from a corpus sentence, but making sure that the example illustrates the same fact as the original sentence (which was recorded for a particular reason)
- wholly composed, in order to illustrate specific facts.

Choosing examples is a very important part of entry-writing, and is further discussed in §9.2.4, §10.8 and §12.3.3.

7.2.8 Vocabulary types: linguistic labels

When an indication of vocabulary type (cf. §6.4.1.4) is given in a print dictionary, this is normally in the form of a 'linguistic label'. Dictionaries will offer in the front or back matter a list of the abbreviations used in these labels.

The first thing to think about is: *What does a label label?* Here are two groups of words that you might be tempted to label 'archaic':

> *chainmail*, *jousting*, *woad*, *alchemist*
> *helpmeet*, *verily*, *greensward*

The first group denotes a person, thing, or activity no longer part of modern life; however there is no other word for any of them, and if we want to talk about them we must use these four words. The words themselves are not archaic. Of the second group, *helpmeet* is an archaic word for 'companion' or 'spouse'; *verily* for 'truly' or 'in truth'; and *greensward* for 'patch of grass'. The concepts denoted by these words are still current, but the words are not, and should be labelled 'archaic'.

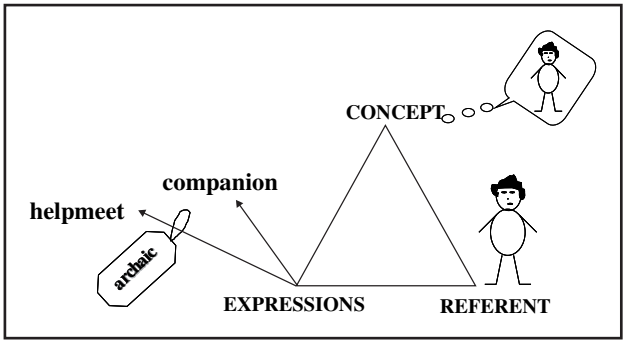


Fig 7.20 Labelling says something about the expression (word or phrase)

The difference between labelling a word and labelling what it refers to is often difficult for new lexicographers. The adapted version (Figure 7.20) of Ogden and Richards' well-known 'meaning triangle'[13] may make things clearer. It illustrates a threefold distinction:

- the 'referent' (a person in the real world)
- the 'concept' (broadly, what you think of when you hear or use either *helpmeet* or *companion*)
- any 'expression' (word or phrase) that refers to this person.

[13] Ogden, C. K. and Richards, I. A. (1923) *The Meaning of Meaning.*

Only an expression can be labelled, not a concept and certainly not a referent. In this case, the expressions differ in the time dimension: *companion* is unmarked, and only *helpmeet*, an archaic word, would be labelled in a dictionary.

7.2.8.1 *Domain* Domain labels (discussed in § 6.4.1.4 and highlighted in Figure 7.21) have an important role to play in lexical databases, particularly those used by computers, where the domain label is useful in word sense disambiguation. In publishers' databases, these labels offer a way of automating lists of specialized vocabulary which can be exploited in a number of ways. As Figure 7.10 showed, however, they are not always instantly comprehensible to dictionary users, and nowadays tend to be used sparingly.

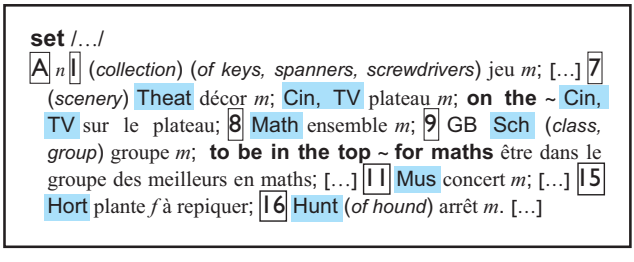

set /…/
A *n* 1 (*collection*) (*of keys, spanners, screwdrivers*) jeu *m*; […] 7 (*scenery*) Theat décor *m*; Cin, TV plateau *m*; **on the ~** Cin, TV sur le plateau; 8 Math ensemble *m*; 9 GB Sch (*class, group*) groupe *m*; **to be in the top ~ for maths** être dans le groupe des meilleurs en maths; […] 11 Mus concert *m*; […] 15 Hort plante *f* à repiquer; 16 Hunt (*of hound*) arrêt *m*. […]

Fig 7.21 Some domain labels in *OHFD-3* (2001)

➔ A 'domain' label indicates that *the item is used when the subject of discussion is* . . . (science, hockey, plumbing, poetry, etc.).

7.2.8.2 *Region* The *ODE* entry in Figure 7.22 contains a number of regional labels (cf. §6.4.1.4): 'fair dos' and 'be set fair' are marked as British English, and 'fair go' as from Australia and New Zealand. Most dictionaries establish one region or a group of regions (in this case, world English) as a default, and mark other items. This is especially useful information for language-learners.

➔ A regional label indicates that *the item is mainly but not exclusively used in* . . . (Britain, the United States, Australia, etc.).

Dialect Figure 7.22 also shows a 'dialect' label (cf. §6.4.1.4). This specific type of regional label indicates that the verb 'to fair', used of the weather, is not standard English. It is more informative if a dialect label is accompanied by a regional label showing where the word is current, for instance 'Scot dialect'. The dialect label is rare in learners' dictionaries, mainly because dialectal vocabulary is rare in these dictionaries.

fair[1] /.../ ▶ **adjective 1** treating people equally without favouritism or discrimination: [...]
5 archaic beautiful: *the fairest of her daughters*. [...]
▶ **verb** [no obj.] dialect (of the weather) become fine: *looks like it's fairing off some*.
– PHRASES **all's fair in love and war** proverb in certain highly charged situations, any method of achieving your objective is justifiable.[...] **fair dos** Brit. informal used to request just treatment or accept that it has been given: *Fair dos – you don't believe I've been idle all this time?* [...] **fair go** Austral. /NZ informal used for emphasis or to request someone to be reasonable or fair: *Fair go! How can I ask a thing like that?* **fair name** dated a good reputation. **the fair** (or **fairer**) **sex** dated or humorous women. **fair's fair** informal used to request just treatment or assert that a situation is just: *Fair's fair – we were here first*. [...] **be set fair** Brit. (of the weather) be fine and likely to stay fine for a time.

Fig 7.22 Various linguistic labels in *ODE-2* (2003)

➔ A dialect label indicates that *the item belongs to the non-standard language of*... (Yorkshire, Devon, etc.).

7.2.8.3 *Register* Register labelling (cf. §6.4.1.4) is perhaps the most common of all in general trade dictionaries. Most dictionaries mark at least two layers of informality ('informal', 'very informal', etc.) and one of formality ('formal'). In Figure 7.22, the phrases 'fair dos', 'fair go', and 'fair's fair' are marked as informal, the first in British English only, and the second in the English of the Antipodes.
➔ A 'register' label shows that *the use of this item indicates a*... (formal, very familiar, etc.) *manner of speech or writing*.

Slang and jargon Slang and jargon labels (§6.4.1.4) constitute a subset of register labels. These labels make more sense if accompanied by some indication of the group of people who use it, for instance 'army slang' or 'computer jargon'.[14]
➔ A slang or jargon label indicates that *the item is non-standard language used by the named group* (naval personnel, computer experts, etc.).

[14] In some dictionaries, 'slang' is considered to be a register label, meaning 'even more informal than very informal'. The Style Guide once again must make these distinctions clear.

Offensive terms The offensive-term label constitutes another subset of register labels. It covers a catch-all group of items which can cause offence of one degree or another (from swear words to extreme racist terms). Labels of this type vary from 'rude' through 'offensive' to 'taboo'. Here again, the Style Guide sets the limits.

➔ An offensive-term label indicates that *the use of this item will cause offence and should normally be avoided.*

7.2.8.4 *Style* One style label (cf. §6.4.1.4) in Figure 7.22 shows that if you refer to women as *the fair sex* either your language is rather old-fashioned or you are trying to be funny. This dictionary also treats 'proverb' as a style label. The most common style label is 'literary', indicating that the word is found in literature but not in conversational language.

➔ A style label indicates that *the item is normally used in a . . .* (literary, newspaper, etc.) *text*.

Box 7.4 Style vs. domain labels

Some people have trouble distinguishing *style labels* from *domain labels*, especially where the dictionary has 'literary' as a style label and 'literature' as a domain label. The word *bounteous* is a word found in literary and poetic texts, but has nothing to do with literature, and so would be labelled with the *style label* 'literary'; the word *sonnet* is a perfectly ordinary word but belongs to the field of literature, and so would be labelled with the *domain label* 'literature'. If you use the 'rule of thumb' practical tips given after each of these sections, you won't make any mistake here.

7.2.8.5 *Time* The *ODE* entry in Figure 7.22 notes two levels of out-of-dateness: 'archaic' in the case of *fair* meaning 'beautiful' (other dictionaries use 'obsolete' for this label), and 'dated' for *the fair sex* (elsewhere 'old' or 'old-fashioned' serves the same purpose). It is particularly useful for language-learners to be warned that an item is no longer in current use among younger speakers: this is the purpose of the 'time' label (cf. §6.4.1.4), which can also be used to mark as 'ephemeral' phrases which have only recently entered the language and are not expected to stick around for long. In the absence of a crystal ball, editors tend to avoid 'ephemeral' labelling.

➔ A 'time' label indicates that *in the dimension of time, the use of this item is* . . . (obsolete, old-fashioned, etc.).

7.2.8.6 *Attitude* Attitude labels such as *pej* (pejorative), *derog* (derogatory), and *apprec* (appreciative), discussed in §6.4.1.4, appear mostly in learners' dictionaries, such as the *OALD* entries seen in Figure 7.23, where the first sense only of each entry is marked in this way. Dictionaries for adult native speakers usually include this kind of information within the definition itself: *ODE* defines *slender* as 'gracefully thin', *CED* as 'slim and well-formed', *AHD* as 'gracefully slim'. *CED* defines *conventional* as 'following the accepted customs and proprieties, esp. in a way that lacks originality'.

slen·der /…/ *adjective* (**slen·derer**, **slen·derest**) **1** (*approving*) (*of people or their bodies*) thin in an attractive or elegant way SYN **slim**: *her slender figure* ◊ *long, slender fingers* **2** thin or narrow: *a glass with a slender stem* **3** small in amount or size and hardly enough: *to win by a slender margin / majority* […]

con·ven·tion·al. /…/ *adjective* **1** (*often disapproving*) tending to follow what is done or considered acceptable by society in general; normal and ordinary, and perhaps not very interesting: *conventional behaviour* […] **2** (*usually before noun*) following what is traditional or the way sth has been done for a long time: *conventional methods* […]

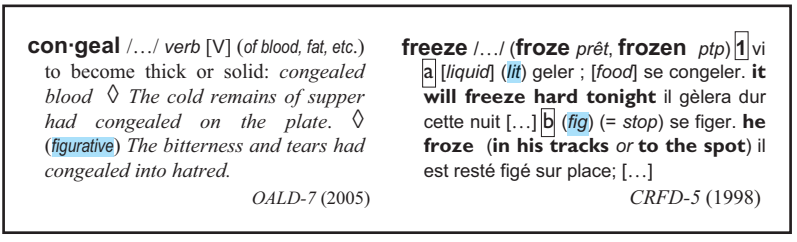
Fig 7.23 Attitude labels in *OALD-7* (2005)

➔ An 'attitude' label indicates that *the use of this word is intended to imply* . . . (approval or disapproval).

7.2.8.7 *Meaning type* Extended meanings occur in many if not all languages, and dictionaries exploit this by assuming that everyone understands the distinction between literal meaning and figurative (or metaphorical) meaning. The most frequent meaning type labels are *lit* (literally) and *fig* (figuratively). They are often used in cases where the sense shift is not so well established as to constitute a new LU, as in the *OALD* entry in Figure 7.24.

➔ A 'meaning type' label indicates that *the item should be interpreted* . . . (literally or figuratively).

7.2.8.8 *Using labels* This section brings together issues to be considered by senior editors when devising a labels policy for a dictionary or database project. Most lexicographers simply have to follow the Style Guide



con·geal /…/ *verb* [V] (*of blood, fat, etc.*) to become thick or solid: *congealed blood* ◊ *The cold remains of supper had congealed on the plate.* ◊ (*figurative*) *The bitterness and tears had congealed into hatred.*

OALD-7 (2005)

freeze /…/ (**froze** *prêt*, **frozen** *ptp*) 1 vi a [*liquid*] (*lit*) geler ; [*food*] se congeler. **it will freeze hard tonight** il gèlera dur cette nuit […] b (*fig*) (= *stop*) se figer. **he froze** (**in his tracks** *or* **to the spot**) il est resté figé sur place; […]

CRFD-5 (1998)

Fig 7.24 Meaning type labels in learners' dictionaries

when applying labels. Since the labels themselves form closed sets, they can be selected from a pull-down menu within dictionary production software.

Devising a labels policy There's quite a lot of work involved in putting together a consistent policy on labels in a dictionary. Some of the issues are:

- which types of label to use, e.g. domain, region, register, etc.
- which labels to use for each type, e.g. 'art', 'architecture', etc. in the *domain* list
- when a label is to be used: the options are . . .
  - on every possible occasion (good for computers)
  - only when it will actively help the users
  - always for some types, when helpful for others
- where the label is to be placed, i.e. before or after the item it marks
- what the scope of the label is (see below)
- how to handle multiple labels on one item (see below).

➔ Remember that labels, involving an additional level of abstraction, are not very informative for human users but very useful for computers.

The scope of labels How far across the surrounding text (in either direction) is the label meant to apply? This always depends on the dictionary's policy on labels, as set out for lexicographers in the Style Guide, and explained for dictionary users in the front matter. Label scope is one of the conventions set up in the dialogue between the lexicographer and the dictionary user. It's a convention which lexicographers follow to the letter, and which most users are probably entirely unaware of.

The simplest case in Figure 7.25 is the *MED* entry, where the label '*computing*' applies to the whole *MED* entry: the word *baud* is a technical term in computing and has no other use.

Another fairly straightforward use of labels is seen in the entry for *conventional* in Figure 7.23, where each label applies to a single LU (or dictionary

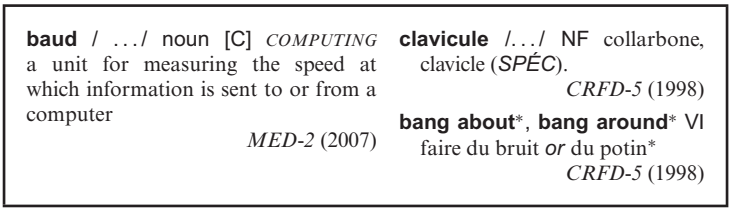


Fig 7.25 Scope of various labels

sense). The label 'often disapproving', coming at the very top of the first LU immediately after the section marker ('1'), applies to everything up to the next section marker ('2'). Similarly, the label 'usually before noun' applies to everything in that LU section. These labels mean that the phrase *conventional methods* has no disapproving overtones, and that you're unlikely to hear **these methods are conventional*.

Labelling in bilingual dictionaries is twice as complex as labelling in monolingual dictionaries, since both an SL example and its TL equivalent need to be labelled if they are not 'unmarked' (§6.4.1.4). This is the case in the entry for *clavicule* from the *CRFD* (Figure 7.25), where the French headword *clavicule* is unmarked but has two equivalents in English, one unmarked (*collarbone*) and one which is a specialist medical term (*clavicle*); the latter carries the 'specialist' label. The Style Guide for this dictionary dictates that a label referring to a TL item comes after it.

Sometimes both SL and TL items need labelling, as in the entry for *bang about/around* in the same dictionary, where both forms of the English phrasal verb are informal (in this dictionary the asterisk marks an informal item) as is the French *faire du potin* (but *faire du bruit* is not, and so isn't labelled). However, you could take another bilingual dictionary down from the shelf and you will find it follows different rules for source- and target-language labelling.

➔ Remember, it's the *positioning* of the label that determines its scope.

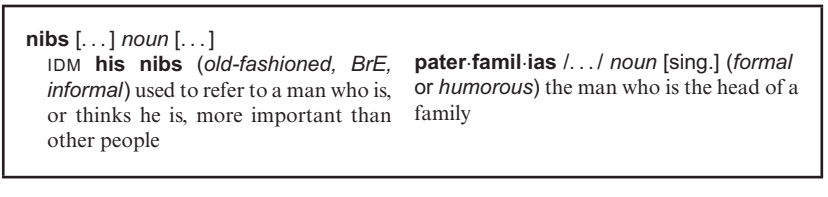


Fig 7.26 Multiple labelling from *OALD-7* (2005)

Multiple labelling When two or more labels are attached to one item, then there are two possible interpretations. A good dictionary will make clear how the labels should be read, and the Style Guide will tell lexicographers how to handle multiple labels. The options for multiple labels (illustrated in Figure 7.26) are:

- to be read as 'X *and* Y *and* Z' (as in *his nibs*)
- to be read as 'X *or* Y' (as in *paterfamilias*).

➔ When you attach two labels to one item, you should always make it clear whether they stand in an 'and' or an 'or' relationship to one another.

7.2.9 Usage

Entry components carrying information about usage are a feature of most dictionaries, with a more significant presence in dictionaries – monolingual and bilingual – for language-learners. Each dictionary has its own approach to usage notes (called variously 'usage', 'synonyms', 'metaphors', 'functional note', 'false friends', etc.) and at the planning stage the editors decide which particular types of usage notes to include. Their aim is to tell their users what they need to know, even when this will not fit the model of the traditional dictionary entry, and also of course to come up with some added value that will give them the edge over their competitors. So no one is the loser here. Teachers in particular find these notes useful in preparing lessons. It's worth comparing the choice over a slew of dictionaries before deciding what to include in your own. We describe two types of usage note in this section, the first with a broad range of relevance throughout the dictionary and the second focusing on the headword of the entry to which it is attached.

7.2.9.1 *Subject-oriented usage note* This type of note has as its focus a group of words relating to one subject, and it is normally cross-referenced from all the headwords it applies to. It's a useful way of avoiding repeating the same information in entries all over the dictionary. One example of this type of usage note is drawn from the *OHFD-3* (2001) and concerns how to translate into French various constructions containing names of countries and continents. Part of that quite long note, located near the entry for *country*, is shown in Figure 7.27.

Countries and continents
■ *Most countries and all continents are used with the definite article in French.*
France is a beautiful country
= la France est un beau pays
I like Canada
= j'aime le Canada
to visit the United States
= visiter les États-Unis
to know Iran
= connaître l'Iran
A very few countries are not:
to visit Israel
= visiter Israel
■ *When in doubt, check in the dictionary.*
■ *All the continent names are feminine in French. Most names of countries are feminine, e.g.* la France, *but some are masculine, e.g.* le Canada. [. . .]
OHFD-3 (2001)

Fig 7.27 Example of subject-oriented usage note

➔ It's a good idea to list topics for these notes at the beginning of a dictionary project, then collect information to go into each during the first year or two of editing, marking entries to be cross-referenced later, when the notes themselves are drafted (a handy job for an interested academic colleague, working with an editor of course).

7.2.9.2 *Local usage note* Local usage notes can contain many different types of information relating specifically to the headword of the entry where they are found. Figure 7.28 contains four examples of these: the sample usage note from the *MED* is fairly standard, pointing out the difference in usage between the headword *although* and its synonym *though*; the *OALD* note on *ask* is more daring, spelling out a wrong usage, scored through to emphasize the point. The second *OALD* note in the *beat* entry contains a useful contrastive account of some of its near-synonyms. The note at the head of the *OHFD into* entry gives the English-speaking user some general advice about how to put this preposition in French.
➔ It's particularly important when writing usage notes to choose the information and the wording according to your reader's language and dictionary skills. In bilingual dictionaries you have to decide first of all whether you are writing the note for the SL or the TL speaker.

although /.../ conjunction ***
1 used for introducing a statement that makes your main statement seem surprising: *Although he's got a good job now, he still complains.* [...]
Though is used with the same meaning as **although**, and is more common in spoken English.

MED-2 (2007)

into /.../
Δ *Into* is used after certain nouns and verbs in English (*way into, change into, stray into* etc). For translations, consult the appropriate noun or verb entry (**way, change, stray** etc).
Into is used in the structure *verb +sb + into + doing* (*to bully somebody into doing*, *to fool somebody into doing*). For translations of these structures see the appropriate verb entry (**bully**, **fool** etc). For translations of expressions like *get into trouble, go into detail, get into debt* etc you should consult the appropriate noun entry (**trouble**, **detail**, **debt** etc).
prep 1 (indicating change of position, location) dans; [...]

OHFD-3 (2001)

ask /... / *verb, noun*
■ ***verb***
▶ **QUESTION 1 ~ (sb) (about sth)** to say or write sth in the form of a question [...]
HELP You cannot say 'ask to sb' : ~~*I asked to my friend what had happened*~~.
▶ **REQUEST 2** to tell sb that you would like them to do sth [...]

OALD-5 (1995)

beat /.../ *verb*, *noun*, *adj.*
■ ***verb*** (**beat, beaten** /bi:tn/)
▶ **IN GAME 1** [VN] **~ sb (at sth)** to defeat sb in a game or competition: *He beat me at chess.* [..]
SYNONYMS
beat
batter ♦ pound ♦ pummel ♦ lash ♦ hammer
All these words mean to hit sb/sth many times, especially hard.
beat to hit sb/sth a lot of times, especially very hard: *Someone was beating at the door.*
batter to hit sb/sth hard a lot of times, especially in way that [..]
lash to hit sb/sth with a lot of force: *The rain lashed at the window.*
NOTE The subject of **lash** is often *rain*, *wind*, *hail*, *sea* or *waves*. [...]

OALD-5 (1995)

Fig 7.28 Local usage notes in learners' dictionaries

7.2.10 Other lemmas within the entry

Within the broad scope of an entry, there are three principal components that carry information about a word related to the entry headword. The first two – secondary headwords and run-ons – tend not to be used so much in learners' dictionaries, the idea being that learners have enough trouble finding what they want without having to burrow around in an entry of a headword that is not the object of their search. The third – cross-references – is fairly standard in most dictionaries.

7.2.10.1 *Secondary headword* Both the secondary headword (also called a subheadword) and the run-on are components whose target is a word

shrug […] 1 N haussement *m* d'épaules; **to give a ~ of contempt** hausser les épaules (en signe) de mépris; […] 2 VI **to ~ (one's shoulders)** hausser les épaules
▶**shrug off** VT SEP [+*suggestion, warning*] dédaigner, faire fi de; [+ *remark*] ignorer, ne pas relever; [+ *infection, a cold*] se débarrasser de.

CRFD-5 (1998)

secondary headword

naked (…) *adj* **1** having the body completely unclothed; undressed. Compare **bare**[1]. **2** having no covering; exposed: *a naked flame.* **3** with no qualification or concealment; stark; plain: *the naked facts.* […] **11b** lacking some essential condition to render valid; incomplete. […] ▶ **'nakedly** *adv* ▶ **'nakedness** *n*

CED-5 (2000)

run-ons

Fig 7.29 Secondary headword and run-on components

or MWE other than the headword of the entry; they both follow on at the end of the entry, often flagged by something like the ▶ symbol in the Collins entries in Figure 7.29. The difference between these components is that the secondary headword heads what is virtually a full entry (only the pronunciation is missing in the *CRFD shrug off* subentry), while nothing but the wordclass is normally given for run-ons. Derived forms of the headword (adjectives formed from noun headwords and the like) do occur as secondary headwords, but not so much nowadays, since the emphasis in dictionaries has shifted away from packing as much information as possible into the entry, regardless of the poor user. Lemmas given secondary headword status are mainly MWEs: phrasal verbs (as *shrug off* here), compounds in which the headword is the first element, and idiomatic MWEs, though all of these are full headwords in many modern dictionaries. Much research has been done in academia in an attempt to discover where people look up MWEs, but no clear-cut view has emerged. (The only certain fact is that native English speakers have no idea what a phrasal verb is, and often hunt in vain for *come out* within the *come* entry.)

➔ What to make into secondary headwords is a problem for the editors at the planning stage: during the writing of the dictionary the Style Guide should tell you what to do here.

7.2.10.2 *Run-on* A run-on is the section of a dictionary entry which holds infrequent derived forms of the headword, such as *nakedly* and *nakedness* in the entry from *CED* in Figure 7.29. There is rarely any indication of the relationship between headword and run-on, but native speakers can hopefully be relied on not to compose sentences like **he leapt nakedly into the pool* or **the candle burned nakedly*, on the basis of a dictionary

entry.[15] There is no indication, either, that the noun *nakedness* is semantically linked with the first sense of the adjective, as well as being open to metaphorical interpretations.[16] It is easy to see how this form of entry could cause problems for language-learners, and run-ons need to be used with care. Ideally, they will only be used in monolingual dictionaries when:

- the word form is infrequent
- its meaning is unambiguously deducible through the application of basic word-formation rules
- its pronunciation can be predicted from the pronunciation of the headword it is attached to
- its grammatical and collocational behaviour is simple and predictable.

Thus, in most monolingual dictionaries, *homelessness* appears as a run-on at *homeless*, because it fulfils all these criteria. But *homeless* itself – though composed by adding the suffix *-less* to the word *home* – is too frequent to be handled in this way, and it also shows signs of unpredictable behaviour (it is often nominalized in the expression *the homeless*). Adverbs formed by adding *-ly* to the related adjective are one of the commonest types of run-on, but care needs to be taken that the meaning and use are unambiguous. Some adverbs of this type are used as intensifiers and may have a different range of collocates from the related adjective: thus *flatly* often modifies *refuse* and *deny*, but *flat* rarely appears with *refusal* or *denial*. Some 'derived' adverbs have several meanings (*thinly* has three senses in *LDOCE-4* and *MED-2*), which don't always correspond closely to the senses of the adjective they are derived from (the use of *thinly* in 'a thinly veiled insult' has no obvious connection with any of the numerous senses of *thin*). Or again, some words double as manner adverbs (*talked frankly about her concerns*) and as sentence adverbs, or 'stance adverbials' (*frankly, I couldn't care less*). So – especially in the case of learners' dictionaries – it's best to avoid anything other than simple manner adverbs (like *accurately*, *acrimoniously*, and *amateurishly*) and unambiguous nominalized forms (like *homelessness*, *pedestrianization*, or *indigence*) in the run-on slot.

[15] The adverb *nakedly* occurs just 14 times in the 100-million-word BNC, normally modifying an adjective: examples are *nakedly financial motives*, and *nakedly behaviouristic theories.*

[16] As in these examples from the 131 occurrences in the BNC: *men and women joyous in nakedness coming together under the full blossom of trees*, and *true religious feelings clothe the nakedness of theory with practice*.

➔ Remember when planning a dictionary that the user should come first: the need for extensive cross-referencing from run-on to the various senses of the main headword makes this component inappropriate for learners' dictionaries.

iron (. . .) *n* **1a** a malleable ductile silvery-white ferromagnetic metallic element [. . .] ♦ *vb* **15** to smooth (clothes or fabric) by removing (creases or wrinkles) using a heated iron [. . .] ♦ See also **iron out, irons**.

necktie (. . .) *n* the US name for **tie** (sense 11)

Persia (. . .) *n* **1** the former name (until 1935) of **Iran**. **2** [. . .]

Iran (. . .) *n* a republic in SW Asia, between the Caspian Sea and the Persian Gulf [. . .] Former name (until 1935) **Persia** [. . .]

Fig 7.30 Cross-references in *CED-5* (2000)

7.2.10.3 *Cross-reference* The cross-reference component tells the user that more information relating to the current headword will be found at the other entry, and as Figure 7.30 shows, this can be done in a number of different ways, both directly and indirectly. In the *iron* entry, users are alerted by the direct cross-reference ('See also . . .') to the presence of two headwords further down the list, the phrasal verb *iron out* and the plural noun *irons*, which might otherwise have escaped their notice. The bold type in the *necktie* entry tells the user that the definition of *necktie* is to be found in sense 11 of the *tie* entry. The entries for the headwords *Persia* and *Iran*, which name the same country across a timespan, carry implicit mutual cross-references. Every dictionary has its own palette of admissible ways of cross-referring from one entry to another. An automated cross-reference check is now the last step before the dictionary is finally put to bed.

➔ It's usually best in the first compiling pass of a dictionary project to include all the cross-references you're likely to need; a lot of them get dropped for reasons of space in the printed book, but they are useful in the draft text and facilitate consistency checking.

7.2.11 The electronic dictionary entry

The advent of the electronic dictionary (henceforth, e-dictionary) has made possible a number of new types of entry component.[17]

[17] The field is developing so quickly that what we write today is almost guaranteed to be out of date by the time this book appears.

Box 7.5 A look at the timeline of e-dictionary development

Past: The first electronic dictionaries consisted simply of the original print text equipped with a search engine; some of these search engines were very basic, but using them was still faster than looking up a book.

Present: The electronic editions of contemporary dictionaries offer a good deal more than the print text (as we see in §7.2.11.1). Search functions have become more powerful, the dictionary may be viewable in more than one mode, and data such as wordclass markers and grammar codes are presented in more user-friendly ways (for example with abbreviated forms fully spelled out, and spoken pronunciations in place of IPA symbols). Most importantly, they increasingly include more content than the print edition, for example by giving access to other dictionaries or by providing additional example sentences.

Future: A new dictionary designed for electronic as well as print publication – a rare bird in the reference publishing world, because of the cost involved – opens exciting possibilities of totally new information presented in new ways. Key features of such a dictionary will be 'customizability' and 'personalizability': in this model, the 'dictionary' is essentially a collection of lexical resources (possibly multilingual), which users can select from and configure according to their needs.

In this section, we offer a basic overview of the dictionary in electronic form:

- first, a quick look at one of the best current examples of the e-dictionary;
- then some instances of how the e-dictionary presents standard information in new and interesting ways;
- and finally, some suggestions for people designing a wholly new e-dictionary.

People tend to think that with the advent of the e-dictionary all our space problems are solved; they propose clever and sophisticated ways of assembling and presenting existing and new information and we all get very excited about this way of producing dictionaries. But we mustn't lose sight of our users: we need to be clear about the difference between doing things just because we can, and doing them because they will be of real value to the user.

➔ Don't assume you can simply give users every fact you know about a word: information overload sets in very rapidly. Devising an e-dictionary calls for smart information management and sensitive design on the part of the editors and the software engineers.

Fig 7.31 The e-LDOCE opening screen after *camp* has been keyed in

7.2.11.1 *Introducing the e-dictionary* The purpose of this section is to introduce the e-dictionary to those readers who are not already familiar with one. (If you already use an e-dictionary, you can probably skip this.) The electronic *LDOCE*, an updated edition of *LDOCE-4* (2003), is at the time of writing a state-of-the-art electronic dictionary. Marketed on a CD-ROM packaged with the print dictionary, it has a pleasantly welcoming appearance and a wealth of well-thought-out features. This e-dictionary's look-up screen contains a different sample entry each time you open it, but your cursor is firmly placed where it should be for you to type in your search word. Figure 7.31 shows the screen after the search word *camp* has been inserted. At once, we see the difference between this dictionary and its print sister: here you have a choice from a list containing not only the

three homograph headwords for *camp* (the noun, verb, and adjective), but all the compound words containing *camp* (*aide-de-camp*, *boot camp*, etc.) which are headwords in their own right in the print dictionary.

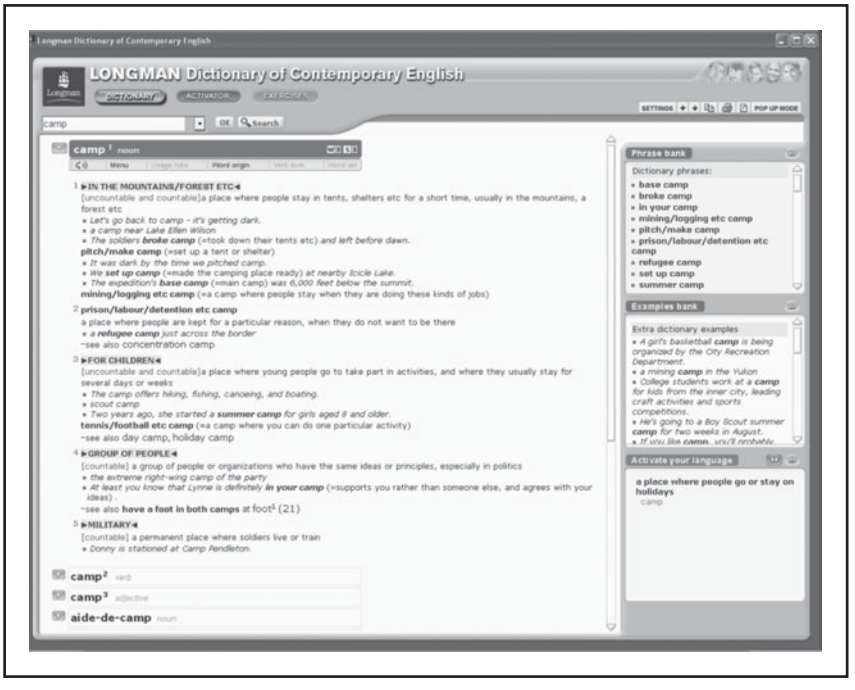


Fig 7.32 The main entry screen for the noun *camp*

When you hit the hyperlink for the noun *camp*, the screen shown in Figure 7.32 appears (note that the other *camp* headwords are still visible at the foot of the screen). Here we see what could be called the primary components of this e-dictionary:

- The main entry, consisting of the various LUs, each containing a definition, grammatical information, examples, and any relevant MWEs. (The contents of this main entry constitute the full print dictionary entry for *camp*[1] (noun) with the exception of the IPA pronunciation; everything else on the screen is added value in the e-dictionary.)
- A list of hyperlinks (below the main entry) to the other *camp* entries in the dictionary.
- A 'Phrase bank' where all the MWEs containing camp are listed; the list contains MWEs from other entries as well as the current one, e.g. *camp fire*, *to pitch camp*, *to camp it up*, *camping gear*, and so on.

- An 'Examples bank' offering more examples of the noun *camp* in action in the corpus.
- An 'Activate your language' hyperlink to material from the *Longman Activator* relating to places where people go on holiday (resort, cruise, tourist attraction, etc.).

The last three features, in frames down the right-hand side of the screen, all carry the option to be opened in a new window.

Icons on this screen link to other types of information. In the top bar are the links to the Longman ACTIVATOR and to a set of EXERCISES for learners of English, with titles like 'Articles', 'Collocations', 'Countable and uncountable nouns', etc.

In the bar containing the headword *camp*[1] (noun), the boxes W3 and S3 indicate that the headword is one of the 3,000 most frequent written and 3,000 most frequent spoken words.

Below that, clicking on the icon offers several options: British or American spoken pronunciations, recording your own pronunciation, or playing it back – all of them added value over the print dictionary. Clicking on 'Menu' in that bar opens a summary of the entry in the form of a list of the mnemonics that introduce each sense ('IN THE MOUNTAINS/FOREST ETC.', 'prison/labour/detention etc. camp', 'FOR CHILDREN' and so on) – not needed perhaps for an entry like this one which fits neatly into one screen, but useful for the very long entries. Clicking on 'Word origin' opens a new window showing the etymology of *camp* (information not in the print dictionary at all). The other options in that bar, 'Usage note', 'Verb form', and 'Word set', are greyed out, showing that there is no *camp*-related material to offer of these types.

So much for the various types of lexical information in this e-dictionary. However the CD-ROM contains other useful material, principally aimed at teachers of English as a foreign language. This includes:

- 'Teachers' Resources': including worksheets on etymology, dictionary training, education-related vocabulary, job-related vocabulary, and vocabulary related to weather and the environment, prepositions, register in language, and many others.
- 'Students' Activities': containing links to three web-based 'activities' that are renewed on a monthly basis: for example, an exercise on 'Related Words' (such as *retailer* and *wholesaler* and *forceful* and *pushy*), or another exercise on 'Suffixes' where the student is asked to

insert into the slot the correct suffix to make sense of the sentence (for instance, *That picture is completely worth____. You wouldn't get a penny for it.*).

- 'Competitions': also web-based and student-oriented.
- 'New words': renewed weekly (currently the new word of the week is the noun *freerunning*).
- 'Articles': about some aspect of the dictionary (renewed monthly, the current title is 'Word combinations in the LDOCE').
- 'CD-ROM': offering some useful facts about the CD, and a guided tour.
- 'Game': web-based, currently about word combinations and offering a prize.

And three sales-oriented topics:

- 'About the dictionary': the e-equivalent of the blurbs on the covers of the print dictionary.
- 'Companion websites': containing hyperlinks to other Longman websites.
- 'Catalogue': a hyperlink to Longman's online catalogue.

7.2.11.2 *New ways of accessing standard information* The e-LDOCE screen shown in Figure 7.32 illustrates a certain hierarchy of entry components that holds good for print dictionaries too. From the e-dictionary (monolingual or bilingual) we see that some components (e.g. wordclass, definitions, translations) are absolutely central, while others (e.g. etymology, pronunciation, inflected forms) are more peripheral. These latter are often hidden in the main screen of the dictionary, available to be called up if wanted.

More importantly, the e-dictionary also offers complex entry components which simply cannot be realized in print form, for instance:

- spoken pronunciations (in addition to the IPA transcriptions)
- search responses combining information from more than one dictionary entry, e.g. all the uses of a specific word in the dictionary, whether these occur in its own entry, or in multiword expressions or illustrative phrases in other entries, or indeed – in bilingual volumes – in the translations in the other 'half' of the dictionary.

The e-LDOCE described above in §7.2.11.1 provides many examples of such complex search responses.

➔ People will love your dictionary if it's easy and quick to use. Some very clever dictionaries take a long time to come up on the screen and won't fold themselves away without asking irritating questions.

7.2.11.3 *New types of information in the entry* The e-dictionary also provides new and different information that can't be contained in a print dictionary, usually for reasons of space: this innovative approach is currently seen at its best in the electronic versions of learners' dictionaries, particularly the monolingual. There are a number of examples of this in the e-LDOCE entry, for instance:

- word origins (etymology)
- complete inflections for every verb (regular or otherwise)
- a group of expressions (in the 'Activate your language' frame) with similar meanings to the headword
- a set of complete sentences drawn from the corpus (in the 'Examples Bank').

Such is the flexibility of the electronic medium that there is no conceivable upper limit to the new components you can devise, and some of these are mentioned in §7.2.11.4 below. The constraints are the cost of development (as always), and also the risk of an entry so complex that users would find it impossible to navigate.

7.2.11.4 *Looking to the future: some ideas* The corollary of the exciting new electronic possibilities is that the user interface is crucial. It's easy to make it too rich and complicated and confusing, to leave the user lost in hypertext space, to bombard learners with facts they don't understand, to make the display so 'attractive' that the screen takes an age to change.

Here are some ideas for the custom-built e-dictionary (for when your publisher says, 'Come up with some brilliant new ideas, expense is no object'):

- Beef up the navigation functions
  Working with hypertext, you have to try to avoid the 'How did I get here?' syndrome. You might think of incorporating:

  - in every new tab, a 'Go to Entry' icon taking you back to the main screen (some good e-dictionaries do this already)
  - an arrowed 'route' or tree diagram in a separate tab or window set top right of screen summarizing previous searches.
- Plan carefully how to display query results
  - Put them in a new tab, or attach an additional 'screen' to the current display, etc.
  - It's also important to indicate clearly when a search draws a blank.
  - Make it easy for the user to handle the dictionary text by including such functions as print, print to file, place on clipboard, etc. (in e-dictionaries that do allow this, the formatting of the resulting text is often very frustrating).
- Build multiple user profiles, and let users customize their e-dictionary
  Users have their own specific needs and skills (and these may change according to the task they are engaged in), so it is important to allow them to decide which information-categories should be displayed by default (and which can be accessed by an additional click).
  - Try to mask advanced material (essential for skilled linguists) from the normal 'learner' user.
  - In an ideal e-world, the user could complete an introductory dialogue, which would be used to set a 'level' of linguistic skills, or an 'interest-profile' (child vs. adult, engineer vs. humanities student, etc.), and to filter the output of searches in the dictionary. But that's a few years away yet.
- Use the user profiles to enhance the e-dictionary
  - Draw up a typology of users, noting a few distinctive user types that focus particularly on users' skills and needs.
  - List operations most likely to be performed by these typical users, including:
    - spelling a word
    - understanding a word
    - differentiating a word from one it's often confused with
    - self-expression in a foreign language
    - understanding how a word or phrase is used 'naturally'
    - translating into their own language
    - translating out of their own language, etc.
  - Consider each of these operations in relation to each of the user types: what functions would most help each type of user?

  - Support new functions by linking dictionary text to SL and TL corpora in various ways:
    - Produce sets of corpus sentences including the headword (if possible, in each of its senses, i.e. a separate set for each LU).
    - Produce lists of words that function semantically like keyword, together with supporting corpus sentences.
    - Produce lists of words that function grammatically like keyword, together with supporting corpus sentences.
  - Include other search conditions to filter the output of the above, on the basis of corpus frequency, any of the labels (domain, style, register, etc.), the language involved, and so on.
- Devise more sophisticated search possibilities
  - Allow users to input (or select from text) a word or phrase (in the case of bilinguals, in either SL or TL) to be used as the 'focus' for the next step, which is . . .
  - Perform some operation on it, such as pronouncing it, translating it, finding it in specific types of corpus text, showing inflected forms, finding near-synonyms, antonyms, etc., finding other corpus contexts (etc.).

➔ When designing an electronic dictionary, remember that it's easy for people to get lost, or distracted, in the maze of different functions that appear on the screen – most users still just want to know what the word means, and (some of them at least) how to use it. Avoid clutter on the screen. Most users will be familiar with search engines like Google, which present the first layer of information in simple, stripped-down form, allowing users to decide which additional information they want to see. This is a good model to keep in mind.

7.3 Entry structure

The senior editors specify the content and layout of the various types of entry in a dictionary (the microstructure) during the planning stage. Their decisions are then spelt out in the Style Guide and implemented by the lexicographers writing the dictionary. This section gives a brief overview of the principal options in microstructure design – it's useful to know how these things come about.

7.3.1 The basic classifying principle: the first 'cut'

The most far-reaching decision relates to the primary 'cut' through the information.[18] Will this be made:

(a) on the basis of grammar (its various wordclasses), or
(b) on the basis of meaning (the major senses of the headword)?

For instance, for the headword *haunt*, the options would be:

(a) to divide it first according to wordclass
This version (the commonest in practice) presents all of the verb uses of *haunt* before any noun uses, regardless of the varying semantic distance between these LUs.
(b) to divide the material first according to the broad sense blocks.
In this version, the entry for *haunt* is divided into broad meaning areas. The second of these – the idea of people frequently returning to a particular place – manifests itself both as a verb and as a noun, so these two uses are grouped together.

Figure 7.33 shows the difference the layout makes in an entry for *haunt*.

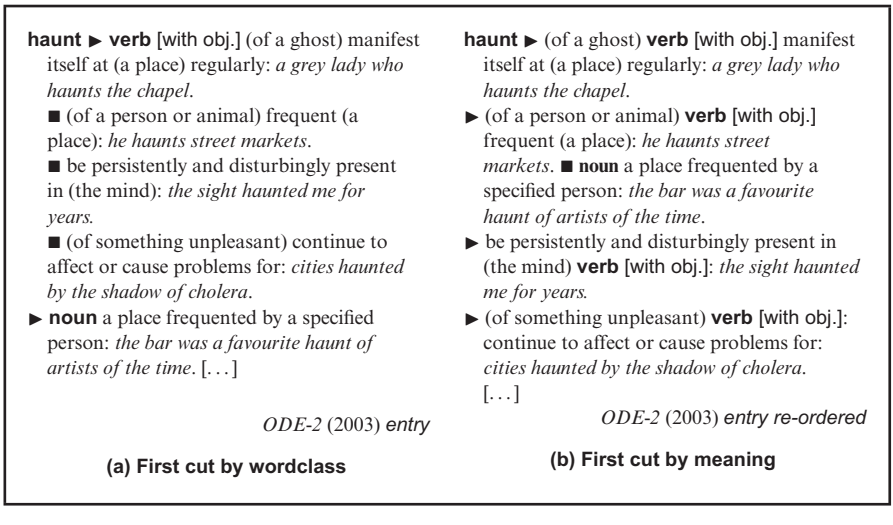


Fig 7.33 Grammar-based and meaning-based organization of the same entry

[18] Some dictionaries split the word into wordclasses and make each major wordclass a homograph headword. The decision on whether to have homograph headwords, and if so the basis on which they should be made, has already been taken at this point, being part of the macrostructure decisions, cf. §6.5.3.

The pros and cons of each format, whether used in monolingual or bilingual dictionaries, are:

(a) Based initially on wordclass:
- It is the more usual way of handling dictionary entries, so most users will be familiar with it.
- As an arbitrary access system (like alphabetical order), it can be applied objectively and systematically.
- There is some psycholinguistic evidence that wordclass is one of the categories we use for storing and accessing words in our mental lexicons (e.g. Aitchison 2003: 112).
- It offers skilled linguists speedy access to information, but . . .
- It's usable only by people who know what nouns, verbs, adjectives, and adverbs are and who can use this information to help them search through a complex entry.[19]
- Its major disadvantage is that very similar meanings may be far apart in the entry, especially when one meaning is realized by two or more wordclasses.
- This frustrates users who do have intuitions about the meaning of the word.

(b) Based initially on meaning:
- It is modern and innovative, but could be off-putting for traditionalists.
- It's instantly intelligible to the speaker of the source language, so it's fine for use in monolingual dictionaries for native speakers (who may be less than confident about terms like 'adjective' and 'verb') or in bilinguals written solely for SL speakers, but . . .
- It is not helpful for language learners or for anyone who doesn't already know the meaning(s) of the word they're looking up.
- Since 'meaning' is a less clear-cut category than wordclass, applying this policy sometimes requires lexicographers to make subjective judgments.

[19] Research into dictionary use by school and university students shows that the majority of these users don't understand these terms, see Atkins (1998). However, some do, and since most people look up words in order to find their meaning, it is arguably not a good idea to base the ordering of the entry on the various meanings of the headword.

- It tends to produce slightly longer entries, and over the whole dictionary could add approximately 2% (or 20 pages in a 1,000-page dictionary).

➔ When you're making decisions like this, which affect the whole impact and appearance of the dictionary, it's as well to do a bit of market research first, to see what your probable readers prefer. (Some lexicographers prefer (b) because it's more satisfyingly logical, but then it's usually only lexicographers who read linearly right through an entry.)

7.3.2 Flat or tiered senses

Dictionaries generally use a 'flat' structure to present the meanings of polysemous words: the various senses (LUs) are simply numbered 1, 2, 3, and so on. But word meanings don't always divide up as neatly as this structure appears to imply. In some cases, two meanings may be closely related, while a third and fourth may be quite distinct. This is often true, for example, of words that exhibit some form of 'regular polysemy' (cf. §5.2.4). Thus, the word *glass* can refer to the substance itself, to a mirror (a dated use), to a drinking container made of glass, or to the contents of a drinking container (*I only drank two glasses of wine*). The last two meanings are related by regular polysemy, and are clearly much closer to one another than they are to the other two. While some dictionaries treat every meaning as equally distinct (the entries for *glass* in *LDOCE* and *AHD*, for example, have a simple flat structure), others use a 'tiered' entry structure which recognizes – and tries to reflect – the variations in 'semantic distance' between a word's various uses. A tiered structure allows us to tuck *subsenses* into 'main' senses, and number them accordingly, e.g. 1a, 1b, 2, 3a, 3b, 3c, 4, and so on. In the entries for *glass* in *ODE* and *MED*, the LU for 'the contents of a glass' is shown as a subsense and nested under the main sense 'a drinking container made from glass'. Figure 7.34 shows how the same semantic content can be presented in flat and tiered entry-structures.

➔ Remember when deciding on 'flat' versus 'tiered' that if your readers are unlikely to notice the difference they won't understand the subtleties of tiered senses: dictionaries for school students usually go for flat structure.

necessary (. . .) *adj* **1** needed to achieve a certain desired effect or result; required. **2** resulting from necessity; inevitable; *the necessary consequences of your action*. **3** Logic. **3a** (of a statement, formula, etc.) true under all interpretations or in all possible circumstances. **3b** (of a proposition) determined to be true by its meaning, so that its denial would be self-contradictory. **3c** (of a property) essential, so that without it its subject would not be the entity it is. **3d** (of an inference) always yielding a true conclusion when its premises are true; **3e** (of a condition) entailed by the truth of some statement or the obtaining of some state of affairs. ♦ Compare **sufficient** (sense 2). **4** Philosophy. (in a nonlogical sense) expressing a law of nature [...]

CED-5 (2000) *entry*
Tiered structure

necessary (. . .) *adj* **1** needed to achieve a certain desired effect or result; required. **2** resulting from necessity; inevitable; *the necessary consequences of your action*. **3** Logic. (of a statement, formula, etc.) true under all interpretations or in all possible circumstances. **4** Logic. (of a proposition) determined to be true by its meaning, so that its denial would be self-contradictory. **5** Logic. (of a property) essential, so that without it its subject would not be the entity it is. **6** Logic. (of an inference) always yielding a true conclusion when its premises are true; **7** Logic. (of a condition) entailed by the truth of some statement or the obtaining of some state of affairs. ♦ Compare **sufficient** (sense 2). **8** Philosophy. (in a nonlogical sense) expressing a law of nature [...]

CED-5 (2000) *entry renumbered*
Flat structure

Fig 7.34 Same partial *CED* entry in tiered and flat structures

7.3.3 Secondary ordering of dictionary senses

Once you've decided on the basic topography (grammar-led or meaning-led sections, flat or tiered entry-structure), you have to make sense of the rest of the entry. A good Style Guide should specify the criteria according to which a word's various meanings will normally be ordered. (You can't be too inflexible about this sort of thing: it's always better to end up with a sensible entry than a weird one that follows the rules blindly.) There are three common ways to choose from:

(1) Historical order
This method presents the senses of a headword in the order in which they entered the language; provided you have adequate information about a word's development over time, this is the easiest system to apply.

(2) Frequency order
The senses are ordered on the basis of their frequency in the corpus. The attraction of this method is its apparent objectivity. Further, it can plausibly be argued that the meanings which are encountered most frequently are the ones that users are most likely to look up – so it makes sense to show them first. This is a persuasive argument in the case of dictionaries aimed at language-learners; though less so if

the readership is mainly native speakers. (These users are probably looking up a more obscure meaning, and might be served better if meanings were presented in reverse frequency order, but publishers are strangely unwilling to enter that territory.) In practice, this frequency-based approach is a good deal less straightforward than it sounds. First, it requires a well-balanced corpus (a major challenge in itself: cf. §3.4.2.3). Second, determining the relative frequencies of the meanings of a polysemous word can never be an exact science because word senses are not objectively stable entities. Lexicographic software is not (at the time of writing) sufficiently sophisticated to give a reliable account of sense frequency, and although it's often possible to identify the most common senses 'manually', there are plenty of cases where this is not easy to do (think of a word like *party*: are political parties more frequently referred to than social ones?).

(3) Semantic order, with 'core' meaning first
A word's core meaning (sometimes referred to as its 'psychologically salient' meaning) is the one that feels, intuitively, to be central to any understanding of how the word works and how its other meanings have developed. The core meaning tends to be the one you think of first, and (a related point) is usually the one you learned first as a child. Thus the core meaning of *reach* refers to stretching out a hand or arm to make contact with something, even though some of the word's other uses may be encountered more often. (A word's core meaning – as in the case of *reach* – may coincide with its original meaning, but this is by no means always the case.). In this ordering system, the core meaning is followed by those meanings that are semantically closest, with more marginal uses appearing later. This is a compromise solution, and the least 'scientific' of the three ways of ordering senses. But (except in the case of historical dictionaries) this is the method that most dictionaries favour, partly because it is relatively easy to apply, and partly because it is felt to give the user the most satisfying account of meaning.

An example of what these options entail is seen in Figures 7.35 and 7.36. In the case of the word *icon*, the original and 'core' meanings coincide, and this sense is given priority in the *ODE*. Meanwhile, the learners' dictionary (*LDOCE*) shows the commonest meaning first, presumably on the grounds that this is the one its users are most likely to look up.

icon /…/ (also **ikon**) **noun** a devotional painting of Christ or another holy figure, typically executed on wood and used ceremonially in the Byzantine and other Eastern Churches.
■ a person or thing regarded as a representative symbol or as worthy of veneration: *this iron-jawed icon of American manhood.* ■ Computing a symbol or graphic representation on a VDU screen of a program, option, or window. […]

ODE-2 (2003)

Ordered with core meaning first

icon /…/ *n* [C] a small sign or picture on a computer screen that is used to start a particular operation: *To open a new file, click on the icon at the top of the screen.* **2** someone famous who is admired by many people and is thought to represent an important idea: *a 60s cultural icon*. **3** also **ikon** a picture or figure of a holy person that is used in worship in the Greek or Russian Orthodox Church.

LDOCE-4 (2003)

Ordered by frequency

Fig 7.35 Different ordering of LUs in entries for *icon*

An alternative interpretation is that the first meaning shown is 'core' for each dictionary's intended user: the main user-group for learners' dictionaries is young adults (typically in the range 16–24 years old), and among this cohort a computer icon is a familiar concept whereas the 'devotional painting' sense may be entirely unknown.

zombie /…/ *n.* & *v.* […] *n.* **1** Orig., a snake-deity in voodoo cults of or deriving from W. Africa and Haiti. Now (esp. in the W. Indies and southern US) a soulless corpse said to have been revived by witchcraft. E19. **2** A dull, apathetic, unresponsive, or unthinkingly acquiescent person. colloq. M20 […]

New Shorter Oxford Dictionary (1993)

zom|bie /…/ zombies **1** You can describe someone as a **zombie** if their face or behaviour shows no feeling, understanding or interest in what is going on around them. N-COUNT
□ *Without sleep you will become a zombie at work.* **2** In horror stories and some religions, a **zombie** is a dead person who has been brought back to life. N-COUNT

COBUILD-5 (2006)

Fig 7.36 Different ordering of LUs in entries for *zombie*

Figure 7.36 shows an example of historical ordering. In the entry for *zombie* in the *New Shorter Oxford Dictionary*, we see a clear chronological progression from what is historically the first meaning (a Voodoo snake deity), its extension to denote a dead body reanimated by witchcraft (now generally felt to be the 'core' meaning), and the transfer from there to denote a slow, dull, apathetic person (currently the most common sense). The entry from *COBUILD*, on the other hand – a dictionary designed for language learners – gives priority to the sense most frequently found in their corpus, and omits the original meaning altogether.

➔ You have to say something in the Style Guide about the order in which LUs should be handled, but hard-and-fast rulings are impossible. Ordering

the LUs in a sensible and coherent manner is a challenge to dictionary writers, but we have never met any dictionary users (as opposed to metalexicographers and computational linguists) who complained of this aspect of our work.

7.3.4 Location of multiword expressions[20]

All dictionaries – monolingual or bilingual – must decide where in the ordering of the entry should go compounds, phrasal verbs (for English etc.), and other MWEs, if they are to be included within the entry of one of their component words. They are often treated as secondary headwords, or may be located in a separate section of the entry, entitled 'Compounds' or 'Phrases'. Another option is to give them a separate entry distinct from any related entry.

leaf /…/ A *n* (*pl* **leaves**) 1 (of plant) feuille *f*; **dock/oak/lettuce** ~ feuille de patience / de chêne / de salade; **autumn leaves** feuilles d'automne; **to come into** ~ se couvrir de feuilles; 2 (of paper) feuille *f*; (of book) page *f*, feuillet *m* spec; 3 (of gold, silver) feuille *f*; 4 (of table) (sliding, removable) rallonge *f*; (hinged) abattant *m*.

B **-leafed, -leaved** (*dans composés*) **red-~** à feuilles rouges; **broad-~** à grandes feuilles.

Idioms **to shake like a** ~ trembler comme une feuille; **to take a ~ out of sb's book** s'inspirer de quelqu'un; **to turn over a new** ~ tourner la page.

Phrasal verb ▪ **leaf through**: ▶ ~ **through** [**sth**] feuilleter [*pages, papers, book, magazine*]; parcourir [*introduction*].

leaf bud *n* bourgeon *m* à feuilles; […]

OHFD-3 (2001)
MWEs in separate blocks
with compounds as headwords

leaf /…/ *n* (*pl* **leaves**) 1 (of plant) feuille *f*; **dock/oak/lettuce** ~ feuille de patience / de chêne / de salade; **autumn leaves** feuilles d'automne; IDM ▶ **to come into** ~ se couvrir de feuilles; **to shake like a** ~ trembler comme une feuille; CPD **leaf bud** *n* bourgeon *m* à feuilles; […] ▶ **-leafed, -leaved**: **red-~** à feuilles rouges; **broad-~** à grandes feuilles.

2 (of paper) feuille *f*; (of book) page *f*, feuillet *m* spec; IDM ▶ **to take a ~ out of sb's book** s'inspirer de quelqu'un; **to turn over a new** ~ tourner la page. PHRV ▶ **leaf through**: ~ **through [sth]** feuilleter [*pages, papers, book, magazine*]; parcourir [*introduction*].

3 (of gold, silver) feuille *f*;

4 (of table) (sliding, removable) rallonge *f*; (hinged) abattant *m*.

OHFD-3 (2001)
entry re-arranged
MWEs within senses

Fig 7.37 Different ways of handling MWEs in an entry for *leaf*

[20] The discussion in this section is focused on print dictionaries: in electronic dictionaries MWEs are often given prominent treatment in which they are attached to the headword entry but not within it. How this is realized depends on the software design.

Figure 7.37 shows two versions of an entry for *leaf*, which – besides having several distinct senses – also figures in the four major types of MWE:

- compounds, where the headword may be the first element (*leaf mould*) or the second (*bay leaf*)
- phrasal verbs, of all types (*leaf through* verb + prepositional particle)
- phrasal idioms, e.g. *to be in leaf*, *to turn over a new leaf*
- combining forms, e.g. *three-leafed*, *black-leaved*.

Deciding how to handle these four types of MWE is complex, and different dictionaries do different things here. Five of the most common options are:

- Make each MWE a headword in its own right.
- Make selected types of MWE headwords in their own right.
- Put all the MWEs within the entry, at the very end in separate blocks for each type of MWE.
- Put the MWEs within the entry, within the 'appropriate' sense, in separate MWE-type blocks.
- Put the MWEs within the entry, within the 'appropriate' sense, without differentiating the MWE type.

Compounds are often headwords in current dictionaries of all types, and learners' dictionaries – monolingual and bilingual – usually show phrasal verbs as secondary headwords, or even headwords, since their users (unlike English native speakers) tend to have some understanding of what phrasal verbs are. There is no way of handling these that does not produce some anomalies. The key is to decide what will cause fewest problems for the user. Here are some pros and cons relating to MWEs as headwords.

- If MWEs such as compounds or phrasal verbs are not given headword status, then you have to rely on the user (who may not know much about the dictionary's source language) burrowing through the long entry for (say) *set* in order to find *set piece*, *set square*, *set to*, *set-to*, *set up*, *set-up*, etc.
- If they appear as headwords, however, the phrasal verb *set up* and its related hyphenated noun *set-up* are neatly positioned side by side.
- If compounds are made into headwords but phrasal verbs are not, then twenty entries or so will separate *set up* from *set-up*.

- If compounds and phrasal verbs are both given headword status, other anomalies occur. For instance, the phrasal verb *go off* will come between *goof* and *goofy*, and hundreds of entries (including *good* and its compounds) will separate *go off* from the main verb *go*.

In the *OHFD-3* (2001) entry, each type of MWE is given a separate dedicated section, located in a fixed order (first combining forms, then phrasal idioms and so on) at the end of the entry, and signalled by a solid triangle. Compounds with *leaf* as the first element figure as headwords below the *leaf* entry (without pronunciations) in the normal alphabetical order.

In the re-arranged entry, the MWEs, including compounds with *leaf* as first element, and combining forms such as *-leafed*, are all tucked into the most appropriate LU, but flagged according to MWE type.

➔ When you're deciding how to handle MWEs, it's a good idea to look at a lot of other dictionaries, think about your user profile, then choose the way that best fits the needs of your most vulnerable user.

Exercises

1 Identifying and naming components

Choose a dictionary entry to work on. Photocopy or scan it and mark off all the entry components used in it. Figure 7.38 gives an idea of what is involved.

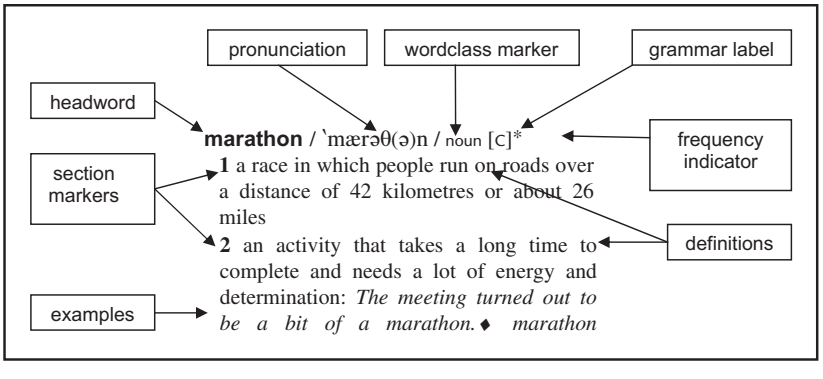


Fig 7.38 Some entry components in the *MED-2* (2007)

2 'Reverse-engineering' a dictionary's Style Guide

This means figuring out the dictionary's main style policies by working backwards from published entries.

- Start by re-reading §4.4, where the Style Guide is introduced.
- Look at the entries for the noun *operator* in Figure 7.39.
- In each entry, identify the various components carrying information about:
  - inflections (cf. §7.2.2.4)
  - grammar (cf. §7.2.6)
  - examples (cf. §7.2.7.2).
- Consider each dictionary's policy for dealing with these three types of information.
- For each dictionary shown in Figure 7.39, draft Style Guide rules which tell the lexicographer how to deal with these three types of information.

operator (ŏp'ə-r-tər) *n.* **1.** One that operates a mechanical device: *a telephone operator*. **2.** The owner or director of a business or industrial concern. **3.** A dealer in stocks or commodities. **4.** A symbol, such as a plus sign, that represents a mathematical operation. **5.** *Informal.* A shrewd and sometimes unscrupulous person who gets what he wants by devious means. **6.** A chromosomal sequence that is the region of an operon responsible for regulation of structural genes.

AHD-2 (1985)

operator /ɒpəreɪtəʳ/ **(operators)**

◆◇◇ N-COUNT

1 An **operator** is a person who connects telephone calls at a telephone exchange or in a place such as an office or hotel. □ *He dialled the operator and put in a call for Rome.*

N-COUNT usu n N

2 An **operator** is a person who is employed to operate or control a machine. □ *...computer operators.*

N-COUNT usu with supp

3 An **operator** is a person or a company that runs a business. (BUSINESS) □ *`Tele-Communications', the nation's largest cable TV operator.*

N-COUNT usu adj N

4 If you call someone a good **operator**, you mean that they are skilful at achieving what they want, often in a slightly dishonest way. (INFORMAL) □ *...one of the shrewdest political operators in the Arab World.*

5 → See also **tour operator.**

COBUILD-5 (2006)

Fig 7.39 The entry for *operator* in two dictionaries

Reading

Recommended reading

Atkins 1993; Landau 2001: 98–152 (*entry components*), and 217–342 (*labels*).
Phraseology: Cowie 1998.
E-dictionaries: Atkins 1996; de Schryver 2003.

Further reading on related topics

Cowie 2001; Hoey 2005; Kilgarriff 1997b; Louw 1993; Mel'čuk 1996; Mel'čuk', Clas, and Polguère 1995; Sinclair 1996; Stubbs 1996, 2001; van der Meer 2000, 2004.

How words work with other words: Benson 1990; Čermak 2006; Coffey 2006; Cowie 1981, 1994, 1999a; Cowie and Howarth 1996; Fontenelle 1992, 1996; Grossmann and Tutin (eds.) 2003; Hanks 2004b; Hanks, Urbschat, and Gehweiler 2006; Hausmann 1989, 1991; Heid 1994, 1998; Kilgarriff 2006b; Lorentzen 1996; Mel'čuk 1988; Moon 1988, 1992, 1996, 1998; Ruppenhofer, Baker, and Fillmore 2002; Siepmann 2005, 2006; van der Meer 1998.

Labels: Fedorova 2004; Norri 1996, 2000; Selva, Verlinde, and Binon 2002; Sharpe 1995.

E-dictionaries: Bogaards and Hannay 2004; de Schryver and Joffe 2004; Duval 1992; Geeraerts 2000; Kay 1983; Rogers and Ahmad 1998; Ruus 2002; Varantola 2002.

Websites

http://www.kuleuven.be/dafles/acces.php : DAFLES (Dictionnaire d'apprentissage du français langue étrangère ou seconde): an innovative online learners' dictionary of French

http://africanlanguages.com/: gives access to a number of online bilingual dictionaries for African languages, including a Kiswahili-English one: (http://africanlang uages.com/swahili/)

This page intentionally left blank

PART II

Analysing the Data

This page intentionally left blank

Introduction to Part II

In the next two chapters, we get down to business. You have your corpus and your corpus-querying software, and you have a good idea about what you need to look for and record. You are now ready to get to grips with the language data, and Chapters 8 and 9 take you through this process in detail. A good way of getting started is to call up a concordance for your headword in one of its wordclasses, taking a sample of (say) 300 or 400 lines. This should give you enough data to develop a working overview of your headword in its various meanings and uses. (If you have access to Word Sketches or similar lexical profiles, these can also form a good starting point for your analysis.)

Chapter 8 focuses on the first 'cut' through the data – the challenging task of identifying the senses of words with more than one meaning – and we suggest a number of practical strategies to help you do this confidently and effectively. Theoretical issues relating to 'word sense disambiguation' are also discussed here, and their practical relevance to the task explained. The outcome of this process is a set of senses, or 'lexical units' (LUs), which then need to be fleshed out. This is the subject of Chapter 9, where we describe the systematic exploration of the corpus data for each of these LUs, classifying the facts to be discovered and suggesting ways of recording these in a relational database, which will serve as a launch-pad for the dictionary proper.

This page intentionally left blank

8

Building the database (1): word senses

8.1 Preliminaries

Ask an English-speaker what *perfect* means, and they will probably give you a simple definition like 'when something is as good as it could possibly be'. But if you ask them what a *party* is, they are likely to say 'well, that depends which sense you mean'. Yet the entry for *party* in *OALD-7* has just four numbered senses (and one multiword expression), while *perfect* has seven. This suggests a disconnect between 'dictionary senses' (the numbered meanings into which many headwords are divided in dictionaries) and 'meanings' as they are perceived by language-users. At the very least, it is clear that the task of assigning meanings to words is different in kind from that of, say, assigning part-of-speech tags. Deciding which wordclass a word belongs to is something humans can do with a high degree of consistency. (And because 'wordclass' is a fairly stable category, machines can be trained to assign POS-tags in a way that reliably replicates what humans do.) But, as we shall see – and as the *party*/*perfect* distinction already suggests – with 'word sense disambiguation' we enter a far more uncertain world. In simple terms, the problem is this:

- Dictionaries typically present words – some words, at any rate – as having several distinct meanings or 'word senses'. It follows that identifying and describing word senses is a major part of what lexicographers are expected to do.
- However, there is little agreement about what word senses are (or even whether they exist). Lexicographers are therefore in the position of having to describe something whose nature is not at all clear.

Not surprisingly, then, it has been observed that 'one of the hardest problems torturing practising lexicographers has always been the question of how to describe the meaning of so-called polysemous words' (van der Meer 2004: 807).

Our objectives in this chapter are to show how dictionary senses can be abstracted from raw language data, and how lexicographers can undertake this task with reasonable confidence. Our goal is to arrive at an inventory of senses for each headword and log them in the database. (As explained in Chapter 4 (§4.2.2), we use 'database' to refer specifically to the structured collection of material assembled during the analysis process, on the basis of which final dictionary entries will be created.) Along the way, we will try to resolve the paradox outlined above, and to explain how the specific requirements of dictionaries can be reconciled with more general truths about how people communicate with one another. Figure 8.1 gives an outline of what this chapter covers.

8.1.1 Why this is important

In every word of extensive use, it was requisite to mark the progress of its meaning, and show by what gradations of intermediate sense it has passed from its primitive to its remote and accidental signification . . . This is . . . not always practicable; kindred senses may be so interwoven, that the perplexity cannot be disentangled, nor any reason be assigned why one should be ranged before the other.

Samuel Johnson, *Preface* (1755)

Most words have only one meaning. A random page in the *Oxford Dictionary of English* throws up numerous single-sense words like *lucrative* (profitable), *Lucullan* (luxurious), *luderick* (a type of fish), *ludic* (playful), *ludo* (a board game), and *lues* (a disease). Most of these words correspond neatly to a specific entity in the real world, and the rest lexicalize unambiguous

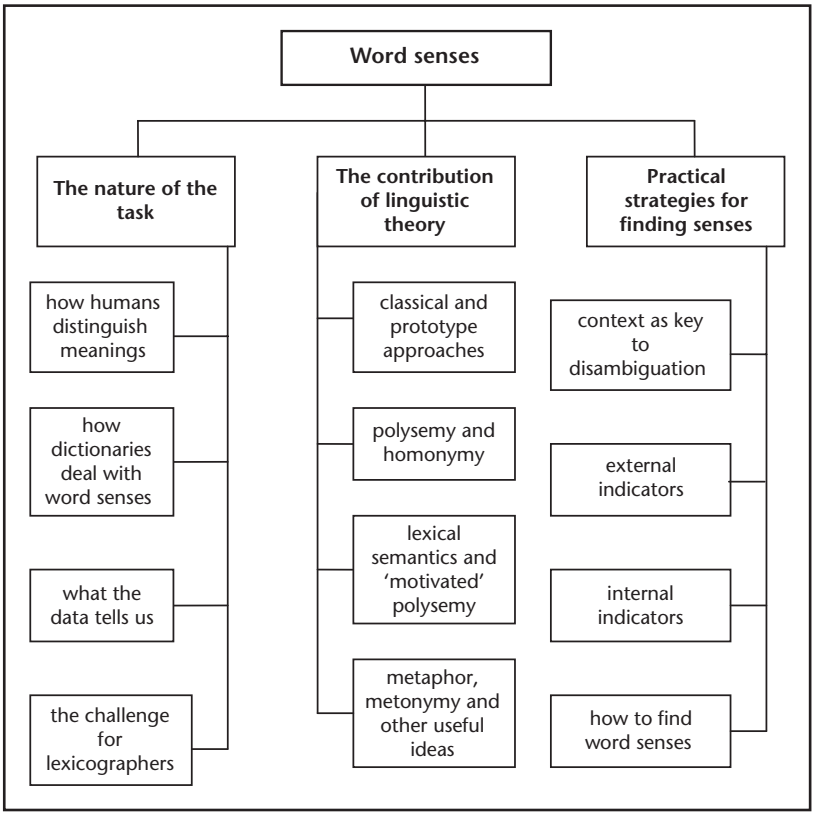


Fig 8.1 Contents of this chapter

concepts. Either way, the meaning is not ambiguous: there is only one possible 'reading' of the word when we encounter it in text. In the average learners' dictionary, around two-thirds of headwords are like this, and the proportion of single-sense words is even higher in dictionaries with larger headword lists. This is because of the strong correlation between frequency and polysemy: the additional headwords in larger dictionaries tend to be infrequent items, and, as a general rule, the rarer a word is, the less likely it is to be semantically or syntactically complex.

But the converse is also true. The more common a word is, the more likely it is to have multiple meanings.[1] And it is precisely these high-frequency words, with their multiple meanings and uses, that make up the bulk of

[1] Moon (1987b: 176) notes that Johnson's entry for *take* (the 54th most frequent English word) lists 134 senses, while the corresponding entry in the first edition of the *OED* has no fewer than 341.

most texts (cf. §3.4.1.1). They form the 'core' vocabulary of the language, so it follows that a major part of the lexicographer's job involves describing these frequent and polysemous items. This in turn means identifying word senses, which is the first stage in the process of building a dictionary entry. Against this background, it is obvious that we need a clear understanding of the issues underlying the division of words into senses, and a set of strategies for performing this task successfully.

8.1.2 Two kinds of polysemy: *party* and *overwhelm*

For the average language-user, there is an implicit assumption that word meanings are fixed entities, and that one dictionary's account of them will be much the same as any other's. A glance at entries for *party* in several dictionaries provides support for this idea. The word is described in identical terms in *LDOCE-4*, *OALD-7*, *MED-2*, and *COBUILD-5*. Each of these dictionaries identifies five lexical units (LUs) – four main senses and one multiword expression:

(1) a political organization (*the Republican Party*)
(2) a social event (*a party to celebrate the end of the semester*)
(3) a group of people involved in a shared activity (*a party of climbers/tourists*)
(4) one of the individuals or organizations in a legal agreement or dispute (*changes can be made only with the agreement of both parties*)
(5) the multiword *be a party to* (*I won't be a party to anything dishonest*).

Larger dictionaries (including *ODE-2*, *MWC-11*, and *OHFD-3*) present exactly the same grouping, and typically add another LU:

(6) (dated or humorous) a person (*an amusing old party*).

There is a high level of convergence among these accounts – but such unanimity is surprisingly rare. A more common scenario is for different dictionaries to divide up the various uses of a polysemous word in different ways, as in the two entries in Figure 8.2. Both dictionaries are aimed at the advanced learner, and both are founded on extensive corpus data – yet they handle *overwhelm* in radically different ways. And each of the other dictionaries in this class gives a somewhat different description. Even so, it is not difficult to find corpus lines that are not adequately accounted for by any of these dictionaries. For example:

overwhelm **1** if someone is overwhelmed by an emotion, they feel it so strongly that they cannot think clearly **2** if work or a problem overwhelms someone or something, it is too much or too difficult to deal with **3** to surprise someone very much, so that they do not know how to react **4** to defeat an army completely **5** if water overwhelms an area of land, it covers it completely and suddenly *LDOCE-4* (2003)	**overwhelm 1** If you **are overwhelmed by** a feeling or event, it affects you very strongly and you do not know how to deal with it. **2** If a group of people **overwhelm** a place or another group, they gain complete control or victory over them. *COBUILD-5* (2006)

Fig 8.2 Senses for *overwhelm* from two dictionaries

Our screens will continue to be overwhelmed by imported products and our national audio-visual industries will suffer.

Indeed, fog can overwhelm the city on as many as 15 days in a winter month.

Within the nucleus of a stable atom, the interproton repulsion is overwhelmed by the strong nuclear force that binds the protons firmly together with neutrons.

They also predicted abortion would overwhelm all other issues in a series of gubernatorial elections this autumn.

In addition to numerous instances like this (which don't obviously match any of the numbered senses in the two dictionaries), we find cases like the following, which seem to straddle two 'different' LUs:

Overwhelmed by the number of donors pushing their desire to lend, recipient governments were frequently unable to sort out their own priorities.

On their last day they were overwhelmed by farewell messages and gifts.

Were the recipients given so much that they didn't know how to deal with it (corresponding to sense 2 in *LDOCE-4*)? Or were they emotionally touched by the donors' generosity (corresponding to sense 1)? The answer is: probably both. To the writer and reader this isn't a problem because writers and readers (and speakers and listeners) don't think in terms of dictionary senses. But lexicographers are obliged to think in these terms.

8.1.3 Lumping and splitting

It will already be clear that the meanings of a word like *overwhelm* can be described at various levels of granularity. We could – as *COBUILD* does here – try to account for the word's uses with just two or three broadbrush

descriptors. But equally, in a more fine-grained analysis, we might identify as many as six or seven senses, each matching a precise context. These two approaches are what lexicographers call 'lumping' and 'splitting'. How far a given dictionary leans towards one approach or the other will depend on the type of dictionary it is, and on the needs and skills of its users (cf. §2.3.1, §2.3.2). But at this point we are building the database. (As noted above, §4.2.2, we use 'database' specifically to denote the structured data collected during the analysis stage of lexicography, which forms the raw materials for eventual dictionary entries.) And at the database stage, it is a good idea to split meanings fairly finely. This is because – when we get to the synthesis stage, where each entry acquires its final form – it is easier to lump related LUs together than to attempt the process in reverse and try to split a coverall sense into smaller units. With a finely split database, publishers have the fullest range of options at their disposal, and can derive a number of dictionaries (larger and smaller, monolingual and bilingual) from a single framework. And in the case of bilingual dictionaries, a database of this type helps to ensure that as many translation solutions as possible are tested out by the translator.

The task of creating final, publishable entries is described in Chapters 10 and 12. In the present chapter, we will describe the process of identifying LUs for recording in a database. Details of the LUs are discussed in Chapter 9.

8.1.4 What this chapter covers

Party and *overwhelm* are at opposite ends of a spectrum. At one end, there is almost complete agreement among dictionaries about 'what the senses of *party* are'. In the case of *overwhelm*, ten different dictionaries will give ten different accounts of its meanings. Between these extremes, communicative events in the real world are mapped onto dictionary senses with varying levels of consensus. This chapter explains why, and suggests how we should respond to this as lexicographers. In the course of this, we will:

- describe the nature of the task facing lexicographers (§8.2)
- discuss relevant theoretical issues, and see what they can contribute (§8.3)
- relate the task to corpus analysis (§8.4)
- outline practical strategies for identifying dictionary senses (§8.5)
- draw some conclusions (§8.6).

8.2 Finding word senses: the nature of the task

In creating a database from which a dictionary (or set of dictionaries) will be derived, our first task is to analyse word forms into distinct meanings, or LUs. This process is often referred to as 'word sense disambiguation', or 'WSD'. WSD is a concept associated particularly with the natural language processing (NLP) community, for whom automation of this process is a major research goal.[2] But the term is equally serviceable in a lexicographic context. However, as we have already seen, 'the trouble with word sense disambiguation is word senses' (Kilgarriff 2006a: 29): there is very little agreement about what word senses are or how broad their scope should be, and no definitive way of knowing when one sense ends and another begins.[3] In this section we outline the nature of the task and the challenge we face as lexicographers, as a prelude to reviewing the contribution of linguistic theory to the business of finding senses.

8.2.1 How many linguists does it take to find a bank?

For decades, linguists, computer scientists, and philosophers have pondered the supposed ambiguity of the noun *bank*.[4] Because of this word's polysemy, it is suggested that an utterance like 'I'm just going to the bank' is capable of two quite different readings: either 'I'm going to the place that handles my money', or 'I'm going to the side of the river'. This is self-evidently ridiculous. In any normal human interaction, there is no ambiguity at all, and if people stopped one another at every turn, to check which of several possible meanings they intended, communication would grind to a halt.[5] An utterance like 'Her heart was beating fast' may – in statistical terms – be open to thousands of possible readings (since each of the words in this sentence has several dictionary senses), but the likelihood of a listener recovering the 'wrong' (unintended) meaning must be close to zero. A far more plausible scenario is one like this:

[2] Automatic WSD is a prerequisite for effective information retrieval, machine translation, content analysis, speech processing, and many other applications.

[3] cf. van der Meer 2006: 604: 'After centuries of practical lexicography, there is still hardly any consensus on how to divide the semantic space of a lexical item'.

[4] See Cruse 2004: 106 for a recent discussion.

[5] Similar points are made by Stock (1984: 134) and Hanks (2000b: 206).

She got up and went into the kitchen. 'Want a drink?' she called. 'No thanks', I said, 'but could you bring me a glass of water?'

Margaret Atwood, *The Edible Woman* (1969)

This interaction may confuse an automatic WSD system (how can a glass of water *not* be 'a drink'?), but it poses no problems for the two participants (or for the reader of the novel). Occasionally, speakers do specify which of several possible meanings they intend. For example:

She was, in the best sense, an old-fashioned family doctor.
After all, at any one time in man's history, there are far more 'ordinary' people (and I do not mean that in a derogatory sense) than those who hit the headlines.
Its moral claims were, in the most literal sense of the word, conservative, in that it enjoined ancient truths and established values.

But this is rare. Despite the pervasiveness of polysemy and the massive potential for ambiguity, humans almost never have a problem with it. Which prompts us to ask: how do human language-users effortlessly perform a task which computers find so difficult? If we can understand how people do this when they communicate, we will be better placed to figure out how to approach the task as lexicographers. The following citations for the polysemous word *icon* provide a clue:

Sotheby's is to auction an icon used in 16th century Russia to assist women in childbirth.
Diana was easily the most influential fashion icon of the 20th century, exhibiting flair and a dash of daring.
When you use the mouse to drag an icon to a new position, the sonic feedback continues.

Here again, there is minimal risk of the writers' meanings being misinterpreted. Anyone reading the full texts from which these extracts are taken has plenty of contextual information to guide them; the third sentence, for example, comes from a newspaper column for computer users, so the reader is primed to expect the 'software' sense of *icon*. But even when these sentences stand alone – like the examples in a dictionary or the short extracts in a concordance line – there is more than enough information to enable us to select the 'right' sense. So, for example, the context in the first sentence tells us that *icon* refers to an object (which rules out the 'person' reading, as in sentence 2); that it is probably an antique of some kind (it is

500 years old, and being auctioned); and that it comes from Russia – all of which leaves no doubt as to the intended meaning.[6]

Context, in other words, is the key, and this will be explored in more detail later (§8.4, §8.5). First, though, we will have another look at how dictionaries handle words with more than one sense.

8.2.2 What dictionaries do

Dictionaries generally divide polysemous words into numbered senses. A conventional dictionary entry consists of 'a list of neatly separated, consecutively numbered lexical meanings' (Geeraerts 1990: 198). This practice dates back at least as far as Johnson, as Figure 8.3 shows.

To RESOU'ND. *v.a.*

1. To echo; to sound back; to celebrate by sound.
   The sweet singer of Israel with his psaltery loudly *resounded* the innumerable benefits of the Almighty Creator. *Peacham.*
   The sound of hymns, wherewith they throne
   Incompass'd shall *resound* thee ever blest. *Milton*
2. To sound; to tell so as to be heard far.
   The man, for wisdom's various arts renown'd,
   Long exercis'd in woes, oh muse! *resound. Pope.*
3. To return sound; to sound with any noise.
   To answer and *resound* far other song. *Milton.*

Fig 8.3 The entry for *resound* in Johnson's *Dictionary* (1755)

It is possible, indeed, that Johnson 'invented' the idea of numbering senses.[7] His immediate predecessor Bailey, at any rate, provides a separate entry for each main meaning (prefiguring the way the *Cambridge Advanced Learners' Dictionary* handles polysemy), while the dictionary produced by the Académie française deals with each separate sense in a new paragraph, without using numbers. Figure 8.4 provides a contemporary example. This convention rests on two (unarticulated) assumptions:

[6] In some cases, successful decoding requires familiarity with a 'sublanguage'. The following sentence includes several very common, highly polysemous words, and could cause problems for an uninitiated reader: 'Flintoff is on strike, and Warne has set an attacking field, with two short legs and a silly point, with a deep third man for back-up'. But for anyone who knows about the sport of cricket, this is all perfectly clear.

[7] 'Sense numbering is seen first in Benjamin Martin's dictionary of 1749 . . . but there is some suggestion that Martin got the idea from Johnson' (whose *Plan* was published in 1747). (Thanks to Rosamund Moon, personal communication.)

keen[1] *adj* **1.** Having a fine, sharp cutting edge or point. **2.** Having or marked by intellectual quickness and acuity. **3.** Acutely sensitive: *a keen ear* **4.** Sharp; vivid; strong: *"His entire body hungered for keen sensation, something exciting"* (Richard Wright). **5.** Intense; piercing: *a keen wind.* **6**. Pungent; acrid: *A keen smell of skunk was left behind.* **7.a.** Ardent; enthusiastic: *a keen chess player.* **b.** Eagerly desirous: *keen on going to Europe in the spring* **8.** *Slang* Great; splendid; fine: *What a keen day!*

Fig 8.4 The entry for *keen* (adjective) in *AHD-4* (2000)

- first, that there is a sort of Platonic inventory of senses 'out there' (so if the dictionary says word W has N senses, it can't possibly have N − 1 or N + 2 senses)
- second, that each sense is mutually exclusive and has clear boundaries (so if a specific occurrence of *keen* is assigned to sense 5, it cannot also belong to sense 6).

The weakness of these assumptions – already clear from the entries for *overwhelm* (Figure 8.2) – is revealed once more in this entry for *keen*. The example in sense 6 – *a keen smell of skunk* – could arguably be explained by the definition 'intense' (sense 5) or by 'sharp; vivid; strong' (sense 4); and it wouldn't be difficult to come up with a single coverall definition which accounted for all three of these senses. Yet this approach to meaning is so firmly established that users come to their dictionaries with the expectation that 'this is what dictionaries do'. The question we must now ask is: how well does this lexicographic convention square with our evidence for how people actually use words when they communicate with one another?

8.2.3 What the linguistic data tells us

The reality turns out to be less clear-cut than the picture presented in dictionaries. Corpus data allows us to observe large numbers of real communicative events. These events convey meanings – speakers and writers don't think in terms of dictionary 'senses' – and as Cruse has observed, 'a lexical unit may be justifiably said to have a different sense in every distinct context in which it occurs' (Cruse 1986: 53). But dictionaries generalize (that is their job), and from the infinite number of individual situations

in which a word appears, lexicographers derive a finite set of LUs which collectively explain how that word contributes to the meaning of all of the individual events.

These LUs are idealized descriptions which instantiate a 'one-to-many' relationship (where one dictionary *sense* matches many language *events*). Sometimes they work very well. In the case of *party*, a simple division into five or six LUs accounts very adequately for at least 99 per cent of the instances you will find in a corpus. But things are rarely this convenient. As we analyse words into LUs, we find many different types of relationship between meanings, for instance:

- where different meanings of the same word form are completely unrelated: compare *punch* (a hit with the fist) and *punch* (a type of drink)
- where a word expresses a single idea, but can be used in different wordclasses: compare *laugh* (verb) and *laugh* (noun)
- where meanings are related (through metaphor, for example) and the relationship is transparent: compare *haunt* (in *her ghost haunts the ruined castle*) and *haunt* (in *journalists who haunt the bars around Westminster*)
- where meanings are (historically) related, but the relationship is no longer apparent: compare *broadcast* (in *a farmer broadcasting seeds*) and *broadcast* (in *a radio station broadcasting the news*)
- where meanings are related through some form of systematic mechanism: compare *she's very friendly* (describing a person) and *she gave a friendly wave/smile* (describing what a person does)
- where one meaning is a specialized application of a more general idea: compare *a white dress with a black belt* (a piece of clothing) and *a black belt in karate* (a belt indicating the wearer's attainment of a particular level in a martial art)
- or where, as Johnson says, 'the shades of meaning . . . pass imperceptibly into each other; so that . . . it is impossible to mark the point of contact' (1755: 5). We saw this in the case of *overwhelm*, where there appears to be 'a single meaning or semantic core underlying [the] various uses'.[8]

[8] Moon 1987b: 174. Moon refers to words like this as 'monosemous' or 'quasi-monosemous'.

These different configurations (and the list here is by no means exhaustive) reflect some of the processes by which a single word form can accumulate multiple meanings. These processes will be discussed in more detail in the next section. For now, though, it must be clear that neither of the assumptions underlying the 'numbered senses' model (that senses are discrete, and that senses are mutually exclusive) can be sustained in the face of the linguistic evidence. And in response to these complicated realities, we find that different dictionaries divide up a word's semantic space in widely differing ways. To some extent, of course, such differences are a function of the dictionary's purpose and target user-group: it is hardly surprising that an unabridged, two-volume historical dictionary like *MW-3* has 119 senses and subsenses of the verb *break*, while Michael West's small proto-learners' dictionary (the 350-page *New Method English Dictionary* published in 1935) manages with just five. But even dictionaries of the same type and size, and with the same target user-group, will often present divergent descriptions of the same word. These disparities reflect the inherent difficulty of the task and the particular 'take' that different lexicographers have arrived at.

8.2.4 Squaring the circle: the challenge for lexicographers

The disjunction between lexicographic practice and linguistic reality has often been commented on. Apresjan, for example, points out that: 'Dictionaries greatly exaggerate the measure of discreteness of meanings, and are inclined to set clear-cut borders where a closer examination . . . reveals only a vague intermediate area of overlapping meanings' (Apresjan 1973: 9). But this shouldn't be taken to imply that lexicographers have failed to grasp how meaning really works. Two centuries before corpora existed, Johnson had already identified the inherent problem with 'dictionary senses' (see above), and his twenty-first-century counterpart concurs: 'The numbered lists of definitions found in dictionaries have helped to create a false picture of what really happens when language is used' (Hanks 2000b: 205).[9] Lexicographers have no problem recognizing that conventional dictionary senses are far from ideal as a device for accounting for meanings. And there may (as

[9] Similarly, Stock (1984: 137): Sometimes, 'meanings blur into each other . . . yet the lexicographer must, given the existing methods of presenting dictionary information, make some sort of job of sorting them out into different . . . numbered meanings'.

we shall see in Chapter 10) be things we can do to improve our presentation of meanings, so that the account we give in dictionaries corresponds more closely to what people do when they use language to communicate. Nevertheless, current dictionary models require that we treat polysemous words in certain ways, and in these circumstances it is important that we be clear about the limitations of the task and that we set realistic goals for ourselves.

There can be nothing 'definitive' about the way we divide words into LUs; indeed James Murray – in one of his more downbeat observations – reckoned that the best any lexicographer could hope for would be that readers would feel, on scanning a multisense dictionary entry, that 'this is not an unreasonable way of exhibiting the facts' (Murray, quoted in Moon 1987a: 86). Against this background, a reasonable (and achievable) objective is that all the members of an editorial team understand the issues and apply the same WSD strategies in a consistent way.

When people communicate, their lexical choices are intuitive, but rule-governed. Similarly, when lexicographers distinguish one sense from another, the process is in the first instance an intuitive one, but if we understand the 'rules' that govern these lexical choices, we will be better placed to do the job of identifying senses. Before moving on to the practical business of dividing words into LUs, we will take a brief look at the theoretical issues which bear on this task.

8.3 The contribution of linguistic theory

Meaning is a big issue. Questions about what words mean and how they acquire multiple meanings have been pondered by linguists and philosophers for centuries. It would be impossible (and probably pointless) to review all the arguments, so what we will do in this section is:

- briefly explain those theoretical ideas that seem especially relevant to the job lexicographers have to do
- identify any promising ways in which these ideas can inform the practical task of WSD.

For lexicographers, the value of this theoretical background is that it reveals the complexity inherent in the notion of polysemy. Words don't just have

'different meanings': there are many *kinds* of different meaning, and if we understand the various mechanisms by which these meanings develop, we will be in a better position to tackle the practical task of finding dictionary senses.

8.3.1 Classical approaches and prototype theory

In classical semantic theory, a discrete meaning is one that embodies a cluster of 'criterial features' – the particular conjunction of 'necessary and sufficient conditions' which uniquely identify that meaning. This approach to the analysis of meaning goes back at least as far as Aristotle's *Metaphysics*. A favourite example is the 'basic' meaning of the polysemous word *bachelor* (discussed in Fillmore 1975), whose criterial features are:

- an adult male (neither women nor young boys can be called bachelors)
- someone who has never married (a widowed or divorced man is not a bachelor).

According to this view (which also underpins traditional approaches to defining: see §10.5.1) anyone who satisfies both conditions is a bachelor, and anyone who doesn't is not: it is a straightforward binary choice. In reality, however, there are many words and meanings which don't conveniently conform to this kind of analysis. Even *bachelor* itself has rough edges. For example:

- A man may be in a long-term relationship and may have children with his partner, but without actually being married: is he a bachelor? (No.)
- A man who has divorced or left a long-term relationship may say something like 'now I'm a bachelor again'. But is he? (Perhaps.)
- If a gay man is not in a long-term relationship could he too consider himself a bachelor? (Why not?)

Changes in social norms have altered the way we look at this word. Intuition suggests that it is used far less often than it once was, and that contemporary uses are more likely to be ironic than unmarked. Thus even the word that is most often invoked to explain classical theory doesn't fit the model especially well. Once you admit the existence of borderline cases, classical theory begins to break down. And there are, as Cruse observes, 'many everyday

words whose meanings cannot be captured by means of a set of necessary and sufficient features' (2004: 128).[10]

As Aitchison points out, words and meanings are not so much 'precision instruments' as 'slippery customers', whose exact boundaries can rarely be drawn with any confidence (Aitchison 2003: 41f.). Wittgenstein famously pondered the word *game*, whose varied instantiations in the real world – for example, chess, football, Grand Theft Auto, hide-and-seek, Monopoly, Ragnarok Online – preclude the possibility of a simple checklist of defining characteristics. 'The category is not structured in terms of shared criterial features, but rather by a criss-crossing network of similarities' (Taylor 1995: 38). Membership of the category *game* is thus explained in terms of 'family resemblances'. When speakers understand and internalize what *game* means, they do this not by learning a list of 'defining' features, but by extrapolating the meaning from all the exemplars they encounter.[11] (Similarly, consider the way small children learn the scope of the concept *dog*, whose exemplars exhibit enormous variation.)

All of this leads us to the notion of a 'prototype' theory of meaning, which is most closely associated with the cognitive scientist Eleanor Rosch. Rosch's experiments (see Box 8.1) suggest that:

- Speakers develop an idea of what represents an ideal exemplar of a category (a 'prototype').
- These prototypes – though amenable to some variation – are to a large extent shared by members of a speech community.
- A prototype functions as a 'cognitive reference point', with other entities seen as belonging to the same category provided they are sufficiently similar to the prototypical member.
- This in turn implies that there can be 'degrees of category membership', so that some members are 'better' exemplars than others.

[10] For useful introductions to classical theory and its limitations, see Taylor 1995, chapter 2; and Cruse 2004, chapter 7 (esp. 127–132).

[11] Anna Wierzbicka is a lone dissenting voice. She rejects the idea that *game* is indefinable in terms of essential components, and proposes a definition which, she believes, applies to anything we would call a game (Wierzbicka 1990: 357–358). This is fair enough as an exercise in theoretical semantics – but the definition she supplies has seven criterial features, and would not be appropriate for a mainstream dictionary aimed at human users.

Box 8.1 Eleanor Rosch and 'prototype theory'

In a number of empirical studies, Rosch developed her idea of prototypes. An early experiment investigated the way that a preliterate people (the Dani of New Guinea) learned a number of concepts (relating to colours and geometrical shapes) which did not exist in their own culture and were therefore not lexicalized in their language (Rosch 1973). She found that when subjects were taught arbitrary names for colours, for example, the names they learned most successfully were those that described primary or 'focal' colours – prompting the notion that not all members of the category 'colour' have equal status. In later experiments with US college students, Rosch pursued this idea of 'degrees of category membership'. Subjects were asked to judge how well a given item represented a particular category: thus for example, *chair* emerged as a strong (or prototypical) member of the category FURNITURE, *bookcase* a somewhat less obvious exemplar, with *mirror* at best a borderline member. Similar tests were carried out with categories such as fruit, clothing, tools, and birds. In the latter case, *robins* and *sparrows* were seen as 'prototypical' birds; *pheasants* and *ducks* were much less 'birdy'; and *ostriches* were barely thought of as birds at all (Rosch 1975). The linguistic (and lexicographic) implications of Rosch's research are that the members of a given category (for lexicographic purposes, think 'word sense') are not all equally typical examples of that category. This tallies with what we find when we analyse words in text: we often see a gradience from highly typical instances to borderline cases – those examples which are difficult to assign to a particular dictionary sense with any confidence.

For useful summaries of Rosch's work, see Lakoff 1987: 39–46; Taylor 1995: 42–46; Aitchison 2003, chapter 5.

This looks an altogether more promising avenue. Thinking of words, and senses of words, in terms of prototypes to which individual language events approximate – sometimes closely, sometimes more loosely – is an approach which is more easily reconciled with the messy linguistic realities described above (see §8.2.3). And the implications for what lexicographers do, both when identifying senses and writing definitions, are profound.

The relevance of all this to the task of finding senses should be clear. With an understanding of prototype theory, and of the inherent (and pervasive) fuzziness of word meaning, we are in a better position to take on the task of identifying and describing dictionary senses. We will approach the job in a pragmatic frame of mind, appreciating the limitations of WSD and

recognizing that there are no absolute truths awaiting discovery. We may also feel that conventional ways of presenting word senses in dictionaries need to be modified to reflect the way the language works. In a detailed analysis of the verb *climb*, Hanks (1994) shows that, while some uses in text are absolutely prototypical (as when a human subject climbs a tree or mountain), many others invoke aspects of this prototype to varying degrees. For example:

- A car may climb up a steep hill.
- A plane may climb into the air after take-off.
- A column of smoke may also climb into the air.
- A plant may climb up a wall.
- A road or path may climb up the side of a hill.
- . . . and so on.

The entry for *climb* in *ODE-2* (2003) aims to reflect these findings in the way it is structured. The entry opens with a broad definition of the prototypical 'ascend' meaning, and this is followed by a series of 'subsenses' describing recurrent uses which are both more specialized and less prototypical (like the ones listed above). The beauty of this approach is that it explains all the typical, frequently-occurring uses of *climb*. It makes no claim to account for all possible textual instantiations of the word, but it is loose enough to accommodate considerable variation at the level of individual language events. We still find the occasional corpus instance whose meaning doesn't precisely map onto any of the subsenses in the *ODE* entry. None of the following, for example, quite fits the description:

> *It would be totally irresponsible to risk losing a club glider by attempting to climb a large cloud.*
> *As the main parties ponder the fateful nature of the developing contest in the capital, London and its doings will climb steadily up their agendas.*
> *Dujon tried to take evasive action from a ball that climbed from just short of a length.*

But it would be wrong to say that the dictionary 'fails' to capture these meanings precisely: rather, it deliberately makes no attempt to do so. It aims instead to provide users with a useful explanation both of the prototypical uses of *climb*, and of the most common variations on this prototype. Armed with this knowledge, no user will have a problem with occasional individual uses that invoke the prototype in slightly different ways.

No doubt there is more to be done in the area of presentation,[12] but a prototype approach to WSD has two major advantages over the classical model:

- It reflects the way people create meanings when they communicate, and thus it goes with the grain of the language, and accommodates creativity and fuzziness.
- It makes the lexicographer's task more manageable, because it allows us to focus on the prototype and its common exploitations, rather than requiring us to predict and account for every possible instantiation of a meaning.

8.3.2 Polysemy and homonymy

Dictionaries have traditionally distinguished polysemous words from homonyms. The word-form *punch* – when functioning as a noun – illustrates the difference. Consider the following instances:

(1) *She gave him a punch in the stomach.* (a hard blow with the fist)
(2) *It lacks the emotional punch of French cinema.* (a forceful, memorable quality)
(3) *Glasses of punch were passed around.* (an alcoholic drink mixed from several ingredients)

The meaning expressed in the second sentence can be interpreted as a metaphorical extension of the physical punch in sentence 1. Sentence 3, on the other hand, has nothing in common with the other two: the meaning it expresses lies in an entirely different semantic area, and the sole point of contact is the (accidental) fact of sharing the same orthographic form. This third use of *punch* has an altogether different history: it entered the language later and from a different source (from the Sanskrit word 'panch', meaning 'five': the drink originally had five ingredients.) So in dictionary terms, *punch*[1] is a polysemous word with two meanings (corresponding to sentences 1 and 2), while the unrelated *punch*[2] (the drink) is a homonym.

[12] See for example van der Meer's (2000) comments on the model used in *ODE*; and the diagrammatic, nonlinear ways of representing the senses of the Dutch word *vers* (roughly equivalent to English 'fresh') proposed by Geeraerts (1990: 202–207), which allow for overlapping and clustering of related meanings.

Distinguishing the different exemplars of the same word-form is quite complex, and dictionaries handle the task in a number of ways.[13] The relevant variables are:

- Homonymy: do we separate *punch*-drink from *punch*-blow with fist?
- Wordclass: do we separate *punch*-noun ('a punch in the stomach') from *punch*-verb ('punched me in the stomach')?
- Homophony: do we separate *bow* (the weapon used with arrows) from *bow* (an act of bending the body forward)? The two words look identical but sound different.
- Capitalization: do we separate *swede* (a vegetable) from *Swede* (a person from Sweden)?

In most dictionaries, homophones and capitalized forms are dealt with as separate entries, or 'homographs'. This means there are different entries for *bow* (bəʊ) and *bow* (baʊ), and for *may* and *May*. Differences based on wordclass are variously handled, the usual options being:

- a single entry with subsections for each wordclass: thus *ODE-2*, *OALD-7*, and *OHFD-2* have a single headword *journey*
- separate homographs for each wordclass: thus *MED-2*, *LDOCE-4*, and *MWC-11* have two separate entries for *journey*-noun and *journey*-verb.

Homonymy, however, is gradually being abandoned as an organizing principle in many types of dictionary, and in the case of learners' dictionaries the process is already complete. *LDOCE-1* had no fewer than nine entries for the word-form *tip* (four nouns and five verbs), whereas *LDOCE-4* has just two, bringing all the noun uses under one homograph and all the verb uses under another. The *COBUILD* dictionaries go further still, providing a single entry for each orthographic form (regardless of capitalization and phonological differences): thus *punch*, *bow*, and *swede*/*Swede* each have just one entry.

Is homonymy still a relevant concept for lexicography? The answer, as always, depends on the intended uses (and target users) of the dictionary. In historical dictionaries, homonymous words always appear as separate entries: describing words' origins and development is central to the function of dictionaries of this type. But the value of homonymy to a synchronic account of meaning is far less clear. For many users, such divisions may

[13] See also the discussion of entry structure above: §7.3.

seem pointless. The connections – or lack of them – among the various uses of a word form will not necessarily be obvious. A user who surmises (wrongly, but not unreasonably) that the drink *punch* is so called because it is strongly alcoholic and 'packs a punch', may be mystified by the fact that this use is given a separate entry (while the diverse meanings of *club* are covered under a single headword).[14] And a rigorous application of homonymy could well cause look-up problems, too. If a dictionary has several noun and several verb homographs for a word like *tip* (and in *MWC-11* there are ten homographs of *tip* in all), how will the user know where to look in order to find subsidiary meanings or multiword expressions like 'on the tip of my tongue' or 'tip someone the wink'?

For all these reasons, homonymy is no longer seen as either helpful or relevant in the way learners' dictionaries organize meanings. As Moon explains (when describing *COBUILD*'s policy): 'Because access to an item is through its orthographic form, and because etymological homonymy depends on knowledge that is not available to the dictionary user before he or she locates the word in the dictionary, it was decided to ignore homonymy completely' (Moon 1987a: 88).

8.3.3 Lexical semantics and 'motivated' polysemy

In his two major books on lexical semantics, D. A. Cruse (1986, 2004) catalogues the varieties of polysemy in exhaustive detail, showing the wide range of ways in which a single word can instantiate several meanings. Though his terminology is occasionally offputting (he identifies phenomena such as 'autoholonymy' and 'meronymic enrichment'), Cruse's work is essential reading for lexicographers who want to understand the mechanisms by which words acquire multiple meanings. The important message is that polysemy comes in many forms, and an appreciation of the various ways in which it arises will help to make the WSD task less daunting. Three concepts have particular value here:

(1) Contextual modulation: consider, for example, the following uses of the word *car*:

> *My car has broken down.*
> *I'm having the car resprayed*
> *What a comfortable car!*

[14] See also Lyons 1977: 552; Rundell 1988: 129.

Our car got a puncture.
I'm going to fill up the car.

In each sentence, the single concept *car* (a private road vehicle for a driver and a small number of passengers) is 'modulated' by the specific context (Cruse 1986: 53; 2004: 118f.). What's happening here is that different features of this concept are foregrounded in different situations. Thus, in the first sentence, the focus is on the car's mechanical parts, in the second on its outer body, and so on. These are not distinct *meanings* of the word: the listener or reader simply settles on 'a reading which is compatible with the context' (Cruse 2004: 118). In Hanks' terms, *car* could be said to have a set of 'meaning potentials' (cf. Hanks 1988), any one of which may be activated in a particular context. It may not in every case be easy to distinguish contextual variation from clearly separate meanings, but it is important to understand the general principle.

(2) Antagonism: one of the characteristics of a genuinely polysemous word (one with two or more distinct senses) is that you cannot simultaneously interpret the word in two or more different ways: the various meanings are said to be *antagonistic*, because 'they compete with one another, and the best one can do is to switch rapidly from one to the other' (Cruse 2004: 106). Thus a sentence like *Here comes the groom* could refer to the man getting married at a wedding or to someone whose job is to take care of horses – but it can't refer to both at the same time. In normal communication, any potential ambiguity would be resolved by the broader context (cf. §8.2.1), but as Cruse points out, 'it is impossible to focus one's attention on both readings at once' (ibid.). Compare this with the *car* examples above: there is no 'antagonism' there, just different perspectives on the same entity. Antagonistic readings are mutually exclusive, and are thus the clearest manifestation of polysemy.[15]

(3) Motivation: as a rule, linguistic behaviour is motivated rather than arbitrary. The underlying 'rules' governing particular phenomena (such as the generation of phrasal verbs) are sometimes hard to retrieve (and in some cases have not yet been identified), but the

[15] Similarly, we saw (§8.1.2) that *party* can refer (among other things) to two types of grouping: a political organization, and a group of tourists, climbers, etc. In a sentence like *She joined the party*, the two 'group' readings are antagonistic, and mutually exclusive.

general principle holds good. So too with word senses: 'There is, by definition, a motivated relationship between polysemous senses' (Cruse 2004: 108). That is, there is always a reason why a word has more than one meaning, and the links between its various senses can usually be explained. Cruse's position is quite a bold one: to the average language-user, the relationship between *club* (society) and *club* (weapon), or between *convention* (conference) and *convention* (custom), may look pretty arbitrary. Nevertheless, in the case of a polysemous word (as opposed to two or more homonyms: §8.3.2), we should always start from the assumption that its various meanings have developed through mechanisms that are both recurrent in the language and fairly well understood. The most productive of these mechanisms are discussed in the next three sections.

8.3.4 Specialization

The word *dog* has a default reading (the domestic pet that barks), but it can also be:

- more general: where it means any animal – wild or domesticated – belonging to the genus *Canis* (such as a wolf or jackal)
- more specialized: where it refers specifically to a male animal (*Is that a dog or a bitch?*).

The second of these processes – 'specialization' – is especially common, and has a number of variations. The various readings of the verb *drink* can be understood in these terms. Consider for example:

She raised the cup to her mouth and drank. (to swallow a liquid)
Jack doesn't drink, so I ordered him a Coke. (to drink alcohol)
She left her first husband because he drank. (to drink alcohol excessively)

Here the use of *drink* becomes progressively narrower (and the term 'narrowing' is sometimes used to describe this process) and thus changes its meaning in significant ways. Something similar happens with the verb *run*:

I had to run to catch the bus.
Sally runs/goes running several times a week.

The concept encoded in the second sentence is not fundamentally different from the 'basic' sense of *run*, but here it implies specific goals, training, and

techniques. (This narrower use is defined in *ODE-2* as 'run as a sport or for exercise', as a subsense of the basic meaning.) Sometimes, specialization is a function of a specialized domain, or sublanguage (cf. §3.4.3.6). The word *walk* is a case in point:

> *Australia's Adam Gilchrist should be applauded for his decision to walk in Tuesday's World Cup semi-final against Sri Lanka.*

The sublanguage here is the sport of cricket: a batsman who believes he has been fairly dismissed and walks off the pitch without waiting for a decision from the umpire is said to 'walk'. (This is regarded as very sporting.) Walking (in the basic sense) is indeed involved here, just as running (in the basic sense) is involved in running for exercise. But this is a distinct (if fairly obscure) LU.

A number of other processes could be seen as forms of specialization. Amelioration and pejoration – where an originally neutral word acquires either a positive or negative 'spin' – are both common. Thus *notorious*, at one time merely synonymous with 'well-known', has undergone pejoration and now combines the notions of 'famous' and 'wicked'. *Crusade* denotes (neutrally) one of the medieval military campaigns to the Holy Land, and, by metaphorical extension, a 'campaign' against something deemed to be bad (such as drug barons, child abuse, or corruption). This is a case of amelioration (metaphorical crusades are seen as noble enterprises), but recent events have put the word off-limits in certain contexts: George W. Bush had to be warned against referring to 'a crusade against terrorism' because, for his putative targets, this conjured up unfortunate associations with the word's original reference to medieval wars waged by Christians against Muslims. In hyperbole, similarly, we can observe specialization from a neutral to an emotionally charged sense, as in *It's boiling in here*, *I've been waiting for ages*, or *Those shoes cost a fortune*.

Deirdre Wilson (forthcoming) explains processes like this in terms of her concept of 'relevance'. The reader or listener naturally seeks to make sense of an utterance, and interprets it accordingly. Thus if someone says *The baby has a temperature*, the search for relevance leads inevitably to the (specialized) reading 'an undesirably high temperature'; the neutral reading – the degree of heat or coldness of something – would make no sense in this context. For Wilson, specialization (or narrowing), and other processes by which the 'literal' meaning of a word can be modified in context, are seen as

'outcomes of a single pragmatic process which fine-tunes the interpretation of virtually every word' in terms of relevance.

8.3.5 Regular polysemy

Another well-documented form of motivation is 'regular polysemy'. Regular polysemy – discussed in more detail in Chapter 5 (§5.2.4) – describes the phenomenon where all the members of a particular semantic set behave in predictably similar ways. A classic example is that words for trees typically function both as countable nouns (referring to the tree itself) and as mass nouns (referring to its wood), thus:

> *Is that an oak or a beech?* [countable]
> *The doors are made of solid oak.* [mass]

The important lexicographic point is that the name of *any* tree which is used as a source of wood is capable of this alternation. This systematicity presents risks and opportunities for the lexicographer: the risk of failing to treat similar items in consistently similar ways, and the opportunity of streamlining the editorial process. Both can be addressed through style policies and through template entries which provide editors with a standard entry format to be applied to every member of a class (cf. §4.5).

Regular polysemy was first described by Apresjan (1973), who sees it as one aspect of 'the search for systematicity in the lexicon' (Apresjan 2002: 91). Apresjan's initial work identified instances of regular polysemy in Russian. But the phenomenon has been shown to operate in many other languages, and there is a substantial inventory of regular polysemy classes in English. These include 'verb alternations', where all the verbs in a particular semantic set exhibit predictable variations in syntactic behaviour. And adjectives that are applied to people for describing how they feel – a very large class of words – also regularly describe people's actions and responses. Thus a Word Sketch (§4.3.1.5) for *angry* includes the following in its list of nouns which the adjective modifies:

- people: *crowd*, *mob*, *man*, *demonstrators*, *fans*, *residents*
- actions and responses: *voice*, *response*, *retort*, *face*, *outburst*, *letter*.

It makes obvious sense that any adjective belonging to this broad semantic set should be handled in the same way in a dictionary (though issues of

relative frequency need to be considered too, as we shall see later – §8.6.3). This is just one area where a systematic approach to the lexicography – reflecting systems in the language – can bring efficiency gains *and* help us produce better dictionaries.

8.3.6 Figurative extensions: metaphor and metonymy

We saw earlier (§8.3.1) how the verb *climb* has a range of meanings. They radiate outwards from a central prototype (e.g. *children climbing trees*) and, in the process, different aspects of the prototype (or different 'meaning potentials') are activated in specific contexts. Thus for example a car can climb a hill (which involves motion) but so can a path or road (the path itself doesn't move, but could be seen as starting at the bottom and proceeding to the top). Many of these uses are sufficiently stable and frequent to be regarded as distinct 'dictionary senses', while others – though still comprehensible by analogy with the prototype – are too marginal and idiosyncratic to qualify for sensehood. At these outer edges, the boundaries can never be quite clear, and a word's lexicographic treatment will depend as much on the size and scope of the dictionary as on the weight of corpus evidence.

Similar processes are at work throughout the lexicon. In a high proportion of cases, the journey from original meaning to 'new' senses involves *figurative* uses of language, where there is an implied resemblance between primary and secondary meanings. The most obvious manifestation of this is where a literal or physical meaning (*they climbed the mountain*) underpins a non-literal or abstract one (*the stock market climbed to a new high*). The names of body parts and organs (*head*, *mouth*, *heart*, *skin*, and so on) exemplify this process especially well: they spawn numerous figurative meanings which variously invoke similarities – real or supposed – with the form, location, or function of the word in its primary sense.

Figurative extensions come in many forms, and some of the more common of these will be discussed here:

(1) Figurative uses in general: *haunt*, *sacrifice*, and *broadcast*. Consider the following concordance lines:

> *Her ghostly presence now haunts the ruined castle.*
> *. . . memories of past sins and traumas which then return to haunt them*
> *Modigliani had been haunted by the fear that he would be trapped in the south*
> *Americans had been rather thin on the ground lately, fearing such terrorists as*

might haunt the dark streets round Notting Hill Gate
As night fell I looked for information elsewhere. I haunted bars and lowish dives . . .

The default reading of *haunt* is that it is what ghosts do when they visit a place. This sense incorporates a number of features or images:

- recurrence (the ghost keeps coming back, haunting isn't something that happens just once)
- a sense of anxiety or menace (the ghost's appearance is unsettling to those who witness it)
- a gloomy or ethereal quality (ghosts don't generally appear in broad daylight).

It is easy to see how the other meanings develop from here, through the activation of some (but not all) of the characteristic features of the core sense. In this case, the original 'ghost' image is powerful enough to pervade the extended meanings. When speakers and writers use *haunt* to talk about guilty or anxious feelings, or about people in dimly-lit bars, a residual 'ghostly' feel is preserved and transmitted to these new meanings.

The word *sacrifice* also exemplifies a literal-to-figurative progression, but in slightly different ways. In the case of *haunt*, most of the features of the core meaning are activated in various ways in the word's other senses. But with many other words, 'only some prototypical features of the literal meaning are transferred in the metaphorical process, while others are suppressed' (Knowles and Moon 2006: 21):

The goat had been sacrificed at the shrine and the meat shared out among the villagers.
In the early days at least, being your own boss means sacrificing your social life.
. . . stiff dialogue and generic plotting which sacrifices all complexity or realism to simplistic human interest
The safety/cost conflict is all too easily settled by sacrificing safety to the god of profit.

In this case, the progression depends not so much on the evocation of images of religious sacrifice as on a shared sense of the *intention* of the 'sacrificer': when you make a sacrifice, you voluntarily give up something of value in order to gain something you value even more highly. The second line exemplifies this idea in a straightforward way: the 'sacrifice' of one's social

life is seen as tough but worthwhile. In many other cases, though (as the last two lines illustrate), there is an additional implication that the sacrifice was wasted, because you gave up far more than you gained. In contexts like these, *sacrifice* demonstrates a common phenomenon, whereby – in the process of figurative extension – a word can acquire semantic features that were not necessarily present in the original meaning.[16]

Finally, the case of *broadcast*, which began life as an adjective, referring to seeds 'scattered abroad over the whole surface, instead of being sown in drills or rows' (*OED*). It started being used as a verb ('to scatter seeds' etc.) in the early nineteenth century, then acquired a figurative use to match that of its Latinate equivalent *disseminate*. In the twenty-first century, however, its most common use by far – whether as a verb or noun – relates to the television and radio industries.[17] The literal-to-figurative process by which *broadcast* acquired these newer meanings is not essentially different from what we saw with *haunt* and *sacrifice*. The difference here is that the original (agricultural) meaning has disappeared from mainstream discourse. By contrast with *haunt* and *sacrifice*, the literal meaning of *broadcast* does not contribute any 'atmosphere' to the word's contemporary uses. So although metaphorical processes are at work, this is a 'dead metaphor', and *broadcast* would not be recognized by the average speaker as an instance of figurative language.

(2) Metaphorical sets: in some cases, a whole set of specialized vocabulary may be generated through metaphor. Computer terminology is a good example. Though some clearly technical terms do exist in this field (such as *hexadecimal*, *graphical user interface*, and *baud rate*), most of the vocabulary familiar to non-specialists recycles existing words. Examples include *mouse*, *crash*, *hibernate*, *architecture*, *bookmark*, *unzip*, and *cookie*. A closer look reveals subsets of computer words which all share a particular metaphorical image, for example:

- the computer is thought of as a person: *virus*, *memory*, *infect*, *client*, *compatible*, *refresh*
- the computer screen is thought of as an office: *clipboard*, *bin*, *desktop*, *wallpaper*, *notepad*, *folder* (not to mention Microsoft's *Office* software).

[16] The word *lottery* illustrates a similar progression: in its literal and figurative uses, the common theme is that the outcome depends on chance. But the figurative meaning carries additional connotations of unfairness and arbitrariness.

[17] This meaning first appeared in 1922.

As these subsets show, a metaphorical schema – once initiated – will often influence the way a whole collection of related notions is lexicalized.[18]

(3) Conceptual metaphor: George Lakoff and Mark Johnson's seminal *Metaphors We Live By* (1980) illustrates the pervasiveness of metaphor in language. The authors see metaphor not so much as a 'decorative' feature of literary or rhetorical registers but as a fundamental cognitive process that shapes the way we form concepts and give them names. This is especially relevant to any discussion of polysemy, because it helps to explain why so many 'new' senses are effectively metaphorical extensions of older ones. Thus, basic spatial concepts (like up/down, in/out) give rise to 'orientational' metaphors. The idea of being *up*, for example, equates with status and power (we talk about people being 'at the *top*', '*climbing* the career ladder', or '*falling* from power'); good health ('at the *peak* of fitness', '*came down* with flu') and good spirits ('on *cloud nine*', '*down* in the dumps'); and moral rectitude ('the moral *high ground*', 'an *upstanding* member of the community'). Similarly, mental processes are typically conceptualized in terms of physical ones: the idea of understanding something, for example, is interpreted by analogy with the physical processes of vision (*I see what you mean*) and of taking hold of something (*I couldn't quite grasp her point*).

In addition to these 'big' metaphors, Lakoff and Johnson identify numerous 'conceptual metaphors' that underlie the way we think about (and lexicalize) a wide range of common activities. A typical example is the conceptual metaphor ILLNESS IS WAR – the idea that illnesses 'attack' the body and must be 'resisted'. This metaphor is reflected in expressions like these:

> *. . . a substance in the blood that aids the body's* ***defences*** *when antibodies combine with* ***invading*** *antigens*
> *People were also worried that electro-magnetic fields could* ***destroy*** *the body's* ***resistance*** *to infection.*
> *The conventional treatment for large tumours is to* ***bombard*** *them with powerful doses of gamma radiation.*
> *Your body can cope with a cold,* ***fight off*** *a serious illness and with time, even mend a broken bone.*
> *This may in part explain the* ***aggressive*** *nature of these cancers.*
> *. . . substances which can be introduced into the human body to* ***combat*** *pathogenic microorganisms*

[18] See also Meyer et al. 1998.

To Lakoff and Johnson (1980: 2), 'most of our ordinary conceptual system is metaphorical in nature', because our thought processes are governed by (indeed, constrained by) the physical realities of our bodies and the world we occupy. Insights like these won't necessarily make the practical task of WSD any easier. But by helping us to perceive underlying systems in the language, they leave us better equipped to make sense of language data.

(4) Metonymy: this is another important mechanism by which word meanings develop. Consider the following:

> *Kuwait was invaded by Iraq in 1990.*
> *Blair's reputation was profoundly damaged by Iraq.*

Each mention of a country here invokes a different kind of meaning. 'Kuwait' refers straightforwardly to the country of that name. The first use of 'Iraq' does not, but there is a simple association between the country and its leaders and armed forces. The second use of 'Iraq' is more complex. The reference here is not to the country, its government, or its army. Rather, it encapsulates a bundle of associations which include the case made for the Anglo-American invasion of Iraq (and the taint of mendacity attaching to it), various aspects of the war's conduct, and the developing outcomes for Iraq and the wider region.

Both references to Iraq are examples of metonymy, a form of figurative language 'involving either part-whole relations . . . or else naming by association' (Knowles and Moon 2006: 47). Cruse (2004: 210f.) provides a large, though not exhaustive, inventory of recurrent forms of metonymy. In most cases, one of two basic mechanisms is at work:

- a focus on one attribute or aspect of something when your intention is to refer to the thing as a whole
- a reference to something as a whole, when your intention is to focus on one of its attributes or aspects.

Examples of the first type ('attribute or aspect for whole') include:

> *There were just too many **mouths** to feed.* (= people, but with the focus on their need for food)
> *He wanted everyone to admire his new **wheels**.* (= his new car)
> *To celebrate, we ordered a couple of bottles of **fizz**.* (= champagne: 'fizziness' is one of its salient features)
> *Do you take **plastic**?* (= credit cards, which are made of plastic)

Examples of the second type ('whole for attribute or aspect ') include:

> *The **kettle's** boiling.* (= the water in the kettle)
> *The **bank** has defended its policy of applying monthly service charges.* (=the representatives of the bank as an institution)
> *Would you mind if I used your **bathroom**?* (= the toilet in your bathroom)
> *The **school** was in mourning for the murdered students.* (= the staff and students of the school)
> *If United win today, they'll gain a place in **Europe**.* (= specifically, in the European champions' league)

Many of these metonymic processes recur so frequently that they can be seen as forms of regular polysemy (cf. §8.3.5, §5.2.4). The interesting lexicographic question is how far a dictionary entry can or should account for such alternations in meaning. As usual, there is no simple answer, but in each case our decision will be guided by much the same criteria as we apply to other aspects of the WSD problem, such as:

- Systematicity: does the metonymy instantiate a recurrent, well-established pattern?
- Frequency: is the extended meaning a common usage?
- Longevity: is the extended use ephemeral or likely to endure?

The second reference to 'Iraq' above – though a frequent enough usage in the first decade of the twenty-first century – fails these entry criteria because it is an idiosyncratic form of metonymy (a kind of journalistic shorthand that depends on a great deal of shared knowledge) and is specific to a particular moment in history. Conversely, the use of placenames associated with government (*the Pentagon*, *the Kremlin*, *Capitol Hill*, *Downing Street*, and so on) to refer to government institutions and their representatives (*Downing Street has refused to comment on these allegations*) passes every test, and most dictionaries include extended senses of these and similar names.

A note of warning: appeals to 'consistency' should not lead us to apply these principles unthinkingly. As Hanks shows (2001), it is not that difficult to take a lexicographically valid instance of metonymy (like 'the whole school applauded') and then imagine scenarios in which more marginal cases can be made to look plausible (like 'the whole forest applauded' or 'the whole bike-shed applauded'). In dealing with metonymy, regular polysemy, and indeed polysemy in general, we need to keep in mind that 'part of the

job of the lexicographer is to distinguish canonical members of a lexical set from ad-hoc members'.[19]

8.3.7 Linguistic theory and its relevance: some conclusions

When our lexicographic analysis leads us to identify a distinct LU, we won't always be able to pinpoint the precise mechanism by which it came into being. Cruse (2004: 209) mentions cases like the 'head' of a bed and the 'back' of a chair, which can be explained in several ways. Is it because this is the part of the bed where our head rests, the part of the chair where our back rests (implying a form of metonymy)? Or are metaphoric processes at work ('head' and 'foot' regularly instantiate notions of 'top' and 'bottom')? What is important here – from the point of view of practical lexicography – is not that we can necessarily say 'I know how X came to mean Y'. What matters is that we recognize that polysemy comes in many forms and arises through many mechanisms, and that it is almost always motivated rather than arbitrary.

Theoretical linguists have explained (or attempted to explain) these phenomena from numerous different angles. Some of the more important of these have been discussed here, but there are many others. Lexical Network Theory, Frame Semantics, the Generative Lexicon, and Lexical Priming[20] (*inter alia*) each makes its own special contribution to the debate. But

[19] See also §8.6.3 on misguided appeals to 'consistency'.

[20] On Lexical Network Theory, see e.g. Norvig, P. 'Building a Large Lexicon with Lexical Network Theory', *Proceedings of the First International Language Acquisition Workshop* (Detroit, MI, 1989); essentially, LNT is a formalism that was proposed as a basis for building large computer lexicons. It envisages a network of senses, with each sense related through links that connect different senses of the same word *and* related senses of other words. A largely pre-corpus theory, it helps to explain (often by invoking the big Lakoffian metaphors) why 'new' senses of a word arise, and in what ways they are different from (or similar to) existing senses. On Frame Semantics in general, see §5.4 above; on the contribution of FrameNet to practical WSD, see Atkins, Rundell, and Sato (2003), esp. 334ff.; on the Generative Lexicon, see Pustejovsky, J. *The Generative Lexicon* (Cambridge, MA: MIT Press, 1998), esp. chapter 3: for Pustejovsky, polysemy is a logical and systematic process, whereby 'subsidiary' and novel senses of a word derive from its core meaning through the application of lexical rules to a rich description of the features of the base meaning; on Lexical Priming, see Hoey (2005), esp. chapter 5. Hoey argues that 'the collocations, semantic associations and colligations a word is primed for will systematically differentiate its polysemous senses' (ibid.: 81).

there is also a good deal of overlap and convergence across the range of theoretical approaches. Learning about these ideas won't necessarily make the process of identifying word senses any easier, but you will tackle the job with greater confidence if you understand the underlying systems, and you will be better equipped to make good judgments in the more marginal, less clear-cut cases.

8.4 Word senses and corpus patterns: context disambiguates

Knowing about theoretical perspectives on WSD helps us understand how and why some words acquire multiple meanings. But finding word senses for a dictionary is, in the end, a practical activity for which we need a practical methodology. The availability of abundant corpus data enables us to characterize the WSD task in quite simple terms. As Rosamund Moon puts it so succinctly: 'Context disambiguates' (Moon 1987a: 87). The lexical and syntactic environment in which a word appears turns out to be the most reliable indicator of the meaning it conveys in any particular instance (when several readings are theoretically possible). This link between meaning and form was one of the earliest 'discoveries' made when linguists and lexicographers began analysing corpus data in a systematic way. Alongside the principle that 'every distinct sense of a word is associated with a distinction in form' (Sinclair, quoted in Moon 1987a: 89), it was also observed that the *amount* of context you need in order to perform this task is – in most cases – surprisingly small. As early as 1984, Stock was struck by the high proportion of short concordance lines that could unambiguously be associated with a particular meaning. This led her to wonder 'what it was in the minimal contexts available that enabled lexicographers to decide without apparent trouble which meaning of a polysemous word was being used' (1984: 132). She gives two examples of the verb *operate*:

> *Human beings will simply be unable to operate them.*
> *They operated but it was too late.*

With corpus-querying software, we always have the option of retrieving more of the original context – but as Stock points out (ibid.: 134), we rarely need to do this. Even in the two minimally contextualized cases above it is clear to any competent user of English that the first example refers to running something like a machine or system, and the second to carrying

out a surgical procedure. Which once more prompts the question: how do we do this? The case of *issue* illustrates the process well, because the word has – fairly recently – acquired a 'new' use, as these examples show:

> *Typically half of the 90 children in Gardner-Betts have mental health issues.*
> *He wasn't hateful, but I could tell he had some issues.*
> *. . . a supplement to support bladder control and health in dogs especially those with incontinence issues*
> *Issues around sexuality can be deeply threatening for young people.*
> *Due to his emotionally chaotic upbringing, he likely does have significant intimacy issues.*
> *It helps to understand . . . how issues around gender, dependency, daily routine, and staff responsibility impinge on the environment.*

It isn't always obvious how this use of *issue* diverges from its familiar sense of 'an important topic or problem for debate or discussion' (*ODE-2*, 2003), and there are plenty of borderline cases. But corpus data supports the view that a *specialization* of the basic meaning has become sufficiently frequent to be regarded as a distinct LU. This is covered in *ODE-2* as a subsense of the first main meaning:

> n (**issues**) informal personal problems or difficulties:
> *emotions and intimacy issues that were largely dealt with through alcohol.*

What are the factors that underlie this 'split' in the basic sense of *issue*? Our first observation relates to the *domain* in which this usage tends to appear: it is used mainly in the broad area of social science (in fields such as psychology, counselling, criminology, and childcare). Apart from this, the clues are in the context:

- It is always (in this use) pluralized, as the *ODE* entry acknowledges.
- It is rarely sentence-initial or clause-initial, but usually occurs in the patterns *have* + *issues* or *with* + *issues*.
- It is often followed by *around* (this use of *around* to mean 'concerning' is itself quite recent, and characteristic of the same discourse types as *issues* in this meaning).
- It is often premodified by another noun (as in *intimacy issues* or *incontinence issues*).

Taken together, this cluster of features provides a clear description of what makes this use of *issue* distinct. This in essence is how practical WSD

proceeds, through a combination of subjective judgments and objective observations.

The next section outlines a set of practical strategies which – taken together – represent a reliable *modus operandi* for identifying LUs for a dictionary.

8.5 Practical strategies for successful WSD

When we say that 'context disambiguates', we use 'context' in its broadest sense, referring both to the *kinds of text* in which a given meaning typically occurs (what we might call 'external' indicators) and to the immediate *lexico-grammatical environment* typically associated with a particular meaning (the 'internal' indicators). These are the features that enable us to resolve potential ambiguities in real communication, and the same features, collectively, provide a methodology for WSD.[21]

8.5.1 External indicators: domain, dialect, and setting

One of the basic requirements of a lexicographic corpus is *diversity*. From the Brown Corpus onwards, corpus-builders have recognized that – since words behave differently in different settings – it is desirable to include texts from the widest possible range of sources. (See §3.4.2.1 for a fuller discussion.) With a well-balanced corpus, we have access to meanings that appear frequently in one type of text but may be rare or non-existent in others. In such cases, the setting alone is the strongest indicator of intended meaning – though there will usually be other evidence to support our analysis. This section deals with some of the external (that is, non-linguistic) features of texts which have a bearing on the way that the writer or speaker uses a polysemous word. Understanding these features will assist us in associating meanings with uses.

8.5.1.1 *Domain* In many cases, the *domain* (or subject matter) of a text determines the meanings that certain polysemous words have in that text. The word *bond* has several possible uses, but for the reader of any of the three extracts below, there is no risk of ambiguity:

[21] Stock 1984 and Moon 1987a cover roughly similar ground.

Once the substrate is bound to the surface of the enzyme, covalent bonds may be formed or broken.
Japan's leading brokerages agreed to stop issuing new shares and convertible bonds for at least a month.
A dominant form of family life began to take shape . . . characterized by close emotional bonds between parents and children.

Each sentence comes from a distinct subject-field: respectively, chemistry, financial journalism, and social science. The texts are full of lexical clues as well, notably the domain-specific modifiers (*covalent*, *convertible*, *emotional*) and other items such as *substrate* and *enzyme*. But essentially, the fact that we are reading a text about chemistry primes us to expect the 'chemistry meaning' of *bond*, and the same applies to the other sentences.[22]

8.5.1.2 *Regional dialect* Another external indicator is *regional dialect*. When American speakers say they are going to *wash up* they mean they are going to wash their hands and face, whereas the same verb in British English refers to the activity of washing plates and cutlery after a meal. Similarly, a South African speaker of English would not be surprised to pass several *robots* while driving to work: *robot* is the usual South African word for what most English speakers call a *traffic light*. Within each speech community, the intended meaning will always be clear, though the risk of misunderstanding may arise in communication between speakers of different dialects. (This – unlike the spurious *bank* scenario mentioned above – is one of the few situations in which there is genuine potential for a polysemous item to be misconstrued.)

8.5.1.3 *Time* We have seen some of the mechanisms by which words can acquire new meanings in the course of time (§8.3). Sometimes, a new meaning completely replaces an older one, and in such cases *time* is the best indicator of intended meaning. When Sir John Middleton informs the Dashwood girls that the only gentleman staying at his house is 'neither very young nor very gay' (Jane Austen, *Sense and Sensibility*, chapter 7), the modern reader knows that Sir John is not commenting on this gentleman's sexual orientation: *gay* did not acquire its 'homosexual' meaning until well

[22] Similarly: the expression *on strike* – which usually refers to industrial stoppages – has a distinct meaning in the context of cricket (roughly equivalent to 'at bat' in baseball). Anyone reading a text on cricket would construe the expression in these terms unless there were compelling evidence for the default interpretation.

into the second half of the twentieth century – but once this meaning arose, it only briefly co-existed with the older use before completely obliterating it. Intelligent readers (including lexicographers) interpret meaning in terms of the norms prevailing at the time when a text is produced (or in the period in which it is set). In less clear-cut cases, our reliance on context will be greater. Hanks (1998) notes a significant shift, since the eighteenth century, in the default meanings of *enthusiasm* and *condescension*, which have undergone, respectively, amelioration and pejoration over the last 200 years. But even if we don't know about these changes, the context leaves little doubt as to the intended meaning: a eulogy for an 18th-century Bishop of Exeter describes him as a 'successful Exposer of Pretence and Enthusiasm' and applauds his 'winning conversation and condescending Deportment'. In a dictionary providing a diachronic description of the language, these older uses have to be accounted for: thus in *ODE*, the entry for *enthusiasm* includes a sense (labelled 'archaic, derogatory') which is defined as 'religious fervour'.

8.5.1.4 *Subcultures* Any speech community – especially one as large as 'English' – will include numerous subcultures. By subculture, we mean any group whose characteristic preoccupations or lifestyles differentiate it from mainstream culture (if there is such a thing). Markers of distinctness include race and ethnicity, age, interests, and ways of earning a living, and subcultures include such groups as criminals and prison inmates, computer geeks, or drug-users. In any cohort of this type, members will be under pressure to conform to approved norms – and these norms include distinctive ways of using language. Teenagers and young adults constitute a classic (if heterogeneous) subculture, with its own distinctive linguistic behaviour. Young people's slang changes rapidly, partly in order to maintain its exclusiveness; the much-cited positive meaning of *wicked* (equivalent to 'great' or 'fantastic') was spurned by all right-thinking teenagers as soon as adults noticed it and – worse – started using it themselves. A more recent instance of 'youthspeak' is the idiosyncratic use of the adjective *random*. In the period 2001–2006, the frequency of *random* increased year on year in text-types dominated by younger people (blogs, message boards, chatrooms, and so on),[23] while its meaning shifted away from the usual sense of 'haphazard' or 'unsystematic', as the following sentences show:

[23] Thanks to Adam Kilgarriff (personal communication) for this information.

That was the worst ending ever [to a film]*: hard-working bloke gets nothing and some random bloke comes in and gets the girl.*
Rock records are becoming ever more soulless . . . a criticism that we'll happily hurl at the new Oasis album (popping up on a random web site near you).
The first is her best: a deceptively spontaneous book about a pair of hotshots whose random affairs keep them running back and forth across the country.

This is not so much a case of specialization as a leaching of semantic content, so that – although the word retains echoes of its original meaning – it has become more like a general marker of vagueness.

How do we deal with this as lexicographers? As always, our (subjective) intuitions and knowledge of the world interact with the (objective) linguistic data. In a well-balanced and well-annotated corpus, it will be clear that uses like these tend to cluster in certain well-defined text-types, and this constitutes evidence for a 'sense' that is specific to one subculture. Our strategy here mirrors what readers and listeners do when they encounter such uses in real life: they (and we) take account of the context of use and (following Wilson's analysis, §8.3.4 above) the search for 'relevance' leads us to an appropriate interpretation of the intended meaning.

8.5.2 Internal indicators

In the last section, we looked at some of the *external* characteristics of texts that influence – and sometimes determine – the meaning assigned to a polysemous word by a writer or speaker (and the meaning inferred by a reader or listener). For lexicographers, text-type features such as domain, time, and regional dialect often provide valuable evidence to support the process of identifying dictionary senses. But many (probably most) of the different uses of multisense words are unmarked in terms of features like these. In order to account for meanings that are found across a range of text-types, we have to rely on *internal* evidence – specifically on evidence of a word's . . .

- syntactic and lexico-grammatical behaviour
- collocational features and selectional restrictions
- colligational preferences.

A single piece of evidence will sometimes be conclusive. Quite often, though, we need to look at a number of indicators. When a particular meaning is instantiated in corpus data, the various instances of that

meaning typically cluster together on the basis of *several* shared features – a unique permutation of semantic and contextual clues.

8.5.2.1 *Syntax and lexicogrammar* The word *friendly* illustrates the way a distinct meaning is often signalled by a change in grammar. Compare the following:

(A) *The enquirer was a friendly, bubbly girl of about twenty-three.*
One more advantage of living in Taiwan is that people here are polite and friendly towards foreigners.
She met me with a friendly smile, shook my hand and introduced me to the class.
The Samos is an established hotel which is well run and has a friendly, relaxed atmosphere.

(B) *It was known, also, that the old lady had been friendly with his mother.*
I had become friendly with Vivien Fish, whose father was Churchill's dentist.
He said he was still friendly with the princess but would do nothing to embarrass her.

In set (A), the word is used about people (describing their manner or personality), and – through a form of regular polysemy (§8.3.5) – it is often also applied to people's behaviour or gestures (*a friendly smile, wave, manner*, etc.) or to situations where people behave in a friendly way (*a friendly atmosphere, place, hotel*). The examples in set (B) are quite different: they refer not to behaviour or personal attributes, but to a relationship of friendship. They indicate a *state*, and the context typically includes words like *be* and *become* (*How long have you been friendly with her?*). But the key indicator here – the thing that decisively marks this out as a separate sense – is the preposition *with*: if you are 'friendly with' someone, you are their friend.

As this example shows, a significant shift in meaning may be encoded in an apparently trivial change in grammar. The examples that follow show similar processes at work in the areas of countability, transitivity, and syntactic patterns.

(1) Countability: in many cases, countable/uncountable alternations reflect types of regular polysemy, such as the 'generic'/'item of' alternation:

I can't do my job without email or *There were 28 emails in my inbox*
an increase in gun-related crime or *Police recorded 243 gun-related crimes*

Sometimes, though, the change in grammar signals a completely different meaning:

She received the sort of welcome normally reserved for royalty (= members of a royal family)
The author's royalties will all go to charity (= income received from a book)

(2) Transitivity: when a verb is polysemous, transitivity is often an indicator of meaning. Some senses require an object, some never have one, while in others the object is optional. Compare:

No drink at all should be the rule, particularly when driving, ***operating*** *machinery or taking certain medicines* (object obligatory)
Surgeons were forced to ***operate*** *after an infection set in* (no object).

In its 'surgical' use, *operate* is always intransitive, and is used either on its own or in a PP with *on*.

Verbs that allow 'object-deletion' don't usually change their meaning through this process: *We went out to celebrate* is no different from *We went out to celebrate her birthday*. But sometimes this change has implications for meaning. Compare:

I borrowed $5 from him
Some firms had to borrow in order to stay in business.

The verb *borrow* normally takes an object, and the object specifies the thing being borrowed (money, someone's car, a book, etc.). But when the object is omitted – as in the second example here – the only possible reading is a specialization of the general sense of *borrow*:

- the thing being borrowed is money (it can't be anything else)
- the lender is a financial institution (not, say, a friend of the borrower)
- the borrower has undertaken a programme of repayments.

(3) Syntactic patterns: the verb *remember* has a number of uses, and each is typically associated with a particular syntactic pattern. Figure 8.5 gives some examples. The list is not exhaustive, and it should be noted that some of the verb's meanings can also be instantiated through a simple V + O structure. But *remember* illustrates well the way that syntactic behaviour often provides evidence for shifts in meaning.

8.5.2.2 *Collocation and selectional restrictions* Not all polysemous words are as syntactically complex as *friendly*, *operate*, or *remember*. Many verbs,

Example	Pattern	Meaning
I remember sitting alone in the cafeteria, slowly drinking my cup of coffee	+ *-ing* form	call to mind a past experience or event
Then she remembered why Nahum had come home in a terrible temper	+ *wh*-clause	call to mind information you knew before
But just remember that you are being judged even before you speak a word	+ *that*-clause	often imperative: keep in mind a relevant fact
He only hoped Jane had remembered to leave the window open	+ *to*-INFIN	do something you undertook to do
Feynman, who died in 1988, is remembered for his many contributions to theoretical physics	+ PP/*for*	usually passive: be known or celebrated for a particular achievement
Please remember me to your mother. I trust that she is well.	+ PP/*to*	usually imperative: convey greetings to

Fig 8.5 Syntactic patterns used with the verb *remember*

for example, only ever appear in the pattern V + O, and in cases like this we have to rely on other forms of 'context' to identify different meanings. In this section we will look at the role of *selectional restrictions* and *collocation* as markers of senses.

Both terms refer to an observable tendency of certain words to occur frequently with certain other words. When we talk about 'selectional restrictions', we mean the general semantic category of items that typically appear as the subjects or objects of a verb, or as the complements of an adjective. A collocation, on the other hand, is a recurrent combination of words, where *one specific lexical item* (the 'node') has an observable tendency to occur with another (the 'collocate'), with a frequency far greater than chance.

To illustrate: the verb *forge* originally referred to the process of making or shaping metal objects through the application of heat (*a blacksmith forging a sword*), and in this use it is sometimes followed by a PP with *from* or *into* (*forging metal into armour*, *forging armour from metal*). This use is still occasionally found, but most of the time *forge* is used in a simple V + O pattern, so we need other kinds of evidence to distinguish its meanings. Lexical-profiling software is an efficient way of gathering the information we need. A Word Sketch for *forge* lists the following words as typical objects:

alliance, banknotes, bonds, friendship, links, metalwork, painting, partnership, passport, relationship, signature, sword, ties, unity

It is immediately clear that (if we leave aside 'sword' and 'metalwork') these objects belong to two broad semantic groupings: 'relationships' and documents' (or anything else that can be copied with criminal intent).

These categories make up the 'selectional restrictions' of *forge*, and they point to two distinct meanings: to create an enduring relationship, and to fraudulently make a copy of something.

With selectional restrictions, once you know the category, any word belonging to that category can fill the relevant slot. Collocation is a less open-ended – and more arbitrary – phenomenon. We talk about people *committing* crimes, but in this case we can't – without sacrificing naturalness – substitute other verbs from the same general category (words like *make*, *do*, *perform*, *carry out*, or *execute*). 'Commit a crime' is a collocation. Fluent speakers recognize this intuitively, and statistical software confirms that the probability of *commit* occurring with *crime* is massively greater than for any other verb with this general meaning. (Of course, some of the items that instantiate a word's selectional restrictions could also be regarded as collocates of the word if they appear very frequently in this combination.) Taken together, these two types of 'words that often co-occur' provide valuable evidence for meaning differences.[24] Most adjectives are light on syntax and can be used both attributively (*an excellent performance*) and predicatively (*her performance was excellent*). For words like this, collocation and selectional restrictions are vital indicators of meaning. The adjective *fresh* is a good example. It is occasionally followed by the preposition *from*, and this use instantiates one of the word's several meanings:

fresh from their victory/triumph/win/success . . .

. . . and so on. For the most part, though, syntax doesn't help us to disambiguate the uses of this adjective. But the word's typical complements point to a number of clear selectional restrictions, including these two sets:

(1) *fruit, vegetables, herbs, fish, peas, tomatoes, parsley*
(2) *perspective, thinking, insight, approach, look*

Set (1) illustrates the word's basic meaning, while set (2) points to a quite distinct use ('original and novel'). Collocation plays a part, too. When combined with words from set (1), *fresh* is occasionally modified by adverbs such as 'deliciously', 'wonderfully', and 'lipsmackingly'. Combinations involving words from set (2), however, have a much stronger tendency to be modified, and in this case the adverb collocates – a quite different set – include 'genuinely', 'completely', 'entirely', 'strikingly', and

[24] For a fuller discussion of collocation, see §9.2.7.

'remarkably'. Each piece of evidence contributes to our analysis and enables us to distinguish dictionary senses with greater confidence.

In many cases, corpus data simply confirms our intuitions and gives them objective support. Sometimes, however, the corpus reveals important distinctions which introspection alone is unlikely to provide. The phrasal verb *build up* illustrates the point nicely. Like the similar verb *increase*, *build up* can be used with or without an object: you can *build something up* (or *increase it*), or something can *build up* (or *increase*). But *build up* isn't as straightforward as it looks, as the two concordance extracts in Figures 8.6 and 8.7 show.

1	thirties, has spent the last five years	building up	a successful business. However, she recen
2	and Lee Grant also scored as Colts	built up	a 6-0 lead. Sam Reed had set them a diffi
3	her own mark and has been steadily	building up	a quality client list. Spence, who made the
4	tive enterprise; over time, researchers	build up	a body of wisdom which tells them which
5	eel for the language. This in turn will	build up	their confidence in English, which they
6	n exposed to the infectious agent and	built up	an immunity to it. It is only in remote

Fig 8.6 Concordances for the transitive use of *build up*

1	aotic conditions. Massive bottlenecks	built up	in the early spring on the railway network
2	cerned that inflationary pressures are	building up	in the region's economy, foreshadowing
3	t followed incident, tension had been	building up	to a new peak. Raymond was accused of
4	the kidneys stop working and poisons	build up	in the blood. The patient will experience
5	we realise that a situation could be	building up	which could lead to the ultimate defeat of
6	ash in 1987. Huge backlogs of work	built up	in the security dealers' back offices, and

Fig 8.7 Concordances for the intransitive use of *build up*

It turns out – and this is the kind of insight which has only become possible through the availability of large corpora – that things which *build up* (intransitive) are overwhelmingly negative and undesirable, whereas things that people *build up* (transitive) are almost always positive. So with two very different semantic classes operating as selectional restrictions, the evidence leads us to conclude that *build up* has two quite different meanings in its transitive and intransitive uses.

8.5.2.3 *Colligational preferences* Michael Hoey, refining a concept introduced by J. R. Firth and applied (implicitly or explicitly) by linguists such as Halliday and Sinclair, sees colligation as 'a midway relation between grammar and collocation' (Hoey 2005: 43). Specifically, for Hoey, it describes the tendency that some words have to favour (or avoid) particular forms or positions. For example: the 'average' countable noun can be either singular or

plural, and can appear in any position in a clause or sentence. So if we find that a particular noun is almost always plural, and is never sentence-initial, then we have a *prima facie* case of colligation – an observable preference for a subset of the available grammatical options. 'Preference' is the key word here: we are not talking about hard-and-fast rules but about quantifiable *norms*. Large corpora enable us to establish norms, so that we can say, for instance, 'verbs that take an object are, on average, passivized in *n* per cent of instances'. If we then observe an individual verb occurring in the passive far more frequently than this, we need to apply the Sinclairian principle mentioned above (§8.4: 'every distinct sense of a word is associated with a distinction in form') and ask ourselves whether this preference has implications for meaning. And very often it does. In its broadest application, colligation can include any of the following:

- in verbs, a marked preference for one particular form or use, such as the imperative, passive, reflexive, or progressive (*-ing* form)
- in nouns, a marked preference for either the singular or plural form, or for modifying other nouns
- in adjectives, a marked preference for either attributive or predicative position, or for comparative or superlative forms
- in any wordclass, a marked preference for one particular position within the sentence or clause
- in any wordclass, a marked preference for appearing in negative (or 'broad negative') constructions: think of words like *compunction*, *remotely*, *afford*, *tenable*, or *budge*
- a strong tendency to *avoid* any of these forms, structures, or positions.

Three of the uses of *remember* we discussed above (§8.5.2.1 and Figure 8.5) exhibit marked colligational preferences (the last, for example, strongly favours the imperative), and this provides additional evidence to take account of when we analyse the verb's meanings. A few further examples will show the relevance of colligation to the task of identifying LUs.

(1) Adjectives: many adjectives can be used both as 'classifiers' (or 'pertainyms') and as descriptive words. Compare:

> *Like the precise astronomical observations of the Maya, these technical achievements proved to be a dead end.*
> *In this country young drivers are paying astronomical fees for insurance coverage.*
> *The price of local property was said to be 'absolutely astronomical'.*

When *astronomical* is a classifier ('related to astronomy': *astronomical observations*), it has an overwhelming preference for attributive position, whereas its descriptive use – referring to a large amount or number – can appear before or after the noun. And while prices can be described as 'absolutely astronomical', classifying adjectives are not generally modified or graded.

(2) Nouns: in its most frequent use, *trial* refers to a legal process presided over by a judge or magistrate. When used in this way, the noun has no obvious preferences for singular or plural form or as subject or object, though it frequently occurs – following a preposition and with no article – in expressions like *on trial*, *committed for trial*, and *detention without trial*. In other meanings, however, there are marked colligational effects:

- in the sense of 'a test carried out before a decision is made', *trial* often functions as a modifier, appearing directly before another noun:

  Once you take out a 60-day trial subscription, you can cancel for any reason
  Avoid any mail order or catalogue offer unless there is a free home trial period or bona fide money-back guarantee.
  Rice and her husband have agreed to a trial separation after work pressure forced their marriage on to the rocks.

- in the sense of 'a sports competition for selecting team members for a major event', the noun is typically pluralized (even when the referent is singular), and almost always pre-modified:

  Cambridge University's Boat Race trials were derailed yesterday when one of the three crews hit a floating sleeper on the Thames.
  Crawford finished 12th in the third round of the World Motor Cycle Trials.

- in the sense of 'a painful or difficult experience', *trial* is typically (though not always) pluralized:

  Adam's wife was expounding on the trials of being the mother of a pre-school-age daughter.
  the happiness we experienced on our wedding day, the early years of our marriage with their trials and uncertainties

(3) Verbs: the two main senses of *acquit* show marked colligational features:

- in its 'legal' use it has a strong preference for the passive
- its other meaning is only invoked when the verb is used reflexively:

The former Scarborough goalkeeper certainly acquitted himself well on his debut.

In its most usual ('applaud') meaning, *cheer* can be either transitive or intransitive. In a less frequent use, meaning 'to encourage', the verb is almost always passive:

I was much cheered by the fact that he expressed unqualified approval for it.

In an interesting case study, Hoey (2005: 82–88) looks at the colligational features of the two meanings of *consequence* ('result' and 'importance'). The 'result' use is at least ten times more frequent than the 'importance' use, and shows a marked preference for being pluralized. In its 'importance' use, *consequence* has a number of clear preferences (this is a brief summary of a detailed study):

- It almost always appears in a PP (*of great/little consequence*).
- It is never the subject of a verb.
- It never occurs with a *specific* deictic (like *these* or *the*), favouring instead words like *some*, *no*, or *any*.
- It has a strong tendency for 'denial', that is for saying that something is *not* important.

In Hoey's terms, these preferences are examples of 'primings': typical features of a word's behaviour which we unconsciously associate with a particular meaning. As Hoey points out, a word doesn't have to conform to all of these primings in order to invoke a meaning, but it will always conform to some of them.

8.5.3 Putting it all together: *argue* again

Along with the other linguistic features discussed in this section, colligation makes an important contribution to the task of identifying senses. The accumulated evidence from all these 'internal indicators' complements our intuitions about meaning and underpins an analysis which is as objective and 'scientific' as it reasonably can be, given the slippery and dynamic nature of word meaning. Before concluding this section, we will look once more (Figure 8.8) at the verb *argue* to show how its four different meanings can be mapped onto four unique permutations of linguistic clues.

In an earlier chapter (§5.5.2.1), we discussed *argue* in the context of 'lexicographic relevance' and the value of a Frame Semantics analysis. Four distinct LUs were identified:

LU-1 'quarrel, dispute' (*don't argue with her*)
LU-2 'maintain, make a case for' (*he argued for a change in tactics*)
LU-3 'indicate, constitute evidence for' (*this argues a lack of support*)
LU-4 'persuade' (*she argued them out of going*).

Now let's do this in reverse: starting from these four meaning areas, we will collect some typical corpus examples for each, and use these to identify those features which, collectively, differentiate one meaning from another.

LU/meaning	corpus examples	linguistic features
LU-1 quarrel	*The teachers and medics were arguing about who has which square inch of my time.* *We spent most of our time in cafes, arguing and holding hands.* *The platoon commander was arguing with a gang of Christian Phalangists.*	– no object, no passive – often in *-ing* form – allows reciprocal use (with two or more subjects) – you argue *with* a person, and *about* (sometimes *over*) an issue
LU-2 maintain, make a case (for)	*Employers in the industry argued strongly for the retention of a statutory levy.* *She argues against the radical feminist view of 'male violence in the hands of the state'.* *He argues the need for a written constitution which is compatible with the rule of law.* *Headland has persuasively argued that there was just not enough food for such groups in the forest itself.* *Of course, it can be argued that readers get the paper that they deserve.* *Originally, France had argued for these plans to be confirmed by popular referendum.*	– usually with a PP (you argue *for* or *against* something) or with a *that*-clause – often modified by an adverb (*persuasively*, *cogently*, *convincingly*, etc.); though some of these collocates are shared with LU-3 – modality is common, in patterns like *It can/could be argued . . .* , *One could argue . . .* – occasionally with a simple noun object – occasionally in the pattern *argue* + *for* + *to*-infinitive – rarely in *-ing* form
LU-3 indicate, constitute evidence for	*The congestion on our roads argues that a serious vehicle tax should be levied.* *These features argue for a local origin.* *This lack is a key factor arguing against the existence of such a relationship.*	– non-human subject – usually with a PP (facts argue *for* or *against* something) or with a *that*-clause – rarely in *-ing* form, typically in simple present
LU-4 persuade	*Don't try to argue him out of it now – it's too late.* *Better not tell her the truth. Better just argue her into going back where she belonged.*	– obligatory PP: you either argue someone *into* something or *out of* it – often in infinitive, after words like *tried to, managed to*

Fig 8.8 Linguistic features of the LUs of *argue*

Notice that LU-2 and LU-3 have quite a lot in common, but the critical difference is that LU-3 (uniquely for this word) requires a non-human subject. As discussed in §5.5.3, text-type information can provide additional support: in the case of *argue*, examples of the first and last LUs are most likely to be found in conversation or fiction, whereas LU-2 and LU-3 are more typical of journalism and academic discourse.

8.6 Conclusions

In this chapter, we have discussed the various mechanisms by which a word can acquire 'new' meanings. Sometimes, the new and old co-exist (like the senses of *party*); sometimes a newer meaning becomes far more common than one it grew out of (as happened with *broadcast*); and sometimes, newer meanings completely replace older ones (as in the case of *nice*, whose original meaning – 'stupid, foolish, senseless', according to the *OED* – has been obsolete for several centuries). We have shown, too, that these processes are motivated rather than arbitrary: we may not fully understand the means by which each individual new sense developed from an older one, but the process is in general systematic and explicable. We have also seen how meanings have considerable elasticity: there are norms (or prototypical uses), and there are exploitations that take these norms as their starting point.

8.6.1 Words and meanings

There is thus plenty of scope for creativity. Following the terrorist attacks on New York City in September 2001, there was a vogue among politicians and commentators for the word *existential*. This doesn't reflect a sudden interest in Kierkegaard or Sartre – rather, the word is used in expressions like this:

> *The paper relates Mr Blair's "visible frustration" that opponents have failed to confront what is described as the "existential threat of global terror".*
> *The Bush administration was confronted by the greatest, existential challenge to its power and authority that any US government has faced since Pearl Harbor.*

The intention, presumably, is to stress that these threats and challenges aren't a matter of conjecture but 'really do exist' and/or that they pose a threat 'to our very existence'. From the speaker's point of view, the

denotation doesn't need to be any more precise than this, but the choice of *existential* – with its philosophical connotations – adds a touch of gravitas.[25] The important point here is that – although nothing in current dictionary entries exactly accounts for these uses of *existential* – it's unlikely that listeners will have any problem decoding the speakers' meaning. As language-users, we encounter things like this on a daily basis, and it is not at all unusual, in any novel, newspaper article, or conversation, to come across words being used in ways which are not explained in dictionaries. A few more examples:

> *We know Andy is always happy to help so we shout 'Andy, can we borrow you pleeeease?' and five minutes or so later he appears.*
> *Of course, Eighteen Visions* [a rock band] *are playing a dangerous game, and when they get things wrong it's all a bit of a car crash.*
> *But the PC model of the world is embedded deeply into Microsoft's corporate DNA.*
> *He'll have a definite case to answer if you can get someone high enough up the food chain to take some notice.*

We don't usually 'borrow' *people*, but the meaning is clear enough. Similarly with the other three examples, which are all figurative extensions. As these cases illustrate, the language system allows us to generate novel meanings (or 'stretch' existing ones), without compromising intelligibility; the cooperative principle in communication will usually ensure that a speaker or writer doesn't exploit a norm to the point of obscurity. In many cases, a novel use comes and goes without leaving much trace. Sometimes (and this may be what happens with the 'new' meaning of *existential*) it enjoys a brief vogue and then disappears. And in other cases, it is taken up and copied by a large enough section of a speech community to qualify as a new meaning. We saw this with the specialized use of *issues* (§8.4) – described in the 11th edition of *Merriam-Webster's Collegiate* (2003) but not in earlier editions – and the figurative use of *food chain* is also now accounted for in many dictionaries.

From the Enlightenment onwards, philosophers have been perplexed – even irritated – by the phenomenon of polysemy. The idea that a single word could have multiple meanings looks messy, and seems necessarily to entail

[25] We discussed this with our colleague Adam Kilgarriff, who notes (personal communication): 'The word has done what words very often do when they come into general discourse from a specialist field, which is to lose their specific denotation, and become words which are used for their connotation (or colour) rather than their denotation.'

ambiguity.[26] But experience tells us that true ambiguity – where one person genuinely misinterprets the meaning intended by another – is exceptionally rare. This is the reality of communication (whatever counterexamples are cooked up by theoretical linguists), and it suggests a more positive take on polysemy. Rather than being a weakness, polysemy could be seen as an elegant design feature which, with maximum economy, enables language to respond to new situations while keeping to a minimum its demands on our short-term memories and processing capacity.

8.6.2 Meanings and 'dictionary senses'

Most people would agree that words have meanings, sometimes multiple meanings. But meanings and dictionary senses aren't the same thing at all. Meanings exist in infinite numbers of discrete communicative events, while the senses in a dictionary represent lexicographers' attempts to impose some order on this babel. We do this by making generalizations (or abstractions) from the mass of available language data. These generalizations aim to make explicit the meaning distinctions which – in normal communication – humans deal with unconsciously and effortlessly. As such, the 'senses' we describe do not have (and do not claim) any special status as 'authoritative' statements about language. Rather, their purpose is to enable dictionary users to associate what they have encountered in a specific context with a particular area of meaning. They function as prompts, in other words, intended – as Bolinger famously said (1965: 572) – to help the reader 'relate the unknown to something known'.

8.6.3 How to find word senses

The process looks something like this. The Lexicographer . . .

(1) analyses instances of usage, typically in concordances or lexical profiles (§4.3.1.2, §4.3.1.5), and
(2) provisionally identifies different word senses (this is the subjective, intuitive part), then
(3) collects good, typical corpus examples for each of these provisional senses. As long as you have plenty of data, one-off oddities can

[26] Thanks to Patrick Hanks (personal communication) for these observations.

usually be ignored, but ambiguous cases (the examples that you can't confidently assign to one or other of your provisional senses) should be stored for further analysis (step 5);

(4) analysing each cluster of examples in turn, the lexicographer identifies the features that are typically associated with it (and that distinguish it from all the other clusters);

(5) finally, our inventory of senses is refined if necessary (which may involve further splitting, or conversely, lumping of closely related clusters) so that all uses of the word that occur frequently in text are fully accounted for.

Each dictionary sense identified through this process has its own unique permutation of the indicators we discussed earlier in this chapter: 'external' ones like text-type and domain, 'internal' ones like syntactic or collocational preferences (see for example the analysis of *argue* in Figure 8.8). These senses, reflecting as they do recurrent phenomena, represent the norms of the language. Any one-off exploitations – examples of usage that can't be straightforwardly assigned to one of the senses – can usually be interpreted by reference to these norms.

Before we conclude this chapter, a couple of potential pitfalls are worth mentioning. A usage becomes a norm – and hence something that deserves to be described in a dictionary – when it is judged to be 'part of the language'. The key to this is recurrence: we confer 'sensehood' on those uses which can be observed to recur independently in a number of different texts. A use may be restricted to one particular domain or subculture (§8.5.1.4), but recurrence within those text-types indicates its presence in the mental lexicons of a significant section of the speech community. However, there is sometimes a danger of overspecifying; of extrapolating from a specific context a dictionary sense which has no independent validity. A striking example of overspecification can be found in the entry for the noun *rot* in Webster's *Third International* (1961), which includes this:

> **6:** the falling of several cricket wickets in quick succession

This 'sense' probably arose because the compilers had a number of citations for *rot* in contexts like these:

> *Once Tendulkar was dismissed, the rot set in.*
> *Warne took the wicket of Alastair Cook, and that's what started the rot.*

It's a fair bet that the compilers inferred from their citations that this expression had some special status in the vocabulary of cricket (about which – being American – they probably knew little). But this is to confuse context and meaning. When *rot* is used in expressions like this, it refers to a process of deterioration and can be applied to all sorts of situations, as the following instances show:

> *It was a nice area until a few years ago, then the rot set in.*
> *Donald shows how the rot set in with Richard Nixon and Watergate.*
> *The composer's father unwittingly started the rot over two centuries ago when he took the six-year-old Mozart to market, as it were.*
> *In years gone by, the Bank's governor would have stopped the rot by calling somebody in to his office for tea.*

Moon (1987b) gives some good examples of this phenomenon. She shows how the temptation to elevate contexts into senses is especially acute in the case of adjectives, where different sets of complements may suggest discrete senses. As she points out, definitions which are overly context-bound 'may misrepresent the nature of the word and destroy its semantic integrity' (181).[27] The availability of large corpora exacerbates the problem. The last word on this topic goes to Patrick Hanks, who points out (2000b: 208) that 'as new citations are amassed, new definitions are added to the dictionary to account for those citations which do not fit the existing definitions. . . . Less commonly is asked the question "Should we perhaps adjust the wording of an existing definition to give a more generalized meaning?".'

Somewhat counterintuitively, another danger lurks in an over-attachment to the notion of consistency. The entry for *whisky* in *OALD-7* defines the drink then adds a second sense:

> [C] a glass of whisky: *a whisky and soda|Two whiskies, please*

This is a classic example of regular polysemy (§5.2.4, §8.3.5). The whole point about regular polysemy is that the alternations it identifies can be applied to any members of a semantic set. Yet if we look at the *OALD*'s entries for *absinthe*, *Calvados*, *crème de menthe*, and *grappa*, we find no mention of the 'glass of' sense. Surely this is inconsistent? Well, yes – but that doesn't necessarily mean it is wrong. In all these cases, the lexicographer will have observed that the countable use of the word is either rare or unattested. S/he will then have concluded that there are better ways of using the

[27] van der Meer (2006) also addresses this issue, showing how some sense distinctions in dictionaries may 'be solely, or at least largely, ascribable to *contexts* of use' (602).

available space in the dictionary, and that anyone who actually encountered an instance of 'two absinthes, please' would be able to infer the meaning by analogy with what happens with commoner words like *whisky*. This kind of attitude, with its emphasis on pragmatic criteria and subjective judgments, infuriates people in the language-engineering community: a computational lexicon has to record somewhere that the 'glass of' alternation is available to *every* member of this semantic class, and NLP-ers expect dictionaries to be equally rigorous. But dictionaries are designed for human readers, and humans are well-equipped to cope with this apparent lack of system. The moral here is that consistency shouldn't be pursued at all costs, and we remind ourselves once again that the dictionary's job is to deal with 'the probable, not the possible'.

8.6.4 Last words

At the beginning of this chapter, we observed a disjunction between the expectations of dictionary users and the behaviour of language-users. On the one hand, dictionary users expect words to be chopped up into 'senses' that are thought to instantiate distinct and mutually exclusive meanings. On the other hand, real communication consists of individual language events, whose participants don't think in terms of 'word senses' – yet seem to handle quite effortlessly the inherent ambiguity of many words, and frequent encounters with 'new' uses which their dictionaries don't account for. The approach outlined here enables us to resolve this conundrum. By focusing on *context*, we can observe the specific patterns of usage that regularly appear when a particular meaning is invoked. The various indicators we have described (§8.5) throw up clusters of examples which all behave in much the same way – and from these clusters we abstract our dictionary senses. In many cases, no single diagnostic test is conclusive, but cumulatively the various features discussed here can underpin a reliable account of meaning. The beauty of this methodology is that it starts and ends with the observable data. It does not rest on any *a priori* theory of meaning; rather it recognizes humans' intuitive (*subjective*) ability to find meaning in communicative events, but complements this (adding an *objective* element) by making explicit the criteria by which senses were identified.

Lexicographers take a realistic and pragmatic view on all this. They know what is expected of them, but are aware of the inherent limitations in the task of word sense disambiguation. The notion of 'word senses' may indeed be a construct of dictionaries, a product of our need to create helpful

distinctions for dictionary users. No description of a word's meaning can ever be 'complete', in the sense of being comprehensive enough to account for every single instantiation of that meaning in human communication. Judgments are relative rather than absolute, and if one dictionary divides a word into six senses, and another into four, neither account is necessarily 'better'. Like most aspects of lexicography, WSD is always somewhat provisional but the process can – through a combination of theoretical insights and practical strategies – be made more systematic, easier to complete, and more likely to deliver satisfactory results.

Exercises

1 Analysing the noun *bite*

Call up a concordance for *bite* as a noun, and make a sample of no more than 300 lines. Then:

- Divide it into broad word senses (no need to be too fine-grained).
- For each sense you have identified:
  - provide a rough description of the meaning (this doesn't need to be a fully-developed dictionary definition)
  - list the contextual clues or 'preferences' that support your sense
  - note one line from the corpus which clearly instantiates this meaning.
- Note any corpus lines which can't unambiguously be assigned to one specific sense.
- Decide how you would group or order these senses in an entry for *bite*, and explain your decisions.

2 Comparing sense divisions in different dictionaries

Choose any two dictionaries of the same general type which are aimed at the same kinds of user. Then:

- Look up the entry for *command* (noun and verb) in each dictionary.
- Compare the two entries, noting:
  - points where the dictionaries agree
  - points where they disagree
  - meanings in one dictionary not accounted for in the other.
- Decide which dictionary's account you prefer, and explain why.

You can also try this with two dictionaries of different types.

Reading

Recommended reading

Apresjan 1973; Cruse 2002, 2004 (chapters 5–12); Hanks 2000b, 2001, 2002, 2004a; Kilgarriff 1997a; Moon 1987a; Rundell 2002; Stock 1984; Taylor 1990.

Further reading on related topics

Atkins 1993; Béjoint 1990; Bolinger 1965, 1975 (chapter 7); Braasch 2006; Cowie 2001; Cruse 1986; Fillmore and Atkins 2000; Geeraerts 1994; Hanks 1988, 1990, 1998, 2000a; Hoey 2005 (chapter 5); Kilgarriff 1998, 2006a; Melčuk 2000; Moon 1987b, 2004; Robins 1987; Rundell 2002; Rundell and Stock 1992; Sinclair 1996; van der Eijk, Alejandro, and Florenza 1995; van der Meer 1999, 2004, 2006; Wierzbicka 1985; Wilson forthcoming.

Prototype theory: Cruse 1990; Geeraerts 1990; Hanks 1994; Lehrer 1990; Rosch 1973, 1975; Taylor 1995; Vandeloise 1990; Wierzbicka 1990.

Regular polysemy: Copestake and Briscoe 1995; Nunberg and Zaenen 1992; Ostler and Atkins 1992.

Metaphor and metonymy: Csábi 2002; Lakoff and Johnson 1980; Knowles 1996; Knowles and Moon 2006; Meyer et al. 1998; Moon 2004; Pinker 1997 (352–385); van der Meer 1996.

9

Building the database (2): the lexical unit

In this chapter we guide you through the process of compiling entries for a monolingual database. The objective is to set out the lexicographic techniques needed for writing database entries, and to do this we need illustrative material. The language used to illustrate the practical points is English, but we hope that the techniques themselves may be adapted to fit the needs of any language.

All dictionaries are of course databases. However, we use this term specifically to denote the preliminary detailed database which was built to hold material recorded during the corpus analysis process (cf. §4.2.2), and from which will be drawn the facts needed for the actual dictionary entries. The focus of Chapter 8 was the lemma, and the task of dividing a polysemous lemma (or headword) into senses (or lexical units). The focus of this chapter is the lexical unit itself (the LU). The structure of the database entry described here, together with the various types of information it holds, relates almost entirely to the LU, that is to say to the headword in one of its senses. Figure 9.1 provides an outline of the contents of this chapter. Each type of information is briefly discussed and examples[1] given of how it may be recorded in a database. Every project will have its own approach to database structure, database fields, and the kinds of item which populate

[1] The examples are drawn from a pilot study for the *New English-Irish Dictionary*. This project was carried out by the Lexicography MasterClass Ltd for Foras na Gaeilge, see http://www.focloir.ie/ . Some of the entries from which extracts have been taken were compiled by Valerie Grundy.

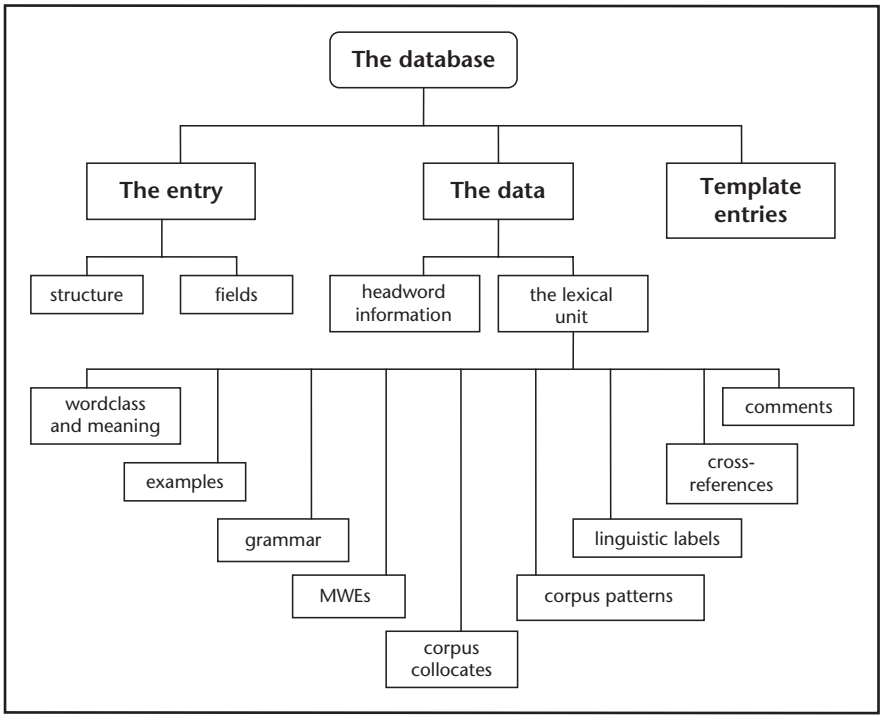


Fig 9.1 Contents of this chapter

them. As long as the item is correctly identified and systematically recorded, the database will serve not only as a launchpad for a single dictionary, but as a source of data for other reference books and textbooks.

9.1 The entry

When I took the first survey of my undertaking, I found our speech copious without order, and energetick without rules: wherever I turned my view, there was perplexity to be disentangled, and confusion to be regulated; choice was to be made out of boundless variety, without any established principle of selection.

Samuel Johnson, *Preface* (1755)

Johnson's words are almost as true today as they were when he wrote them. Like him, we are faced in the corpus with language 'copious without order and energetick without rules': our job is to disentangle the perplexity, and regulate the confusion. However, one advantage we have over Dr. Johnson is the concept of 'lexicographic relevance', which does give us an 'established principle of selection'. Our CQS of choice, the Sketch Engine (cf. §4.3.1),

puts lexicographic relevance into practice, and the results are already more orderly than the raw corpus data, or even the concordanced output. With a carefully designed and well-structured database we can go further, and record the lexicographically relevant facts about a headword in a form that allows someone else to use them as the basis for a dictionary entry or other reference work. The various fields in the database bear the names of the type of data they are designed to hold, so that they effectively act as a prompt to lexicographers studying the corpus and collecting significant facts about their headwords. Some of the fields in the database (for instance, morphological inflections) can be filled automatically, and these are omitted from consideration; consequently the database described here is not exhaustive. As computers get smarter, more and more of the fields will succumb to automatic population direct from the corpus. However, lexicographers need to be aware of the types of data that are important to the description of a word in a dictionary, and recording facts in a structured database is excellent training.

Much of the discussion in Chapter 7 on the dictionary microstructure is relevant here. The difference between the initial database (the subject of the present chapter) and the finished dictionary (Chapters 10 and 12) is that the database is much more detailed in all respects, allowing dictionary editors to select from it what is needed for their particular users. Database and dictionary differ in structure as well as size, but in both cases the actual fields, their names and what goes into them, depend on the policy set out in their individual Style Guides (cf. §4.4). However, most of the fields discussed in this chapter will find their place in the first stage of any dictionary-writing process.[2]

9.1.1 Entry structure

A word, or *lemma*, has one entry in the database; this entry consists of one or more subentries, the first of which (the 'leader') holds information attaching to the lemma itself rather than to any of the senses. If the headword has only one sense then this subentry is fleshed out to contain all the information necessary for a full description of the word. Otherwise, each one of the lemma's lexical units (LUs) has its own subentry, as does each multiword expression (MWE) in which it participates. Every fact recorded

[2] Since the various linguistic phenomena cannot be illustrated with reference to any single word, the material will be drawn from analyses of various lemmas.

in the database is described in terms of its relationship with the individual LU that forms the focus of the subentry.

9.1.2 The fields in a lexical entry

The entry components examined in §7.2 are reflected in the database fields of the initial analysis. The most important of these are listed in the table in Box 9.1, which will serve as a reference point to help you understand this chapter; in the points below, numbers in parentheses refer to that table. Some of these fields will be automatically generated by database software, but for clarity's sake they are still listed and discussed here as though they required manual completion. The fields are listed in the table in roughly the order they will be used as the entry is built, but since some of them will be needed several times within the same subentry it's not possible to give them any set order. The rest of this chapter is devoted to describing these fields, one by one, and showing how they are used. The issues discussed are illustrated with material from lexical entries (cf. §6.6.1), as opposed to those for abbreviations, grammatical words, and encyclopedic material. The design of grammatical word entries depends largely on the language and wordclass involved. In the case of the abbreviations and encyclopedic material, the structure is normally very pared down, and not difficult to use or understand.

As much as possible of the material entered into the database is formalized, in order to make the whole as systematic (and therefore useful) as possible. The types of data in the database are outlined here:

- The most structured types of data – e.g. wordclass (7), construction (11) – consist of pre-ordained codes denoting a particular grammatical category, selected by the database editor from a list of options and assigned to facts identified in the corpus. The use of such codes means that much of the database is searchable by computer.
- Similarly selected from a pre-ordained list are the linguistic labels (21), which the editor uses to show divergence from default unmarked vocabulary, and MWE-type (16) assigned by the editor from a closed list of options.
- All the material in the example fields (10) is taken direct from the corpus, as are the contents of the collocate field, which are flagged by the corpus frequency program.

Box 9.1 The database entry

	Field	Explanation
1	HEADWORD	the first field in every entry and subentry
2	HOMOGRAPH #	
3	VARIANT FORM	these fields normally occur only once, in the main entry, rarely in the subentry for an LU
4	INFLECTED FORM	
5	FULL FORM	
6	LU #	often known as the 'sense number', this is a unique identifier of the subentry
7	WORDCLASS	of the headword in this LU, e.g. *noun, verb*, etc., cf. §9.2.3
8	GRAMMAR	additional grammatical information, and other miscellaneous information, cf. §7.2.6.3, §9.2.5
9	MEANING	an informal description of the meaning of the headword in this sense (polished definitions are for the dictionary entry proper), or of the MWE if that is the focus of the subentry, cf. §9.2.3
10	EXAMPLE	a sentence extracted from the corpus, illustrating the fact (construction, collocate etc.) that it is attached to, cf. §9.2.4
11	CONSTRUCTION	a grammar field, recording (usually in a pre-ordained code) one of the headword's lexicographically relevant co-constituents (cf. §9.2.5), e.g. *NP*, *Vinf*, *cl-wh*, etc.
12	NULL INSTANTIATION	
13	NULLINST-TYPE	cf. §9.2.5.5
14	NULLINST-SEMANTICS	
15	NULLINST-SYNTAX	
16	MWE-TYPE	type of MWE that forms the focus of the subentry, e.g. *idiom*, *phrasal verb*, *compound*, etc., cf. §9.2.6
17	MWE	the actual multiword expression being recorded, e.g. in the entry for *potato* this might contain the MWE *a hot potato* (in the sense of a sensitive issue), cf. §9.2.6
18	COLLOCATE	a word which is a statistically significant collocate of the headword, cf. §9.2.7
19	COLLOCATE-TYPE	the lexical set to which a group of collocates belongs, cf. §9.2.7.3
20	CORPUS PATTERN	an informal description of some recurrent patterning found in the corpus, often used to record semantic prosody, cf. §9.2.8

(*cont.*)

Box 9.1 (*Continued*)

	Field	Explanation
21	Label: REGION (*or* DOMAIN, REGISTER, STYLE etc.)	the name of the type of label to be inserted (cf. §7.2.8); the value would be *Australia* for an Australian English item, or *Peru* for Peruvian Spanish, etc. The actual label types offered to the lexicographer depend on dictionary policy, set out in the Style Guide, cf. §9.2.9.
22	CROSS-REFERENCE	cf. §9.2.10
23	COMMENT	any informal note made for the benefit of editors using the database subsequently, cf. §9.2.11

- Certain kinds of material are supplied by editors from their own knowledge of the language, e.g. variant form (3), inflected form (4), nullinst-type (13), collocate type (19), etc.
- Other types of data consist of free-form text written by the editor, e.g. meaning (9), nullinst-semantics (14), collocate-type (19), corpus pattern (20), or comment (23), etc.

The entry has a certain syntax: the principal fields may be assembled in several ways, but a valid ordering must be maintained throughout. (The software keeps you on track here.) Only a COMMENT (23) may be inserted anywhere in the entry.

9.2 Data

Every fact in an entry is described in terms of its relationship with the headword. A database entry should be at least two or three times bigger than the final dictionary entry. It should be so rich that the people using it to build the final dictionary entry rarely if ever need to go back to the corpus. Because it is so structured, it is easy for dictionary editors to skim the database entry, sift out what's not needed, and construct the dictionary entry from the facts best suited to their particular dictionary.

As with dictionaries, the database entry has to hold more than some users really need if it is to be of any use to other, more demanding, users. Grammatical and collocational facts are always essential, as are examples

of the headword in the context of its significant collocates. However, if you're writing a monolingual dictionary entry you probably won't need many examples of the headword in a 'normal' context. If on the other hand you are writing a bilingual dictionary, then you have to test your translations against as many contexts as possible, so the more examples the better.

The database distils the multifarious facts found in the corpus into an easily accessible store of lexical information. To be useful for a bilingual dictionary, it must contain a wealth of different contexts in which the headword is found. It must offer the translation editor a way of matching an SL item (word or multiword expression) with its TL equivalent, if such exists, taking account of such parameters as:

- the lexical units of the lemma (the various senses the word can have); and for each LU . . .
- its semantic content or basic meaning
- its semantic scope: how specific or general this meaning can be – the LU's metaphorical extensibility
- its morphological properties
- its inherent grammatical properties (part of speech, gender, etc.)
- its valency: the way it combines syntactically with the semantically significant words in its context
- its participation in idiomatic phrases
- its significant collocates: the specific words frequently found with it in the major grammatical relationships (verb–object, subject–verb, adjective–noun, etc.)
- the functions it performs in the language, as for instance an itemizer of mass nouns or a collective of plural nouns, a support verb, etc.
- important grammatical and semantic patterns of behaviour found in corpus data
- other aspects of meaning and use, such as style, register, region, pragmatic force, etc.
- the various source texts from which the corpus citations are taken, together with their textual properties such as medium, genre, date, authorship, and so on.

The database editor must systematically look for and record as many of these linguistic phenomena as are relevant to the lemma being analysed.

➔ Remember that building the database is a wholly *monolingual* exercise, even if your database is likely to be used first for a bilingual dictionary.

If you are to give an objective account of a word's behaviour in the language surveyed, the database must be target-language neutral. As soon as you start to think 'I know how this is translated', you start to make selection decisions for the wrong reasons.

➔ When you're compiling a database entry, the slogan is (unlike in exams) *When in doubt don't leave it out*.

A small fact that may seem insignificant to you could be important to the editor finalizing the dictionary entry. Be generous with examples from the corpus, especially if your dictionary project is a bilingual one. The translators and editors who will use your work as a basis for their dictionary entry need a lot of contexts for every use of your headword.

9.2.1 What is a headword?

This is not a problem in the case of the vast majority of English words: they will clearly be headwords. However, if you are writing a Style Guide, there are some tricky points on which you have to give as clear guidance as possible (though there is always room for doubt). Principal among these issues for Style Guide writers are:

(1) Participial adjectives
It is often difficult to decide whether adjectival uses of a present or past participle should be treated within the entry for the verb, or as headwords in their own right. They tend to be given headword status when there is evidence of their normal use in both attributive and predicative position, as for instance in the case of words like *surprised*, *broken*, *amazing*, and *disturbing*.
➔ A useful rule of thumb is: if it works with a modifier, such as *very* or *absolutely*, make it a headword.

(2) Gerunds (nouns in *-ing*)
When the gerund form has a meaning clearly distinct from that of the verb, and/or when this noun use is very frequent, then it should be given headword status.
➔ The rarer the noun in *-ing*, the less likely it is to be a headword.

(3) Hyphenated combining forms
These are often entered as one LU within the entry of an associated headword, so that *-haired* (as in *brown-haired*) is in the *hair*

entry, *-splashed* (*mud-splashed*) in the *splash* entry, *-nosed* (*long-nosed*) under *nose*, and so on.

➔ These are easy to recognize, just explain clearly in the Style Guide how to handle them.

(4) Compound prepositions

Examples are *in spite of*, *according to, owing to*. If you're writing the Style Guide you have to be clear about how to handle these.

➔ Here again, say exactly what to do with them. It's a good idea to list them all, as there are so few.

(5) Plural headwords

The canonical form of some nouns is a plural, in particular 'clothing' words like *clothes*, *jeans*, and *overalls.* Some plural nouns, like *trousers* and *pyjamas*, have singular forms (*trouser leg, pyjama top*), which compounds the problem. In the case of other plural nouns, like *glasses* (for seeing with), *arms* (weapons), *ceramics*, *proceeds*, *troops*, etc., the singular forms (*glass*, *arm*, etc.) have a quite distinct meaning and often belong to a different wordclass. When you're writing a Style Guide you need to tease out all these issues and for each type give clear guidance on what form should have headword status.

9.2.2 Headword information in the 'leader' record

Every entry and subentry begins with a statement of the headword in the HEADWORD field: this is what links all the information about one word together in one entry. In it goes the *canonical form* of the headword: the singular of nouns, the infinitive of verbs, the uninflected form of adjectives and adverbs, and so on. A small set of fields – HOMOGRAPH NUMBER, VARIANT FORM, FULL FORM, and INFLECTED FORM – normally occur only once in each entry, in a 'leader' record which precedes the various LU subentries. The homograph number is manually entered, and the other three fields hold information often relating to the inherent properties of the lemma (cf. §5.5.1), including the full form of an abbreviation headword, any regional variant spelling of the headword, and the headword's morphological inflections. The last-named is normally automatically supplied from lists already existing elsewhere. Figure 9.2 gives some examples of these fields in action in the leader record for *weave*, *aluminium*, and *EU*. Note that in the case of *aluminium* the two forms are labelled with the relevant regions.

HEADWORD	**weave**
HOMOGRAPH NUMBER	1
INFLECTED FORM	wove (past), woven (past participle)
MEANING	interlace (threads etc.) to make fabric
HEADWORD	**aluminium**
REGION	GB
VARIANT FORM	aluminum
REGION	US
HEADWORD	**EU**
FULL FORM	European Union

Fig 9.2 Fields and data from the initial 'leader' record

9.2.3 Wordclass and meaning of the LU

The first field in an LU subentry is the HEADWORD, as already explained. Next comes the LU # (each LU having a unique identification number), and after that any of the fields discussed in §9.2.2, if required. For instance, irregular inflections are sometimes restricted to one LU, and thus the INFLECTED FORM may be needed within a subentry. This happens in the case of the headword *weave*, shown in Figure 9.3, where both the strong and weak forms of the simple past tense are found for the 'move in and out' LU, as opposed to the LU meaning 'make a fabric by interlacing threads etc.', where only the strong past tense is found.

1	minutes before the final whistle, he	weaved	his way infield
2	dizzy with pride and vodka, she	weaved	back to Drew's car
3	parked behind the dustbins – she	weaved	towards it, finding her vision blurring
4	As the ambulance	weaved	its way dramatically through the . . .
5	A helicopter tracked the car as it	weaved	in and out of traffic near Bristol
6	he yelled something as he	weaved	from side to side across the road
7	They	wove	off through the theatre crowd
8	Herr Nordern	wove	unsteadily across the living room
9	you	wove	your way along the arcade
10	The car	wove	through the traffic on
11	They	wove	in between the cars that lined the drive
12	the demonstrators	wove	through the downtown area

Fig 9.3 Strong and weak past tense forms of *weave* in similar contexts

Similarly, when two LUs of a noun headword have different plural forms (see Figure 9.4) these are recorded in the INFLECTED FORM field, as in Figure 9.5, where the two plurals of *mouse* are shown: *mice* for the 'animal' LU and *mouses* for the 'computer' LU, which is quite common in speech.

1	difficult to keep clean, and	mice	and cockroaches everywhere
2	the lifespan of normal	mice	in laboratory tests
3	He wondered if the	mice	were white.
4	cereals must be protected from	mice	and other vermin
5	gerbils, white	mice	and hamsters
6	your mouse mat – what you run your	mouses	along
7	. . . software, we're getting three	mouses	for those whose sight is impaired
8	internet and television, fingering our	mouses	and remotes, clicking and zapping
9	operate their computers and	mouses	with their feet
10	the screens and printouts and	mouses	

Fig 9.4 Two plurals of *mouse* in different senses

HEADWORD	**mouse**	HEADWORD	**mouse**
LU #	1	LU #	2
WORDCLASS	noun	WORDCLASS	noun
INFLECTED FORM	mice (pl)	INFLECTED FORM	mouses, mice (plural)
MEANING	small rodent	MEANING	computer device
HEADWORD	**weave**	HEADWORD	**weave**
HOMOGRAPH #	1	HOMOGRAPH #	2
WORDCLASS	verb	WORDCLASS	verb
INFLECTED FORM	wove (pt), woven (ptp)	INFLECTED FORM	wove or weaved (pt), weaved (ptp)
MEANING	interlace (threads etc.) to make fabric	MEANING	move in and out of obstacles

Fig 9.5 The first few fields in four LUs

Every subentry begins with a statement of the wordclass of the LU (one of its 'inherent properties'; cf. §5.5.1), and an informal description of its meaning. The WORDCLASS field contains one of the nine or so principal parts of speech: *adjective*, *adverb*, *conjunction*, *determiner*, *interjection*, *noun*, *preposition*, *pronoun*, and *verb.* The contents of this field depend on the dictionary policy: some projects demand more detailed subclasses at this point, such as *noun:count*, *pronoun:personal*, *determiner:article*, *verb:transitive*, and so on. It's often helpful to use subclasses of nouns, pronouns, and determiners from the start of the analysis, but it is usually more rewarding not to pre-classify the verbs but instead to note their valency in the form of constructions (cf. Box 9.2 below). Assigning the common subclasses (*transitive*, *intransitive*, etc.) is then done at dictionary entry stage.

The description of the LU's sense in the MEANING field has to be detailed enough to differentiate it from other LUs, but doesn't need the polished wording of a definition in the final dictionary entry. The meaning description is of use only to the editors who use the database either for

translation purposes or as the launchpad of a monolingual dictionary. Figure 9.5 shows the initial fields in some LU subentries.

9.2.4 Examples

The EXAMPLE field can be placed almost anywhere in the entry. It is there to hold a corpus sentence that illustrates *one specific fact*: the example is inserted immediately after the fact it illustrates. This can be any kind of fact – the meaning of the LU, a grammatical construction associated with it, one of its corpus collocates, and so on. It's easy to see from the field names what that fact is: in Figure 9.6, the facts being illustrated are MEANING and INFLECTED FORM, and in the case of the latter, all the inflected forms recorded in the entry are exemplified.

Database examples and dictionary examples are quite distinct. The function of the example in a dictionary entry, together with what's important when you're choosing them, is discussed in §10.8 (monolingual dictionaries) and §12.3.3 (bilingual dictionaries). Only database examples, collected in the course of corpus analysis, concern us now.

➔ The general rule is that every fact entered into the database should have its supporting corpus example beside it.

In principle, your example should consist of a complete sentence. In print dictionaries there is rarely room for such a luxury, but with the electronic database it's different. The full sentence will be welcomed by the editors who use the database. However, there's no rigid rule about this, and in practice it's often possible – and permissible – to shorten a corpus sentence without losing anything of value. For instance, the second example sentence in the first *mouse* entry in Figure 9.6 was originally

> *Residents of the building said it was infested with rats, **mice** and roaches[, and that it sometimes lacked electricity].*

and the bracketed section was not included. Similarly, the example in the *weave*[1] entry was abridged from

> *The thread is spooled on an enormous reeling-machine [(nituchha)] before being **woven** on a primitive loom into men's shirts and trousers[, household linen and curtains (pologa) to protect the sleeper from the mosquitoes that are so prevalent in the marshes of the western plains].*

You can have literally hundreds of examples of one use of your headword in your corpus and not find a single one that is perfect. You're looking for a sentence that matches the following description:

HEADWORD	**mouse**
LU #	1
WORDCLASS	noun
INFLECTED FORM	mice (pl)
MEANING	small rodent
EXAMPLE	She saw the little **mouse** lying dead in its corner covered in dust.
EXAMPLE	Residents of the building said it was infested with rats, **mice** and roaches.

HEADWORD	**mouse**
LU #	2
WORDCLASS	noun
INFLECTED FORM	mouses, mice (pl)
MEANING	computer device
EXAMPLE	I am left-handed and find it hard to 'right click' with the **mouse**.
EXAMPLE	These will allow you to daisy-chain devices such as scanners, **mice**, keyboards and more.
EXAMPLE	That's what they use up the school – proper computers, the screens and printouts and **mouses.**

HEADWORD	**weave**
HOMOGRAPH #	1
WORDCLASS	verb
INFLECTED FORM	wove (pt), woven (ptp)
EXAMPLE	Turn it into a utopian commune where women **weave** medieval tapestries.
EXAMPLE	Her name was Amadé, and it was she that **wove** the baskets.
EXAMPLE	The thread is spooled on an enormous reeling-machine before being **woven** on a primitive loom into men's shirts and trousers.

HEADWORD	**weave**
HOMOGRAPH #	2
WORDCLASS	verb
INFLECTED FORM	wove or weaved (pt), weaved (ptp)
MEANING	move in and out of obstacles
EXAMPLE	You **weave** uncertainly along the road at first, but soon you pick up speed and rush smoothly forward.
EXAMPLE	The car **wove** through the traffic on Hyde Park Corner and purred up Park Lane, the grime-grey hotels flashing by on the right.
EXAMPLE	A helicopter tracked the car as it **weaved** in and out of traffic near Bristol.
EXAMPLE	He had seen her only once as she'd **weaved** her way across the gardens.

Fig 9.6 Examples illustrating the headword's inflections and meaning

- It's short.
- It provides an 'informative context' for the headword, i.e. the sentence itself helps you to understand what the headword means.
- It has no words in it that are more difficult to understand than the headword.
- It doesn't include words at variance with the register, style, region, etc. of the headword.
- It doesn't contain the name of a real person, living or dead. (Everyone in the dictionary business can quote at least one case of a pejorative reference – often in a corpus-derived example – being removed from a dictionary minutes before publication.)
- It has no foreign words in it.

The reason for insisting that every fact should be accompanied by its own illustrative example is because the body of example sentences in a headword's entry should provide everything needed by the editor who comes along later and extracts a dictionary entry from the database. That person will abridge or adapt the long corpus sentences, making them into examples suitable for the users of the dictionary in question. The database examples are there to provide a *model* for the dictionary examples (cf. §10.8 and §12.3.3 for more detail).

➔ Two examples are usually better than one, and three are often better than two. There is no space restriction on database material, so if you find several really good sentences exemplifying the same fact, put them all in.

9.2.5 Grammar in the database

Grammatical information in the LU subentry consists of three types of facts:

- the wordclass of the headword (in the WORDCLASS field, cf. §9.2.3 above)
- additional grammatical information about it, recorded in the GRAMMAR field (cf. §7.2.6.3)
- details of its syntactic environment, recorded in the CONSTRUCTION field (cf. §7.2.6.2), and supported by corpus sentences illustrating each grammatical pattern.

The latter include all of the lexicographically relevant co-constituents of the clause in which the headword is found. These are the *constructions* which

every speaker of the language must know in order to use a word correctly, and which consequently are given considerable prominence in dictionaries, both monolingual and bilingual, for language-learners. They are recorded in the CONSTRUCTION field, often in the form of a code. The actual codes to be used in any particular database should be detailed in the project's Style Guide (cf. §4.4), together with examples of the construction denoted by each one. In projects using dictionary-writing software, the agreed codes will be listed in a menu offered to editors when they have to insert one of them into a CONSTRUCTION field.

A table of lexicographically relevant co-constituents for each of the four major wordclasses is given in the following sections. In these tables, the first column shows an abbreviated code for each item (showing the *phrase type* and where necessary its *grammatical function*, cf. §5.5.2); the second column explains the first; and the third gives an example of one or two words for which the item is lexicographically relevant. We shall now consider each of the four major wordclasses in turn.

9.2.5.1 *Verbs* This section covers lexically simple (single-word) verbs; lexically complex verbs, also known as 'phrasal verbs', and support verb constructions are discussed with other types of multiword expressions in §9.2.6. However, in the absence of any objective, watertight definition of a phrasal verb (a term beloved of editors of learners' dictionaries but little known to English native speakers), we include in the table in Box 9.2 a different way to record phrasal verbs, by means of specific particles[3] which may be followed by prepositional phrases; the relevant codes are 'Part-*specific*', 'PP-*specific*', 'Part-*specific* NP' and 'Part-*specific* PP-*specific*'.

Verb constructions The lexicographically relevant co-constituents of verb headwords are given in Box 9.2. together with examples of each, where the exemplifying verbs are in bold print. These co-constituents are the 'constructions' to be noted and recorded when your headword is a verb.

Verbs in the corpus The verb *watch* in the sense of 'look at with attention' will be our demo headword in this section. Finding the constructions that you need to record in the database is a skill that comes with practice. Corpus query software nowadays can speed up this task a lot, as you can see from the Word Sketch[4] in Figure 9.7.

[3] A 'particle' in this description is an adverbial particle.

[4] See http://www.sketchengine.co.uk/ .

Box 9.2 Constructions for verb headwords

A list of lexicographically relevant co-constituents of verb headwords.

Code	Constituent	Examples
AJP	adjective phrase	you **seem** *sad*, he **looks** *taller than you*
AVP	adverb phrase	he **ran** *home*
cl-if	whether/if clause	I **wonder** *whether he will be there*, Do you **know** *if she was there*?
cl-that	indicative clause with *that*	I **hear** *that he's arrived*
cl-(that)	indicative clause without *that*	I **hear** *he's arrived*
cl-that-cond	conditional clause with *that*	he **wishes** *that she would go away*
cl-(that)-cond	conditional clause without *that*	he **wishes** *she would go away*
cl-that-subj	subjunctive clause with *that*	they **demanded** *that he obey them*
cl-(that)-subj	subjunctive clause without *that*	they **demanded** *he obey them*
cl-wh	clause with *what, when, how, where, why*	I **forgot** *what to say*, she **guessed** *when you had arrived*, I **know** *how you feel*
it+	anticipatory 'it' construction	*it* **seemed** there was a mistake.
NP	all types of noun phrase	I **like** *honey*, I **heard** *a story about a man named Jed*, I **dropped** *the lid of my vitamin jar*
NP AJP	noun phrase + adjective phrase	**paint** *it green*, we **found** *it very dull*
NP AVP	noun phrase + adverb phrase	we **took** *him away*
NP cl-that	noun phrase + indicative clause with *that*	**tell** *her that he's here*
NP cl-(that)	noun phrase + indicative clause without *that*	**tell** *her he's here*
NP NP	noun phrase + noun phrase	**show** *him the cheese*, **give** *her a book*, *she* **sewed** *him a shirt*
NP Part-*specific*	noun phrase + named particle	**look** *the word up*, **help** *me down*
NP PP-*specific*	noun phrase + preposition phrase with named preposition	**push** *it through the hall*
NP Vinf	noun phrase + infinitive verb without *to*	**make** *him leave*, she **let** *him go*
NP Vinf-to	noun phrase + infinitive verb with *to*	we **want** *you to leave*, they **dared** *him to do it*
NP Ving	noun phrase + gerund	*she* ***watched*** *the children playing*, I **heard** *him leaving*

Box 9.2 (*Continued*)

Code	Constituent	Examples
Part-*specific*	named particle	it **died** *out*
Part-*specific* NP	named particle + noun phrase	**look** *up the word*, he **took** *off his hat*
Part-*specific* PP-*specific*	named particle + preposition phrase with named preposition	**put** *up with it*, **come** *up with a good idea*
PP-*specific*	prepositional phrase with named preposition, e.g. PP-*at*, etc.	they **looked** *at the screen*, that **depends** *on the situation*
PP-*specific* cl-wh	prepositional phrase with named preposition + wh-clause	he **enquired** *about which train I was taking*
PP-*specific* Vinf-to	prepositional phrase with named preposition with infinitive with *to*	I would **prefer** *for him to go*, they **looked** *to him to do it*
PP-*specific* NP Ving	prepositional phrase with named preposition with noun phrase with gerund	we **counted** *on him fixing it for us*
PP-*specific* Ving	prepositional phrase with named preposition with gerund	don't **insist** *on doing it*, I **thought** *of going*
Quo	quote	*'Get out of here!'* she **shouted**
Vinf	infinitive verb without *to*	you **can** *go*, you **needn't** *do* it
Vinf-to	infinitive verb with *to*	I **love** *to see her laugh*, I **tried** *to go*
Ving	gerund	she **likes** *ironing*, I **hate** *washing dishes*
wh Vinf-to	wh-word with infinitive with *to*	I didn't **know** *what to say*, **watch** *how to do it*

The Word Sketch in Figure 9.7 highlights many valuable constructions to be noted in the database. The most important source of these lies in the block entitled 'unary rel[ationship]s'. The most interesting are shown in Figure 9.8, where the codes used in the Sketch Engine are listed in order of significant frequency in the context of the verb *watch*, and an example is given of each construction. However, since the Word Sketch tables are based on corpus frequency, many essential constructions cannot be found in this way. Every lexicographer should be able to look at a set of concordances for a verb and pick out from them the important constructions to be put into the database and if appropriate to be used in the dictionary entry too. Figure 9.9 shows a set of selected and abridged concordances for the verb *watch* in the sense of 'look at with attention'.

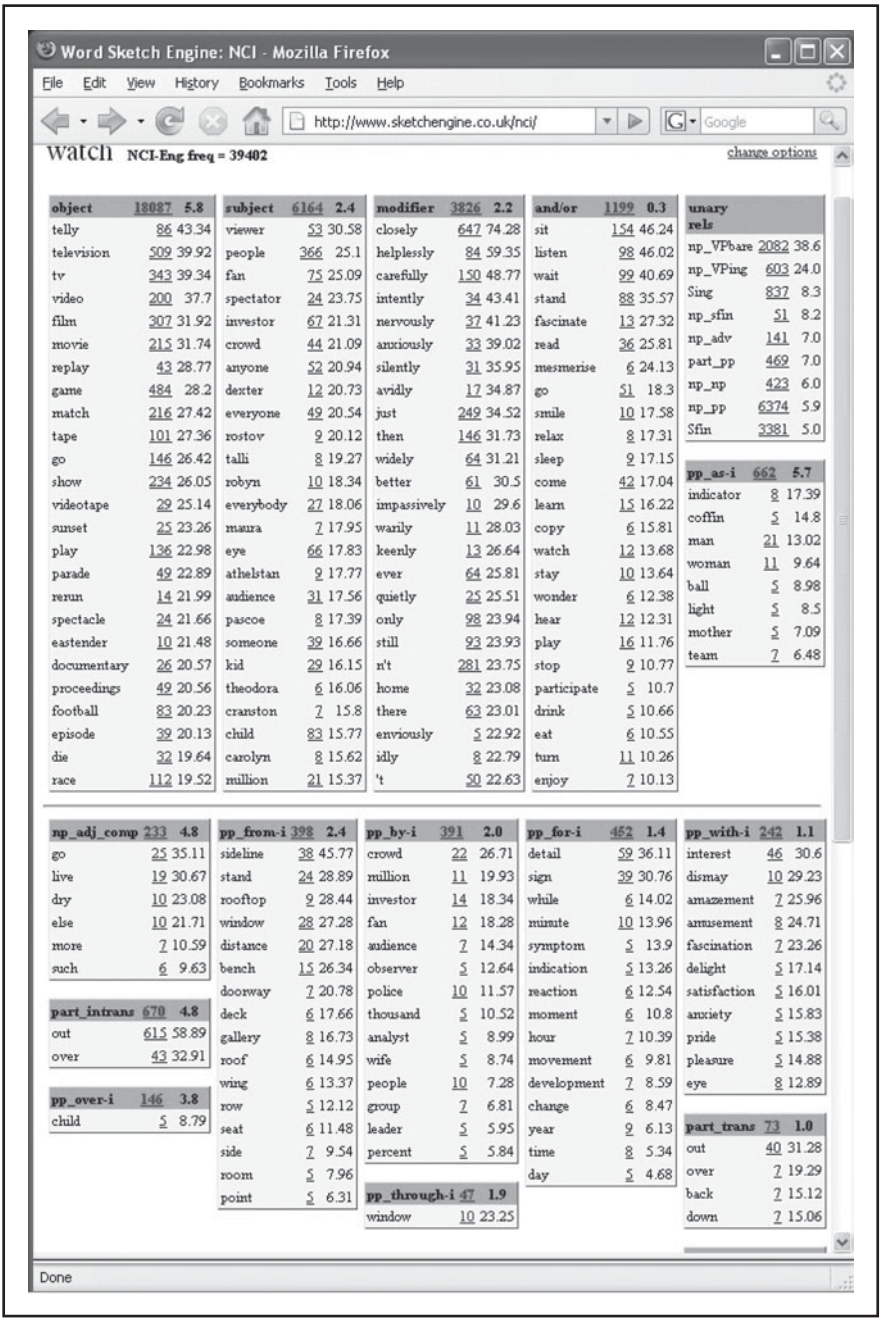


Fig 9.7 Word Sketch of the verb *watch*

Code	Example of construction
np_VPbare	*I watched* **the other passengers go on** *towards the passport control*
np_VPing	*we stood there, watching* **smoke rising** *from Panama City*
Sing	*we watched* **the seagulls whirling**
np_sfin	*he saw her watching* **him**
np_adv	**I** *just stood and watched* **him out of sight**

Fig 9.8 Some constructions from the Word Sketch of the verb *watch*

1		"Watch	how people react," he says.
2	"This is it," Emily said,	watching	for Mungo's reaction.
3	. . . how I felt waiting and	watching	for your taxi to turn into the drive.
4	He	watched	Joe heave his bulk out of the chair.
5	He hid in the bushes and	watched	for his dad to leave for work.
6	Helen stood and	watched	through the binoculars.
7	Here you can	watch	how to cook them in the culinary theatre.
8	I stayed to	watch	through the window.
9	Mary stood on the narrow path and	watched	as the estate-car tore down the slope
10	Monica	watched	the two men, fascinated.
11	Pass them together across the back,	watching	how the needles move.
12	She said she had	watched	what happened but they hadn't spotted her.
13	They could only	watch	as the child stood petrified with fright.
14	This is also a good place to	watch	for buzzards.
15	Twenty-two athletes spend days	watching	their teammates do all the work.
16	Visitors	watch	horses going through their morning work .
17	We all	watched	her, unsure of what to say.
18	People in the crowd were	watching,	curiously.
19	We stood by the rail	watching	the luggage being unloaded.
20	You have to	watch	what Sam does.

Fig 9.9 Selected concordances for the verb *watch*

When you're building a database entry, and find yourself faced with a set of concordances like those in Figure 9.9, you have to record in the database each of the constructions you find in the concordances, together with example sentences.

Verbs in the database Two database fields come into play to record the facts found in the concordances for *watch*. They are CONSTRUCTION and EXAMPLE (numbers 11 and 10 respectively in the checklist in Box 9.1). You record separately every construction that you find, together with one, two, or more examples, as in Figure 9.10, where the figures in the left-hand column have been added for ease of reference; for the same reason we highlight in bold the part of each example that instantiates the construction it is illustrating.

1	CONSTRUCTION	**0** (*zero*)
1a	EXAMPLE	Helen stood and watched through the binoculars.
1b	EXAMPLE	People in the crowd were watching, curiously.
2	CONSTRUCTION	**PP-for**
2a	EXAMPLE	This is also a good place to watch **for** buzzards.
2b	EXAMPLE	At about 7 a.m. he walked towards his house and hid in the bushes across the road and watched **for** his dad to leave for work.
3	CONSTRUCTION	**NP**
3a	EXAMPLE	Monica watched **the two men**, fascinated.
3b	EXAMPLE	We all watched **her**, unsure of what to say.
4	CONSTRUCTION	**NP Vinf**
4a	EXAMPLE	He watched **Joe heave** his bulk out of the chair.
4b	EXAMPLE	Twenty-two athletes spend five days for the most part watching **their teammates do** all the work, and at the end of it all, everyone is quite happy to settle for a draw.
5	CONSTRUCTION	**NP Ving**
5a	EXAMPLE	Everything was grey, wet and colourless as we stood by the rail watching **the luggage being unloaded** into the custom sheds.
5b	EXAMPLE	A typical training programme of the modern thoroughbred is explained and visitors watch **strings of horses going** through their morning work and gallops.
6	CONSTRUCTION	**cl-wh/what**
6a	EXAMPLE	She said she had watched **what happened** but the men had not spotted her.
6b	EXAMPLE	You have to watch **what Sam does**.
7	CONSTRUCTION	**cl-wh/how**
7a	EXAMPLE	Set the carriages for circular knitting and pass them together across the back, watching **how the needles move**.
7b	EXAMPLE	"Watch **how people react**," he says.
8	CONSTRUCTION	**wh VPinf-to**
8a	EXAMPLE	Here you can discover the latest developments in commercially grown mushrooms, asparagus and chicory on a grand scale, and watch **how to cook them** in the culinary theatre.

Fig 9.10 Verb constructions recorded for *watch* in the database

The first thing to notice about the database extract in Figure 9.10 is the fact that all the examples are full sentences drawn from the corpus. An effort has been made to select short sentences where possible, but – apart from removing irrelevant and distracting material – no changes have been made to the corpus data. As well as furnishing subsequent users of the database with genuine corpus extracts, this has the advantage of accelerating the database building (it's always faster to copy and paste a whole sentence than to try to edit it), and a smart customized dictionary-writing system can do this very effectively.

In the table in Figure 9.10 the first construction is recorded as '0' (zero): it has no object or other complementation. This of course is the 'absolute' use of the verb, considered an intransitive in traditional grammar. Note that at the database stage, in this model, there is no need to use 'intransitive' and

'transitive' in the description of verb behaviour, although in the dictionary proper *watch* in constructions 1 and 2 might be marked *intransitive*, and in constructions 3 etc. *transitive*. Constructions 4–8 inclusive are more complex, but each one is essential to the full description of the verb in this sense. Such a summary of a verb's valency, recorded as in Figure 9.10 for one LU of *watch*, provides all the constructions a dictionary editor needs to know about when writing an entry for a specific dictionary.

9.2.5.2 *Nouns* This section covers lexically simple (single-word) nouns: compound nouns are handled with other types of multiword expressions in §9.2.6.

Countability is a very basic aspect of noun behaviour in the context of learners' dictionaries, whether monolingual or bilingual.[5] It's also very significant in language use in general, and often indicates a shift in sense: cf. *Look at the little lamb!* and *Do you eat lamb?* For these reasons, we find it useful in the database to distinguish between countable nouns (marked 'C') and uncountable, or mass, nouns ('U'). When working on nouns with corpus data, you need to be aware of the mass–count regular polysemy relationship (cf. §5.2.4, and Figure 5.14 in particular). This is a feature of many English nouns, some of which appear in Figure 9.11. Take the case of the word *coffee*. This word, in the sense of the substance, either solid or

(MASS	*nu*	She doesn't drink coffee.
(UNIT	*nc*	Three coffees and two teas, please.
(TYPE	*nc*	They stock three coffees from Kenya.
(ANIMAL (etc.)	*nc*	There's a lamb.
(ITS MEAT	*nu*	Have some more lamb.
(FOOD: ITEM	*nc*	He put six potatoes into a bag.
(FOOD: MASS	*nu*	Have some more potato.
(TREE	*nc*	She stood by a tall pine.
(ITS WOOD	*nu*	The desk was made of pine.
(SPECIFIC INSTITUTION	*nc*	There are two schools in that district.
(GENERIC	*nu*	School begins at 9 o'clock each day.
(MEAL	*nc*	He won't eat his dinner.
(OCCASION OF MEAL	*nu*	Dinner at eight. We met at dinner.

Fig 9.11 Some instances of regular polysemy in mass–count nouns

[5] A big problem for many learners of English is the use or omission of the definite article in relation to uncountable nouns.

liquid, is uncountable; however when it is used to denote a single unit of coffee (in a cup, for example) it becomes countable; a similar phenomenon occurs when the word is used to mean 'type of coffee'. These three senses would normally be considered as different LUs in the database: what you do with them in a dictionary depends (as always) on the type of dictionary. The other instances of regular polysemy shown in Figure 9.11 also involve the mass–count distinction, and the Style Guide must make it clear how these should be handled, both in database and dictionary.

Two semantic subtypes of nouns are of particular interest to lexicographers: 'itemizers' and 'collectives'. These are often very salient in corpus data and may be considered to represent a special kind of collocate; they are discussed in §9.2.7.3.

Noun constructions The lexicographically relevant co-constituents of noun headwords are given in Box 9.3, together with examples of each, where the exemplifying nouns are in bold print. These co-constituents are the 'constructions' to be noted and recorded when your headword is a noun.

Nouns in the corpus The noun *reason* in the sense of 'cause, explanation, justification' will be our demo headword in this section. Figure 9.12 shows a set of selected concordances, edited to make it easier to see the construction in each. The next step is to record these constructions in the database.

1	This was another	reason	for Lydia to dislike her.
2	They were not required to provide any	reason	for their action.
3	He did so in an attempt to find the	reason	why this happened
4	There's no	reason	why you can't have a normal job.
5	What, then, was the real	reason	behind the decision?
6	This is happening for totally different	reasons.	
7	The main	reason	for his failure to win the contract was . . .
8	We also had other	reasons	for banning meat imports
9	All the more	reason	for Lorraine to look for a bigger home.
10	It could be demolished for safety	reasons.	
11	This stopped them – for the good	reason	that I was not prepared to go on.
12	It's the main	reason	for choosing that restaurant.
13	We will concentrate on the	reason	that this ended up as it did.
14	The figures were changed for political	reasons.	
15	However, that is no	reason	to go back on the accord, he said.
16	He sees no	reason	to change the plan
17	For technical	reasons	it was necessary to stay in the classroom
18	The Iranians have no obvious	reason	to underestimate the number of refugees.
19	He stressed he was leaving for "family	reasons".	
20	"I think the	reason	that people like my work is because . . . "
21	One	reason	for his reluctance to do it is that . . .
22	She forgot all about his	reasons	for being there.
23	Her appearance was the major	reason	for the large turnout.

Fig 9.12 Selected concordances for the noun *reason*

Nouns in the database As they were for verbs, the database fields CONSTRUCTION and EXAMPLE are used to note the facts found in the concordances for *reason*. Figure 9.13 shows how this is done: here again, the figures in the left-hand column have been added for ease of reference, and for the same reason we highlight in bold the part of each example that instantiates the construction it is illustrating.

Box 9.3 Constructions for noun headwords

A list of lexicographically relevant co-constituents of noun headwords.

Code	Constituent	Examples
AJP	adjective phrase	a *happy* **man**, the **mayor** *elect*
AJ-pert	pertainym (adjective meaning 'pertaining to X', never predicative)	*marital* **bliss**
AVP-post-mod	adverb phrase as post-modifier of headword	the **journey** *home*
cl-if	whether/if clause	the **question** *whether he would go*
cl-that	indicative clause with *that*	their **knowledge** *that he had done it*, the **news** *that he had arrived*
cl-that-cond	conditional clause with *that*	your **wish** *that he were still alive*
cl-that-subj	subjunctive clause with *that*	their **request** *that he go with them*
cl-(that)	indicative clause without *that*	the **reason** *she went*
cl-wh	clause with *what*, *when*, *how*, *where*, *why*	the **reason** *why he left*, the **question** *when to go*
it+	anticipatory 'it' construction	*it*'s a **mistake** to think about it, *it*'s **fun** swimming in the sea
N-mod	headword modified by another noun	the *forgery* **allegations**, *cancer* **treatment**
N-premod	headword as pre-modifier of another noun	**journey** *time*, **road** *accident*
Part-*specific*-post-mod	named particle as post-modifier of headword	a **night** *out*, a **day** *off*
PP-*specific*	prepositional phrase with named preposition, e.g. PP-*at*, PP-*by*, etc.	after a **look** *at the screen*, a **refusal** *by my sister*, a **letter** *from home*, an **exchange** *with his partner*

(*cont.*)

Box 9.3 (*Continued*)

Code	Constituent	Examples
PP-*specific* cl-wh	prepositional phrase with named preposition with wh-clause	**questions** *about what online courses are offered*, **concerns** *about who to support*
PP-*for* Vinf-to	the *for* + infinitive-*to* construction	their **wish** *for him to be there*, her **anxiety** *for him to go*
PP-*specific* NP Ving	prepositional phrase with named preposition with noun phrase with gerund	the **thought** *of him going*
PP-*specific* Ving	prepositional phrase with named preposition with gerund as object	the **thought** *of going*
Supp-PP-*specific*	headword is object of named ('support') preposition	*on* **fire**, *in* **charge**, *over* **budget**
Vinf-to	infinitive verb with *to*	his **desire** *to be* present, her **need** *to behave* well

Here again all the examples are complete sentences drawn from the corpus. The various constructions shown in Figure 9.13 are highlighted in bold in the example sentences which follow each. They are fairly straightforward – only construction no. 11 requires a comment. It notes the use of a 'support preposition' with the noun headword. This term, from FrameNet, is explained in §3.4.1.1 of the FrameNet 'Book' pp. 54ff. (available for download from http://framenet.icsi.berkeley.edu/) and has been adapted for use in this way. Support prepositions often appear semi-arbitrary. Language learners must know which to use with any particular noun, and for this reason, these prepositions are an important aspect of a noun's behaviour, and must be recorded in the database. Some examples from the corpus of support prepositions and their nouns are highlighted in bold in Figure 9.14.

9.2.5.3 *Adjectives* This section covers lexically simple (single-word) adjectives and hyphenated compounds. Two-word compound adjectives such as *sky blue* and *stone cold* are considered as multiword expressions (cf. §9.2.6). Adjectives pose a number of problems for language learners and facts relating to these problems must be set out clearly in the database so that subsequent dictionary entries can include the solutions. The principal problems are:

1	CONSTRUCTION	**PP-for**
1a	EXAMPLE	The statement said NEC had claimed that 'the main reason **for Cray's failure** to win the contract' was its 'inability to meet technical requirements' set by the University Corporation of Atmospheric Research (UCAR), which made a tentative deal with NEC for NCAR.
1b	EXAMPLE	Until now the classifying officials were not always required to provide any reason **for their action** or even to identify themselves.
1c	EXAMPLE	But it is tempting to suspect that one reason for **Mr Major's reluctance** to shake out his cabinet is that he cannot think who to put there instead.
1d	EXAMPLE	Her appearance was a major reason **for the large turnout**.
2	CONSTRUCTION	**PP-for Vinf-to**
2a	EXAMPLE	This was unkind of Lydia, for Betty was not popular with men, which was another reason **for Lydia to dislike** her, since Lydia was one of those women who find something contaminating in ugliness and prefer to mingle only with those who are at least as attractive as themselves.
2b	EXAMPLE	He's had to be kept apart from the hamsters ever since one bit him on the nose – all the more reason **for Lorraine to look** for a bigger home.
2c	EXAMPLE	This principle holds also for joint activity with Israel, and therefore I see no reason **for Israel to be concerned**.
3	CONSTRUCTION	**PP-for Ving**
3a	EXAMPLE	We also had other reasons **for banning** meat imports.
3b	EXAMPLE	she forgot all about his reasons **for being** there.
3c	EXAMPLE	This is something that the British harp on about incessantly, using it as the main reason **for choosing** one restaurant over another, but do we really understand what it means and are we consistent in our assessment of perceived value?
4	CONSTRUCTION	**PP-for NP Ving**
4a	EXAMPLE	There are many reasons **for parties failing** to produce intended changes in outcomes.
4b	EXAMPLE	But there are other reasons **for farmers switching** to autumn cereals than a desire to curb nitrate pollution, and this is likely to be the limit of what the 'voluntary approach' can achieve.
4c	EXAMPLE	Nevertheless, this notion of the station as strong point continued, and was partly the reason **for many being built** outside the communities they served.
5	CONSTRUCTION	**Vinf-to**
5a	EXAMPLE	His estimate that 140,000 Iraqi Shias had sought refuge in Iran is exactly three times the figure given by the Iranian interior minister (although the Iranians have no obvious reason **to underestimate** the number of refugees).
5b	EXAMPLE	However, that is no reason **to go back** on the accord, he said.
5c	EXAMPLE	K-State sees nothing wrong with its plan and no reason **to change** it.
6	CONSTRUCTION	**PP-behind**
6a	EXAMPLE	Christian went on to explain the reasons **behind his momentous decision.**
6b	EXAMPLE	What, then, was the real reason **behind the decision** taken by the small coterie of Yeltsin's close associates, whom many Russians call the 'family politburo'?

(*cont.*)

Fig 9.13 Noun constructions recorded for *reason* in the database

7	CONSTRUCTION	**cl-that**
7a	EXAMPLE	I think the reason **that people are attracted to my work** is because I do not try to simply paint a portrait.
7b	EXAMPLE	So we will concentrate here on the two truly essential elements of the story: the decision to make a move against the Pike and the reason **that this move ended up taking the form that it did**.
8	CONSTRUCTION	**cl-wh**
8a	EXAMPLE	But, you know, there may be many other reasons **why it is important for him to get to the top so quickly**.
8b	EXAMPLE	This is another reason **why a well-advised employer who offers you an ex gratia payment is likely to insist that you are prevented in law from making any further claim against him**.
8c	EXAMPLE	For me, it made perfect sense and I thought 'there's no reason **why you can't have a normal job and have a fulfilling life as an artist as well**'.
9	CONSTRUCTION	**AJ-pert-premod**
9a	EXAMPLE	This rang alarm bells, for if, as was likely, the owner did not comply with this, the interior of the church, if not the facade, could be demolished for **secular** reasons.
9b	EXAMPLE	The figures had been changed for **political** reasons.
9c	EXAMPLE	No observer was present during the home observations but, for **technical** reasons, it was found necessary to have an observer in the classroom.
10	CONSTRUCTION	**N-premod**
10a	EXAMPLE	There were direct refusals to answer questions 'for **security** reasons', occasional resorts to lying, and frequent use of coded conversations.
10b	EXAMPLE	Of course, if alcohol is banned for **safety** reasons, dismissal may be the only option, when the circumstances have been fully investigated.
10c	EXAMPLE	He had been a member of the Cabinet since January 1981 [see p. 30708], and stressed that he was leaving for '**family** reasons' and that there was no disagreement between him and the Prime Minister.
11	CONSTRUCTION	**Supp-PP-for**
11a	EXAMPLE	This is happening **for** totally different reasons.
11b	EXAMPLE	The minute stopped the Cabinet committee dead in its tracks – **for** the good reason that I was not prepared to go on.

Fig 9.13 *Continued*

It had been **on loan** to the museum since 1960.
Houses and cars were **on fire** in the town.
He produced copies of documents **in his possession**.
He left London **for reasons** of health.
A troop of commando, with Lieutenant Smith **in command**.
Some people are more **at risk** than others.
The brigade had been **under attack** for four days.

Fig 9.14 Some support prepositions and their nouns

(1) Participial adjectives: e.g. *deserted*, *distressing*, *confused*. In the absence of any objective way of distinguishing in every case between verbal and adjectival use, the Style Guide for a database will usually advise you to treat these forms as full headword adjectives if they occur in the corpus modified by *very* or in other clearly adjectival uses (cf. §9.2.1).

(2) Gradability: does the adjective allow a comparative and superlative or not? If not, note that fact. If so, how are the comparative and superlative formed? The options in English are of course (a) irregular forms (*good*, *better*, *best*), (b) by adding *-er* and *-est* (*quick*, *quicker*, *quickest*) or (c) using *more* and *most*. Adjectives following (a) or (b) will have this noted in the INFLECTED FORM field (cf. §9.2.2).

➔ For the *more* and *most* group, try to include in the examples at the top of the LU entry at least one or two which show the headword in a comparative or superlative form.

(3) Pertainyms: many adjectives that are not gradable are pertainyms. This useful term comes from dictionary definitions beginning 'pertaining to X'. Pertainyms have no comparative or superlative forms, and cannot be used predicatively; e.g. *their economic policy* but not **a more economic policy* or **that policy is economic*.

➔ This should be recorded in the GRAMMAR field.

(4) Syntactic role: can the adjective be used both attributively (*sunny morning*) and predicatively (*the morning was sunny*), or is it attributive only (*a mere child*), or predicative only (*I want to be alone*), or is it a post-modifier (*mayor elect*)?

➔ This again should be recorded in the GRAMMAR field.

(5) Word order of pre-modifiers: this belongs in a grammar, not a dictionary database. Its position in a group of adjectives pre-modifying the same noun cannot be given in every adjective entry.

➔ For common adjectives it is good to include one or two instances of this phenomenon in the examples, simply because it is such a problem for language learners.

Adjective constructions The lexicographically relevant co-constituents of adjective headwords are given in Box 9.4. together with examples of each, where the exemplifying adjectives are in bold print. These co-constituents are the 'constructions' to be noted and recorded when your headword is an adjective.

Box 9.4 Constructions for adjective headwords

A list of lexicographically relevant co-constituents of adjective headwords.

Code	Constituent	Examples
AJP-premod	adjective phrase as pre-modifier	*pale* **green**, *light* **blue**
AVP-premod	adverb phrase as pre-modifier	*minimally* **cooperative**, *significantly* **different**
cl-if	whether/if clause	**uncertain** *whether I could go*, **unsure** *if he would be there*
cl-that	indicative clause with *that*	I was **happy** *that he would help*, I'm **sure** *that you will understand*
cl-that-subj	subjunctive clause with *that*	it's **important** *that he come*, **insistent** *that he join in*
cl-(that)	indicative clause without *that*	I was **happy** *he would help*, I'm **sure** *you will understand*
cl-wh	clause with *what*, *when*, *how*, *where*, *why*	**curious** *where he was*, **curious** *what to expect at the show*
N-premod	noun as pre-modifier	*heat* **sensitive**, *additive* **free**, *blood* **red**
PP-*for* Vinf-to	the *for* + infinitive-*to* construction	**possible** *for you to do it*
PP-*specific*	prepositional phrase with named preposition	**amazed** *at all this*, **happy** *with what he had*, **delighted** *for all of you*
PP-*specific* cl-wh	prepositional phrase with named preposition with wh-clause	**curious** *about where I can find that information*, **curious** *about what to expect during the event*
PP-*specific* NP Ving	prepositional phrase with named preposition with noun phrase with gerund	**aware** *of him laughing*
PP-*specific* Ving	prepositional phrase with named preposition with gerund	**tired** *of living*, **interested** *in knowing about it*
Vinf-to	verb phrase infinitive with *to*	**happy** *to know*, **eager** *to do it*
Ving	gerundive verb phrase	**busy** *repairing the radio*

Adjectives in the corpus The adjective *happy* will be our demo headword in this section. Figure 9.15 shows a set of selected concordances for this LU, edited in order to make it easier to find the construction in each.

Adjectives in the database As they were for verbs, the database fields CONSTRUCTION and EXAMPLE are used to note the facts found in the concordances for *happy*. Figure 9.16 shows how this is done. Here again, the figures in the left-hand column have been added for ease of reference, and for the same reason we highlight in bold the part of each example that instantiates the construction it is illustrating.

1	Some teachers are least	happy	about teaching poetry to this age group
2	Also I am not	happy	about the state of maintenance on vehicles.
3	He is demob	happy.	
4	He felt suddenly	happy	at being alive.
5	And she said she was	happy	doing it.
6	Natasha's aunt is absolutely	happy	for her to stay with us
7	He's	happy	for the opportunity.
8	We are really	happy	for you.
9	As long as the occupants are	happy	huddling in a spartan hut . . .
10	. . . stray bullets from firing by trigger	happy	individuals.
11	"I'm	happy	it's over," said Jon.
12	A sniper who could have been a gun	happy	soul anywhere had killed.
13	We're very	happy	that it's working.
14	They are exhausted but so	happy	to be free.
15	Many Italians will be	happy	to see the presidency pass to others.
16	I'm so	happy	to be here and playing for Lothar.
17	And they were perfectly	happy	to be left.
18	The rail staff have been quite	happy	to let me have a break in my journey.
19	Odd-Knut was not	happy	with the name.
20	It seems he's not too	happy	with your appearance.

Fig 9.15 Selected concordances for the adjective *happy*

1	CONSTRUCTION	**PP-*for***
1a	EXAMPLE	We are confident that you will now receive the joy you deserve, and we are really happy **for you**.
1b	EXAMPLE	I think the reality of self-sufficiency is better than that of dependency, so I was happy **for my son**.
1c	EXAMPLE	'He's happy **for the opportunity**,' Williams said.
2	CONSTRUCTION	**PP-*for* Vinf-to**
2a	EXAMPLE	You were quite happy **for me to come** home on the bus.
2b	EXAMPLE	Natasha's aunt is absolutely happy **for her to stay** with us and is desperately trying to get some information about her for me.
2c	EXAMPLE	He was quite happy **for Willi to sit** through some of Therese's practice sessions, for Willi's overwhelming admiration for Therese's voice was doing wonders.
3	CONSTRUCTION	**PP-*at***
3a	EXAMPLE	He felt suddenly happy **at** the prospect.
3b	EXAMPLE	Oh, so far from paying any extra for another television set, you are not happy **at** what we're paying at the moment?
3c	EXAMPLE	Dorothea stayed in the sunlit kitchen with her tea, happy **at the letter** – her loss of purpose, her anxiety dissipated like a past illness, already wondered at, the symptoms forgotten.
4	CONSTRUCTION	**PP-*about***
4a	EXAMPLE	Also I am not happy **about the state** of maintenance on vehicles.
4b	EXAMPLE	Evidence suggests that some teachers are least happy **about poetry classes for** this age group, in comparison with the other main literary genres.
4c	EXAMPLE	David wasn't very happy **about this**, but I think in a way it made him realise that I was really straightforward and that I really did care.

(*cont.*)

Fig 9.16 Adjective constructions recorded for *happy* in the database

5	CONSTRUCTION	**PP-*with***
5a	EXAMPLE	It seems he's not too happy **with your appearance**.
5b	EXAMPLE	"To me, this says one out of every three flight attendants are not happy **with the way** things are," she said.
5c	EXAMPLE	We called them huskies, as you might expect, but Odd-Knut was not happy **with the name**.
6	CONSTRUCTION	**cl-that**
6a	EXAMPLE	We're very happy **that it's working**.
6b	EXAMPLE	Dr. Briant wishes me to make clear at the outset that he is not entirely happy **that this matter should have become a subject of public discussion**.
6c	EXAMPLE	Yesterday was a peak day and the lads had gone home at the end of it worn down but happy **that the work was coming along at last**.
7	CONSTRUCTION	**cl-(that)**
7a	EXAMPLE	'I'm happy **it's over**,' said Jon, who dumped a bucket of confetti on Wilkens as the final seconds ticked away.
7b	EXAMPLE	'We are happy **we could help in this case**,' said Arriva's regional manager David Judson.
7c	EXAMPLE	It was the same with the dogs: Jim was happy **they had a real Irish wolfhound**, but it was left to Thomas to clear up after it.
8	CONSTRUCTION	**Vinf-to**
8a	EXAMPLE	I'm so happy **to be** here and playing for Lothar has been great.
8b	EXAMPLE	They are used to all sorts of emergencies, but there has never been anything like this: their own people, prepared to face appalling hardships, possible arrest or even death to get out, arriving exhausted but so happy **to be** free.
8c	EXAMPLE	Many Italians will be happy **to see** the presidency pass to others.
9	CONSTRUCTION	**Ving**
9a	EXAMPLE	And she said she was happy **doing** it.
9b	EXAMPLE	He says this was partly due to counsel and partly because she was happy **busying** herself with her domestic duties.
9c	EXAMPLE	Still, as long as the occupants are happy **huddling** in a spartan hut with the fantasy that they are men of the wilds, then who can criticise?
10	CONSTRUCTION	**N-premod**
10a	EXAMPLE	In two separate instances at least four persons were reported to have been killed in Peshawar and Karachi after being hit by stray bullets as a result of aerial firing by **trigger** happy individuals.
10b	EXAMPLE	In a flash a sniper, who could have been any **gun** happy soul anywhere, had killed.
10c	EXAMPLE	He is **demob** happy and there is something of the anonymity of the confessional in this dimly lit train compartment, lurching slowly over the snowy plateau of the Kola Peninsula.
11	CONSTRUCTION	**AVP-premod**
11a	EXAMPLE	The rail staff at Colchester have been **quite** happy to let me have a break in my journey.
11b	EXAMPLE	And they were **perfectly** happy to be left.
11c	EXAMPLE	We played to a standard I was **reasonably** happy with.

Fig 9.16 *Continued*

9.2.5.4 *Adverbs* This section covers lexically simple (single-word) adverbs: compound adverbs such as *at once*, *upside down*, and *all right* are multiword expressions (cf. §9.2.6). Adverbs pose some problems for language-learners and facts relating to these problems must be set out clearly in the database so that subsequent dictionary entries can include the solutions. The principal problems are:

1. Is the adverb gradable (with comparative and superlative forms) or not? If it is, how are the comparative and superlative formed? The options of course are (a) irregular forms (*well*, *better*, *best*), (b) by adding *-er* and *-est* (*fast*, *faster*, *fastest*), or (c) using *more* and *most*.
   ➔ Adverbs following (a) or (b) will have this noted in the INFLECTED FORM field (cf. §9.2.2). For the *more* and *most* group, it is helpful to include in the examples at the top of the LU entry at least one or two which show the headword in a comparative or superlative form.
2. Can the adverb be used predicatively (e.g. *he's home*, *they're upstairs*)?
   ➔ If your Style Guide doesn't give you a formal way of recording it, you should include one or two examples showing this use, together with a note in a COMMENT field.
3. Where should the adverb go in the clause or sentence?
   ➔ Although this is something that language-learners study in grammar lessons, if the headword position is variable, e.g. if it is a sentence adverb, then it is useful to include one or two examples showing the various options.

Box 9.5 Constructions for adverb headwords

A list of lexicographically relevant co-constituents of adverb headwords.

Code	Constituent	Examples
AJP	modifying an adjective phrase	**happily** *unaware*, **completely** *covered in mud*
AVP	modifying an adverb phrase	**very** *significantly*
CL	modifying a clause (or sentence)	**hopefully**, *it won't rain*
PP	modifying a PP	**directly** *into the net*
VP	modifying a verb	*ran* **fast**, *sang* **beautifully**

Adverb constructions The lexicographically relevant co-constituents of adverb headwords are given in Box 9.5. together with examples of each, where the exemplifying adverbs are in bold print. These co-constituents are

the 'constructions' to be noted and recorded when your headword is an adverb.

Adverbs in the corpus The adverb *seriously* will be our demo headword in this section. Figure 9.17 shows a set of selected concordances for this word,[6] which have been edited in order to make it easier to find the construction in each.

1	Jules added more	seriously,	"You may find yourself in a difficult position."
2	Although I hurt you very	seriously,	I pray that you won't destroy my life
3	Let's try to think	seriously	about this matter.
4	Japan has promised to	seriously	consider the government's demand.
5	A man was	seriously	hurt when a tree fell on him.
6	Something was going	seriously	wrong.
7	. . . the lad whose brother was	seriously	hurt.
8	. . . dispel rumours that he is	seriously	ill or dying.
9	It was already	seriously	over budget.
10	It started to go	seriously	downhill in the 1960s.

Fig 9.17 Selected concordances for the adverb *seriously*

Adverbs in the database As they were for verbs, the database fields CONSTRUCTION and EXAMPLE are used to note the facts found in the concordances for *seriously*. Figure 9.18 shows how this is done: here again,

1	CONSTRUCTION	**AJP**
1a	EXAMPLE	Oh, you know, the lad whose dad was convicted of drunk driving and whose little brother was seriously **hurt**.
1b	EXAMPLE	However, it's not unusual for family members to comment on Deng's health to dispel rumours that he is seriously **ill** or dying.
1c	EXAMPLE	A civil defence worker in the central city of Matanzas said a man was seriously **hurt** there when a tree fell on him.
2	CONSTRUCTION	**VP**
2a	EXAMPLE	Let's try **to think** seriously about this matter.
2b	EXAMPLE	Jules laughed, but then **added** more seriously, "Nevertheless, you may find yourself in a difficult position before much longer, Alice."
2c	EXAMPLE	Although I **hurt** you very seriously, I pray to you that you won't destroy my life
3	CONSTRUCTION	**AVP**
3a	EXAMPLE	Though the area itself was always somewhat seedy, it started to go seriously **downhill** in the 1960s, struck by urban blight.
3b	EXAMPLE	And it may be, for instance, erm that he may even have to intervene at the modification stage if something was going seriously **wrong**.
4	CONSTRUCTION	**PP**
4a	EXAMPLE	It was already seriously **over budget**.

Fig 9.18 Adverb constructions recorded for *seriously* in the database

[6] This adverb will break down into at least two LUs (manner: *he added seriously*, and degree: *seriously ill*); both LUs are represented in these concordances.

the figures in the left-hand column have been added for ease of reference, and for the same reason we highlight in bold the part of each example that instantiates the construction it is illustrating.

9.2.5.5 *Identifying and recording complements for one headword: a case study* Sometimes it's difficult to decide how to record the constructions associated with a headword when they appear in the corpus in varying configurations, and this is where a knowledge of the principles of frame semantics can be very useful. Let's take for example the case of the verb *cook*: Figure 9.19 shows some selected and abridged corpus lines to serve as a little case study on using complements in sense analysis, and in recording them, including those which at first sight appear to be 'optional'.

1	I	cook	for my family every night
2	Olga bought steak and he	cooked	it.
3	She made fresh coffee and	cooked	me a man-sized breakfast
4	We have a girl who comes in and	cooks	lunch during the week.
5	Come round to my flat and I'll	cook	you a meal.
6	I wish John would	cook	every meal for me.
7		Cook	the fish in salted water until tender.
8	She can wash,	cook,	iron and sew.
9	Flip the pancake over and	cook	for another 30 seconds.
10	I've	cooked	him dinner and he doesn't want it.
11	The youngest daughter-in-law has	cooked	the meal.
12	Add onions and	cook	until they begin to soften.
13	Mary likes to	cook	for her guests.
14	The toast is thin and will	cook	very quickly.
15	What had they done while this chicken was	cooking	?
16	They promised they would come in and	cook	lunches for visitors.

Fig 9.19 Some concordances for the verb *cook*

Analysing the corpus lines From the frame-semantics-lite perspective, there are two principal frame elements that interest us in our initial approach to this verb:

- COOK: the person cooking
- FOOD: the food being cooked.

It's easy to identify the phrases in the concordance lines which instantiate these central[7] frame elements, and these are labelled in the table in Figure 9.20.

[7] We use 'central' rather than 'core' here, since in FrameNet terms the 'core frame elements' for the verb *cook* are those expressing the person doing the cooking, and the object being cooked.

#	Example
1	[COOK **I**] cook for my family every night
2	Olga bought steak and [COOK **he**] cooked [FOOD **it**].
3	[COOK **She**] made fresh coffee and cooked me [FOOD **a man-sized breakfast**].
4	We have a girl [COOK **who**] comes in and cooks [FOOD **lunch**] during the week.
5	Come round to my flat and [COOK **I**]'ll cook you [FOOD **a meal**].
6	I wish [COOK **John**] could cook [FOOD **every meal**] for me.
7	Cook [FOOD **the fish**] in salted water until tender.
8	[COOK **She**] can wash, cook, iron and sew.
9	Flip the pancake over and cook for another 30 seconds.
10	[COOK **I**]'ve cooked him [FOOD **dinner**] and he doesn't want it.
11	[COOK **The youngest daughter-in-law**] has cooked [FOOD **the meal**].
12	Add onions and cook until they begin to soften.
13	[COOK **Mary**] likes to cook for her guests.
14	[FOOD **The toast**] is thin and will cook very quickly.
15	What had they done while [FOOD **this chicken**] was cooking?
16	They promised [COOK **they**] would come in and cook [FOOD **lunches**] for visitors.

Fig 9.20 Frame elements COOK and FOOD instantiated in the *cook* concordances

When we study these lines in the light of the two frame elements, we begin to notice some apparent anomalies:

- Neither of these 'central' frame elements is expressed in lines 9 and 12.
- The frame element FOOD is not expressed in lines 1, 8 or 13.
- The frame element COOK is not expressed in lines 14 or 15.
- The frame element COOK is the subject of the verb in the lines where it is expressed, except for lines 14 and 15, where it does not appear at all.
- The frame element FOOD is the object of the verb in the lines where it is expressed, except for lines 14 and 15, where it is the subject.
- Sometimes the verb is ditransitive (as in lines 5 and 10).

These anomalies make us think there must be more than one sense of the headword *cook* (i.e. more than one LU) in these concordance lines. We could distinguish three in all:

LU-1 Of a person, heat (a raw food substance) in order to change it into a more edible form (e.g. lines 2, 7, 9, and 12). In this LU, when they are expressed, the person cooking (COOK) is the subject of the verb, and the food substance (FOOD) is the object.

LU-2 Of a raw food substance, change during the heating process into a more edible form (e.g. lines 14 and 15). Here the food substance is the subject of the verb, and the cook is not expressed.

LU-3 Of a person, prepare (a meal) by cooking food substances. Here both the cook and the meal may be expressed, but are not always made explicit.

This 'splitting' into three LUs of the broadbrush sense of *cook* makes it clear that the two original 'central' frame elements are not enough to account for the various usages seen in the concordances. We can now posit three central frame elements, i.e.

- COOK: the person cooking (as in LUs 1 and 3)
- FOOD: the raw food being cooked (as in LUs 1 and 2)
- MEAL: the meal or a dish being prepared (as in LU-3).

To these must be added a fourth, in order to account for the ditransitive uses in lines 5 and 10:

- RECIPIENT: someone for whom the meal is being prepared (as in LU-3)

These four frame elements represent the central semantic roles[8] instantiated in the contexts of *cook*, and our dictionary entry must be able to specify how they are variously expressed, and also which, if any, of them may be omitted, and the circumstances under which this can happen. When the *cook* concordances are split into the three LUs, the four frame elements are variously expressed by the phrases highlighted in Figure 9.21.

The way in which the four central frame elements are realized in the three LUs of the headword *cook* highlights the 'valency patterns'[9] which make up the 'valency description' of each of the three LUs is as shown in Figure 9.22.

Identifying the constructions to be recorded In the foregoing section we saw how analysing concordances in frame semantics terms gave the valency patterns for each LU of the headword. From there it is easy to get to a list of the essential complementation patterns, or 'constructions' in our terminology, for each LU. When you're recording a verb's complementation in the database, you don't of course include the subject of the verb. You do, however, include a 'zero' complement, where none is found in the context

[8] The RECIPIENT semantic role has of course a larger scope in the language than simply as an element in any single frame (cf. *cage me a peacock*); however, the benefactive NP NP (indirect and direct objects) construction is a central element in lexicographic analysis, and must be recorded wherever it occurs.

[9] The terms *valency pattern* and *valency description* are explained in §5.4.3

2 Olga bought steak and [COOK **he**] cooked [FOOD **it**].
7 Cook [FOOD **the fish**] in salted water until tender.
9 Flip the pancake over and cook for another 30 seconds.
12 Add onions and cook until they begin to soften.
LU-1 : (of person) heat (raw foodstuff) in order to make it more edible

14 [FOOD **The toast**] is thin and will cook very quickly.
15 What had they done while [FOOD **this chicken**] was cooking?
LU-2 : (of food) change during heating process

1 [COOK **I**] cook [RECIPIENT **for my family**] every night
3 [COOK **She**] made fresh coffee and cooked [RECIPIENT **me**] [MEAL **a man-sized breakfast**].
4 We have a girl [COOK **who**] comes in and cooks [MEAL **lunch**] during the week.
5 Come round to my flat and [COOK **I**]'ll cook [RECIPIENT **you**] [MEAL **a meal**].
6 I wish [COOK **John**] could cook [MEAL **every meal**] [RECIPIENT **for me**].
8 [COOK **She**] can wash, cook, iron and sew.
10 [COOK **I**]'ve cooked [RECIPIENT **him**] [MEAL **dinner**] and he doesn't want it.
11 [COOK **The youngest daughter-in-law**] has cooked [MEAL **the meal**].
13 [COOK **Mary**] likes to cook [RECIPIENT **for her guests**].
16 They promised [COOK **they**] would come in and cook [MEAL **lunches**] [RECIPIENT **for visitors**].
LU-3 : (of person) prepare (a meal or dish)

Fig 9.21 Expression of central frame elements in the LUs of the verb *cook*

(examples 2, 7) COOK/NP/subject FOOD/NP/object
LU-1 : (of person) heat (raw foodstuff) etc.

(examples 14, 15) FOOD/NP/subject
LU-2 : (of food) change during heating

(example 8) COOK/NP/subject
(examples 4, 11) COOK/NP/subject MEAL/NP/object
(examples 3, 5, 10) COOK/NP/subject RECIPIENT/NP/complement MEAL/NP/object
(examples 6, 16) COOK/NP/subject MEAL/NP/object RECIPIENT/PP-for/complement
(examples 1, 13) COOK/NP/subject RECIPIENT/PP-for/complement
LU-3 : (of person) prepare (a meal or dish)

Fig 9.22 Valency descriptions of the three LUs of the verb *cook*

of the headword. Figure 9.23 shows the constructions to be noted in the database for each of the three LUs of the verb *cook*.

LU-1 : heat raw foodstuff etc.
constructions 0 (*zero*) examples 9, 12
NP examples 2, 7

LU-2 : (of food) change etc.
construction 0 (*zero*) examples 14, 15

LU-3 : prepare a meal etc.

constructions			
0 (*zero*)			(example 8)
NP			(examples 4, 11)
NP	NP		(examples 3, 5, 10)
NP	PP-for		(examples 6, 16)
PP-for			(examples 1, 13)

Fig 9.23 Constructions associated with each of the LUs of the verb *cook*

Recording complementation in the database It isn't too difficult to record in the CONSTRUCTION field the simple constructions associated with the first two LUs of the verb *cook*, together with examples.[10] This is done in the usual way, and the results are shown in Figures 9.24 and 9.25. Note that, while the absence of any complementation is recorded as a zero, not all zeros are the same, some being more zero than others (cf. Box 9.6).

Of this verb, the LU with the most complicated set of constructions is clearly the third. Any attempt to collapse these into a statement showing options by means of bracketing is guaranteed to fail; all of the complements recorded are 'optional', in the sense that the verb can be used with none, as in *he can't cook*. However, a description of the complementation such as

[(NP) (NP) (PP-for)]

is not accurate, since there is no instance of

[NP NP PP-for]

in the corpus; nor can we accept such a usage as possible. Therefore, the only way to record these constructions in the database is to record each *pattern* individually. The result is the database record of the constructions associated with LU-3 of *cook* that appears in Figure 9.26.

➔ When it comes to collecting corpus examples in the database, more is definitely better. When in doubt, don't leave it out.

Omission of the direct object: 'null instantiation' This phenomenon (described in §3.2.3 of the online FrameNet manual at http://framenet.icsi.berkeley.edu/) occurs when a 'core frame element' (a FrameNet term) is not expressed; such an omission is always conceptually salient. (In some cases more than one frame element may be missing from a single context, but that should not concern us here.) Box 9.6 holds a brief outline of the three types of null instantiation. Although the theory allows for various grammatical relationships between the headword and the missing item, the most useful one for lexicographers is that of the omission of the direct object of a transitive verb. The verb *cook* can also be used as an illustration of this phenomenon.

[10] Note that when the CONSTRUCTION field appears at the top of the entry, before any examples, this means that it applies to the whole LU (i.e. it is an obligatory complement of the headword in that sense). This is illustrated in Figure 9.45, where its position shows that *consign* requires the prepositional phrase with *to*.

Box 9.6 Null Instantiation

Frame semantics recognizes three types of null instantiation.

Constructional Null Instantiation (CNI)

This is the easiest to understand, and has no place in conventional lexicography, since the omission is part of the grammar of the language. CNI is to be found in all imperative uses of verbs, where it is a normal feature of the language: for instance, in sentences like *Go home!* or *Cook the chicken thoroughly*, the subject of the verb is omitted.

The other two types of null instantiation are of much more interest to us, and should be recorded in a lexicographic database, as they are not part of the grammar of the language. They are facts about a word that need to be known if it is to be used correctly or fully understood.

Indefinite Null Instantiation (INI)

In the sentence *Can you knit?* the verb *knit* looks intransitive, and is often treated as such in dictionaries, although its sense is not intransitive, and it may need to be translated by a transitive verb with a non-specific object. In the case of *Can you knit?* the verb's direct object is not expressed, and we do not need to know what might be knitted in order to understand the sentence. We can make the 'general' sense more explicit by inserting the non-specific pronoun *anything* as the object of the verb: *Can you knit anything?* This omission of a central semantic role (a 'core frame element' in frame semantics terms) leading to a 'general' interpretation is a case of INI. Another example of INI is the omission of the core frame element TOPIC in the 'quarrelling' sense of *argue* (cf. §5.5.2.1), as in the sentence *Stop arguing!* There is no need to know what is being argued about in order to understand that sentence. Here again, we can make it more explicit by inserting a non-specific pronoun as the topic, e.g. *Stop arguing about everything!* A lexicographer writing a dictionary entry for *argue* will want to note that, while the INI-type omission of the TOPIC (expressed as PP-about) is a property of the verb in the 'quarrelling' sense, it does not operate in the 'reasoning' sense. Instead of *They were arguing for a revision of the agreement* you cannot say **They were arguing* (omitting the PP-for). A dictionary entry for *argue* should make that clear. The fact of the INI must be recorded in the database.

Definite Null Instantiation (DNI)

DNI may be exemplified by the behaviour of the verb *blame*. When I say *I blame John*, then both you and I know what I blame John for. Without that

Box 9.6 (*Continued*)

knowledge, you could not understand the sentence (and indeed I wouldn't have formulated it). When I want to make the reason for the blame more explicit, the non-specific pronoun will not fit, cf. **I blame John for something.* A specific reason must be given: *I blame John for all the problems*, or *I blame John for that.* Here the omission of the core frame element REASON (expressed as PP-for) is an instance of DNI. Note that only the REASON may be omitted. You can't, say, leave out the person being blamed and say **I blame for that* or **I blame that.* A dictionary entry for *blame* should make that clear. The fact of the DNI must be recorded in the database.

Although the database record of the constructions associated with *cook* (shown in Figures 9.24, 9.25, and 9.26) is detailed and complex, it still does not capture everything we know about this sense of the verb. Yet to be recorded and explained is the apparent omission of the direct object, and in particular the fact that there are subtle differences in our interpretation of this missing object. Two of the senses of cook, LU-1 and LU-3, offer a nice case study of this phenomenon. LU-2, where the subject of cook is the actual food, cannot be used transitively and so cannot be an instance of direct object omission. Both LU-1 and LU-3 can, however: they are both essentially transitive in meaning, expressing as they do the idea of someone cooking either a piece of food (LU-1), or a meal (LU-3). Both appear in contexts where the object is not expressed:

1	CONSTRUCTION	**0** (*zero*)
1a	EXAMPLE	Flip the pancake over and cook for another 30 seconds.
1b	EXAMPLE	Add onions and cook until they begin to soften.
2	CONSTRUCTION	**NP**
2a	EXAMPLE	Olga bought steak and he cooked **it.**
2b	EXAMPLE	Cook **the fish** in salted water until tender.

Fig 9.24 The constructions recorded in the database for *cook* LU-1

1	CONSTRUCTION	**0** (*zero*)
1a	EXAMPLE	The toast is thin and will cook very quickly.
1b	EXAMPLE	What had they done while this chicken was cooking?

Fig 9.25 The constructions recorded in the database for *cook* LU-2

1	CONSTRUCTION	**0** (*zero*)
1a	EXAMPLE	She can wash, cook, iron and sew.
1b	EXAMPLE	And have you forgotten you're down to cook tonight?
2	CONSTRUCTION	**NP**
2a	EXAMPLE	We have a girl who comes in and cooks **lunch** during the week.
2b	EXAMPLE	The youngest daughter-in-law has cooked **the meal**.
3	CONSTRUCTION	**NP NP**
3a	EXAMPLE	Come round to my flat and I'll cook **you a meal**.
3b	EXAMPLE	I've cooked **him dinner** and he doesn't want it.
4	CONSTRUCTION	**NP PP-for**
4a	EXAMPLE	I wish John could cook **every meal for me**.
4b	EXAMPLE	They said they would come in and cook **lunches for visitors**.
5	CONSTRUCTION	**PP-for**
5a	EXAMPLE	I cook **for my family** every night.
5b	EXAMPLE	Mary likes to cook **for her guests**.

Fig 9.26 The constructions recorded in the database for *cook* LU-3

(1) *add onions and cook until they begin to soften* (LU-1)
(2) *have you forgotten you're down to cook tonight?* (LU-3).

Yet if you wanted to express the missing objects in these cases you would have to say

(1) *add onions and cook them until they begin to soften*
(2) *have you forgotten you're down to cook something tonight?*

From these two examples we see that in the case of (1) the missing object of the verb is specific, while in the case of (2) it is non-specific. Both of these examples illustrate the phenomenon of 'null instantiation', where a frame element is missing but understood in the context. Example (1) is an illustration of 'definite null instantiation' (DNI) and example (2) illustrates 'indefinite null instantiation' (INI): these terms are more fully explained in Box 9.6. INI is quite a common occurrence in English, and is associated with certain semantic classes of verbs such as verbs of creation (e.g. *cook*, *sew*, *bake*, *make*) or verbs of ingesting (e.g. *eat*, *drink*, *chew*, *swallow*). DNI is less frequent, and is linked to a specific word, not a class of words. Some more examples of both INI and DNI are given in Figure 9.27, where the headword verb is in bold type. In this table the instances of INI (examples 1–3 inclusive) are all readily understood without further context: many other verbs in English behave like *sew*, *eat*, and *drink*. The instances of DNI, however, cannot be understood unless it is already known what is referred to.

#	Example	NI type	Missing expression
1	She taught her granddaughter how to **sew**.	INI	NP (*anything that can be sewn*)
2	'Have you **eaten**?' said Lucy.	INI	NP (*anything*)
3	Their crews were . . . **drinking** from mugs and watching the passing scene.	INI	NP (*any kind of drinkable liquid*)
4	Bernard neither smoked nor **drank**.	DNI	NP (*alcohol*)
5	She made clothes, **baked**, brewed beer . . .	DNI	NP (*bread, cakes etc.*)
6	'Why wasn't I **told**?' he grumbled peevishly.	DNI	PP-about (*about that*)
7	'He wouldn't **dare**,' she said angrily.	DNI	VP (*to do what had been mentioned*)
8	"Help me up and I'll **try**," she said.	DNI	VP (*to do what had been mentioned*)

Fig 9.27 Some examples of indefinite and definite null instantiation

Box 9.7 Discussion points regarding examples in Figure 9.27

- Note the difference in interpretation of the uses of *drink* without an object in examples 3 (non-specific: INI) and 4 (specific: DNI).
- The verb *bake* in example 5 is normally transitive, and when used without its object has also a very specific meaning (when you read that sentence you don't think of her baking potatoes or meat, but rather bread or cakes).
- Dictionaries often treat the uses shown in examples 4 and 5 as separate LUs of *drink* and *bake*.
- Examples 6 to 8 inclusive are there to show that DNI can refer to the omission of other essential semantic roles ('core frame elements') as well as the objects of transitive verbs.
- In the case of the verb *tell* (example 6) the missing item is what he should have been told about, and had it been expressed it would have been in the form of a prepositional phrase with *about*. (The omission here of the 'person telling' is an instance of CNI, licensed by the grammar of the passive form in English.)
- In examples 7 and 8, the infinitive complements of the verbs *dare* and *try* are omitted: these would express the actions previously referred to.

To summarize: although a common phenomenon in English, null instantiation rarely appears in our dictionaries. And yet INI and DNI are often significant features of a word's behaviour with real value to the dictionary user. The omission of the object of a verb, leading to many contexts in which the verb is apparently intransitive, is of particular interest to the language-learner, and the interpretation of such contexts is of value to applications such as machine-assisted translation and information retrieval.

➔ Because this point has great potential value for learners' dictionaries, it's important to get it right in the database.

Null instantiation in the database The place to record INI and DNI in the database is together with the constructions that trigger their interpretation, since null instantiation is the non-expression in grammatical and lexical terms of a core semantic role. Thus, for the verb *cook*, in the case of two of the three 'intransitive' uses shown in Figures 9.24, 9.25, and 9.26, the CONSTRUCTION part of the entry would be expanded to include INI and DNI information, as shown in Figure 9.28.

LU-1 : heat raw foodstuff etc.		
1	CONSTRUCTION	0 (*zero*)
1a	EXAMPLE	Flip the pancake over and cook for another 30 seconds.
1b	EXAMPLE	Add onions and cook until they begin to soften.
1c	NULLINST-TYPE	DNI
1d	NULLINST-SEMANTICS	omitted = what is to be cooked
1e	NULLINST-SYNTAX	NP
LU-2 : (of food) change through heating etc.		
1	CONSTRUCTION	0 (*zero*)
1a	EXAMPLE	The toast is thin and will cook very quickly.
1b	EXAMPLE	What had they done while this chicken was cooking?
LU-3 : prepare a meal etc.		
1	CONSTRUCTION	0 (*zero*)
1a	EXAMPLE	She can wash, cook, iron and sew.
1b	EXAMPLE	And have you forgotten you're down to cook tonight?
1c	NULLINST-TYPE	INI
1d	NULLINST-SEMANTICS	omitted = 'anything'
1e	NULLINST-SYNTAX	NP

Fig 9.28 Null instantiations recorded for intransitive uses of *cook*

As is shown in Figure 9.28, null instantiations must be noted immediately *after* the relevant construction and its examples. Three facts should be recorded:

- whether it is INI or DNI (in the field NULLINST-TYPE)
- what exactly is omitted from the semantics (in NULLINST-SEMANTICS)
- what exactly is omitted from the syntactic context (in NULLINST-SYNTAX).

The first and the last of these facts are noted in terms of database codes (usually to be selected from pull-down lists in the dictionary writing system); the second, however, is an informal description which will satisfy

editors using the database, although in this form it is not machine-readable.[11] The thinking behind these records is as follows:

LU-1 This type of DNI is to be found principally in cookery books and other instruction manuals (e.g. *insert screw and tighten*; *loosen cap and remove*) and is known as the 'instructional imperative'. It is always worth recording this usage in the database.

LU-2 This is a genuine intransitive use of *cook*, and so there is no null instantiation to be recorded.

LU-3 This type of object-omission is quite common, but worth recording in the database whenever it is found. It is sometimes known as the 'absolute' use of a transitive verb.

There are several ways of dealing with the various types of null instantiation in the dictionary proper, and the Style Guide must give clear instructions on this point.

➔ Try to note both INI and DNI systematically in the database. This is something that language learners cannot know unless dictionaries tell them about it. Without this knowledge, they can neither use the word flexibly nor understand it in all of its contexts.

9.2.6 Multiword Expressions (MWEs)

In an English database it's often helpful to distinguish five types of MWE, each of which will have its own input style. The main types of MWEs found in English are discussed in some detail in §6.2.2, and taken up again in the context of dictionary entry components in §7.2.7.1. They are:

- idioms
- collocations
- phrasal verbs
- compounds
- support verb constructions.

9.2.6.1 *Finding MWEs in the corpus* One of the principal problems in database- and dictionary-editing is to decide how to define, for the Style Guide, a multiword expression (MWE), and how to distinguish one type

[11] In the FrameNet database, however, this is expressed in terms of the actual missing frame element.

from another.[12] This has to be done in some way if MWEs are to be handled systematically, but no one has yet produced a set of watertight criteria to apply as a means of identifying the various types and handling them systematically. This is a problem which greatly exercises dictionary editors, but fortunately does not seem to worry human dictionary users, although it can cause difficulties when the dictionaries are used as input to computer lexicons.

Of the MWEs discussed in §6.2.2, 'idioms' and 'collocations' present particular problems. It's easy to distinguish them from the other MWEs, but less easy to decide which is which when you want to record them in the database. They lie along a gradient, or cline, and distinguishing between them can be so difficult that many Style Guides don't even expect you to try. However, sometimes you are asked to make this distinction and you have to do your best. (There's no absolute right and wrong here.) The phrases at the 'upper' end of the gradient are those which present no problems of identification:[13] phrases like *to bite someone's head off* (snap at them angrily) and *to give something a clean bill of health* (report favourably on something after examination) are quite clearly 'idioms'. In both cases the following is true:

(1) The MWE is a fixed or semi-fixed group of words.
(2) Its meaning is more than the sum of its parts.
(3) It is complex enough to need as its companions a number of corpus examples and perhaps also some grammatical information or other facts.
(4) It would be out of place in the database entry if it were to be inserted into one of the senses of the headword (whether that headword is *bite* or *head*, or *clean* or *bill* or *health*).

Admittedly, that list of 'properties of phrases to be considered idioms' is hardly scientific or objective, but as a rule of thumb it has worked for many lexicographic teams. Most idioms are clearly visible in concordances sorted alphabetically on either left or right context of the keyword (as in Figure 9.29), while others are less salient (as in Figure 9.30).

[12] This is often called 'Phraseology' in linguistics and lexicographic literature. See Cowie 1999a for a brief introduction to the lexicographer's problems.

[13] So much so that these and only these are used to illustrate the many many papers on idiom by theoretical linguists, who are single-handedly keeping alive old favourites like *to rain cats and dogs* and *to kick the bucket*. It is a very long time since either of us heard these in day-to-day discourse.

He was . . . trained as a craftsman, and quite a handsome chap **into the**	**bargain**.
British Gas . . . have provided enough bottled gas for two years **into the**	**bargain**.
'You certainly make me feel so,' he said 'And stupid **into the**	**bargain**'.
Not only shelve the files in question, but pay him a monthly stipend **into the**	**bargain**.
There are possibilities for a lot of enjoyment to be had **into the**	**bargain**.
. . . just about every colour of foliage and leaf shape, and most are evergreen **into the**	**bargain**.
Hard work, ample food and a neat change of clothes **into the**	**bargain**.

Fig 9.29 Some corpus sentences for *into the bargain*

Which came first, the	chicken	or the egg?
I never loved others because I never loved myself – that's a	chicken	and an egg!
They concede the	chicken	and egg possibility – do you smack a child because he is delinquent, or is he delinquent because he is smacked?
The person bemoaning the high incidence of mental illness among the unemployed might wonder if he himself is the	chicken	or the egg
'Long or short term:	chicken	or egg?'

Fig 9.30 Some corpus sentences for the *chicken and egg* idiom

The semantically transparent phrases at the other end of the gradient (usually very easy to spot in corpus data) are quite clearly 'collocations' rather than 'idioms' – for instance *better luck next time!* or *from hour to hour*. Of the list of idiom properties above, only (1) applies to these phrases. But we're dealing with groups of words that occur significantly frequently in a corpus of modern English, although their meaning is transparent, and this phenomenon must be recorded if the database is to be comprehensive. They are particularly useful for dictionaries (mono- or bilingual) for language-learners, who can understand them easily, but need to learn to use them fluently. Moreover, more often than not these collocations will generate an idiomatic translation in a bilingual dictionary.

➔ If in doubt about the status of a phrase, enter it into the database as a collocation. That way, it won't disappear among the standard-usage example sentences, and editors of a bilingual dictionary will have their attention drawn to the particular context in which the headword is used.

9.2.6.2 *Recording MWEs in the database* When you want to enter an MWE in the database or dictionary the first thing to do is to decide on

its canonical form. This is the most basic, the most 'unmarked' form, the one in which it is 'declared' in a dictionary, and the most natural way to refer to it in conversation or writing (as for instance in *Do you know what* '*bite someone's head off' means?*).

What to enter: canonical forms The canonical forms of the words of the language give us the headwords of a dictionary, for example the singular of nouns, the infinitive without 'to' of verbs, and so on. Some nouns (*pyjamas*, *trousers*) have no singular; some verbs (*may*, *can*) have no infinitive form, and the Style Guide must of course give guidance on these. A very few idioms have no canonical form, and here again you'll need some help from the Style Guide on how to handle them. A well-known instance is the 'chicken and egg' idiom illustrated in the corpus sentences in Figure 9.30, which nicely show its meaning. While the first sentence is clearly the source of the idiom, and necessary in order to explain the uses that follow, you can't really use this as the canonical form, because as well as *chicken and egg* the idiom appears in the corpus (and could be looked up by a user) as *chicken or egg*, *a chicken and an egg*, and *the chicken or the egg*.

Why didn't I	hold my tongue	
Stella was forced to	hold her tongue	when Dotty spoke . . .
All the way home she	held her tongue,	answering in monosyllables.
we shall watch no longer, nor	hold our tongues	for fear of hurting you.
	Hold your tongue,	woman!
Speakers should not	hold their tongues	for fear of writs.
He'd learned a great deal about	holding his tongue,	even under injustice.

***to hold one's tongue* = to say nothing although you want to speak**

I think she's	pulling my leg,	so I ask her again.
It depends on who's	pulling your leg.	
She had thought that he was	pulling her leg.	
I said I was only	pulling your leg.	
Of course they started	pulling his leg	then.
'Get along out,' she said, 'and don't be	pulling our legs!'	
We all know you're	pulling their leg	the whole time.

***to pull someone's leg* = to tell someone something that is not true, as a joke**

Fig 9.31 Two idioms requiring different possessives in the canonical form

Many idioms like those in Figure 9.31 contain a possessive, and in a surprising number of dictionaries they appear as *to hold someone's tongue*, or *to pull one's leg.* This implies that somebody else can hold your tongue, or you can pull your own leg, both rather unlikely scenarios. The correct

canonical forms are of course *to hold one's tongue* and *to pull someone's leg.*[14]

➔ When you are entering the canonical form of an idiom with a possessive remember the rule of thumb: *to hold one's (own) tongue*, and *to pull someone (else)'s leg*.

How to enter the MWE Once you've identified an MWE in the corpus, you have to decide how to enter it into the database. The data fields MWE-TYPE and MWE are used here. What you do with an MWE in the database is not necessarily what will eventually be done with it in the dictionary proper. Both database and dictionary Style Guides should give clear guidance about the status of the MWE within an entry: whether to make it into a separate LU (subentry) always, never, or only in specific circumstances.

➔ If you're writing the Style Guide for database or dictionary, make sure that the various types of MWE are correctly identified, and systematically recorded.

When it comes to deciding where to enter MWEs in database or dictionary, there are a number of options for English. The principal are given below.

Idioms and collocations From the wide choice here, you have (or the Style Guide has) to choose one of the following:

(1) Enter the MWE under the first or only lexical (not grammatical) word, i.e. *into the bargain* in the *bargain* entry; *to be hot on something* in the *hot* entry; *to pull someone's leg* in the *pull* entry.
(2) Enter it under the least frequent lexical word, the one expected to have the shortest dictionary entry, i.e. *to open the floodgates* at *floodgates*.
(3) Enter it under the first or only noun in the phrase, i.e. *to rain cats and dogs* in the *cat* entry; *to pull someone's leg* in the *leg* entry; *big deal* in the *deal* entry.
(4) Enter it under the first or only verb in the phrase, i.e. *to rain cats and dogs* in the *rain* entry; *to twist and turn* in the *twist* entry.
(5) Enter it as a headword in its own right, i.e. individual main entries for *into the bargain*, *be hot on*, *pull someone's leg*, *rain cats and dogs*, *big deal*, etc.

[14] Learners' dictionaries generally distinguish between the two options by using *your* and *someone's*, but this style is rather chatty for most other dictionaries.

Of these alternatives,[15] (5) is the least likely, especially in an electronic database, where any MWE can be found automatically wherever it is entered, as long as it is tagged correctly. It's also rarely chosen as the style in general language dictionaries, but of course in dictionaries of idioms this is the standard headword form. Examples are given in Figures 9.32 and 9.33 of two different ways of entering idioms into the database; the methods they illustrate will work for collocations too.

HEADWORD	**bargain**
LU #	**1**
WORDCLASS	noun
MEANING	an agreement between people to do certain things
EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
MWE	**into the bargain**
MWE-TYPE	**idiom**
MEANING	in addition, on top of everything else
EXAMPLE	He must have been an eligible enough bachelor: the son of a fairly prosperous artisan family, trained as a craftsman, and quite a handsome chap into the bargain.
EXAMPLE	In consideration of this, not only would they shelve the files in question, but pay him a monthly stipend into the bargain.
HEADWORD	**bargain**
LU #	**2**
WORDCLASS	noun
MEANING	something bought for less than usual price
EXAMPLE	That second-hand adjustable table was a real bargain.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**3**
WORDCLASS	verb
MEANING	negotiate the terms/conditions of something
EXAMPLE	Buyers will bargain hard to cut the cost of the house they want, but dig in their heels rather than reduce their own asking price.
(*etc. etc.*)	

Fig 9.32 Method 1: the idiom within an LU

Compounds Here the principal options are:

(1) Enter the MWE under the first element, i.e. *sky blue* in the *sky* entry.
(2) Enter it as a separate LU under the second element, i.e. *sky blue* as LU subentry in the *blue* entry.

[15] Members of some speech communities are believed to prefer certain search strategies over others (for instance, it is often said that German users will look for a phrase first under the noun). A lot of academic research has been carried out with a view to discovering where dictionary users expect to find various types of MWE. See the reading list at the end of the chapter for some references.

HEADWORD	**bargain**
LU #	**1**
WORDCLASS	noun
MEANING	an agreement between people to do certain things
EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**2**
WORDCLASS	noun
MEANING	something bought for less than usual price
EXAMPLE	That second-hand adjustable table was a real bargain.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**3**
WORDCLASS	verb
MEANING	negotiate the terms/conditions of something
EXAMPLE	Buyers will bargain hard to cut the cost of the house they want, but dig in their heels rather than reduce their own asking price.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**4**
MWE	**into the bargain**
MWE-TYPE	**idiom**
MEANING	in addition, on top of everything else
EXAMPLE	He must have been an eligible enough bachelor: the son of a fairly prosperous artisan family, trained as a craftsman, and quite a handsome chap into the bargain.
EXAMPLE	In consideration of this, not only would they shelve the files in question, but pay him a monthly stipend into the bargain.

Fig 9.33 Method 2: the idiom as stand-alone LU

(3) Enter it as a headword in its own right, i.e. *sky blue* as *sky blue* main entry.

All of the above alternatives will function perfectly well in the database. There can however be a certain amount of confusion when it comes to recording compounds of which the first element is a noun, principally in noun + noun MWEs. It is not always clear which noun + noun pair should be recorded in the database, as may be seen from the examples of *beach* + noun pairs shown in Figure 9.34.

About 33 per cent of all instances of the noun *beach* in the corpus show it modifying another noun, so deciding which of them to record as compounds in the database is a very real problem. Frequency of the compound in the corpus is the basic criterion for recording this compound, but it is not enough. At first sight, of the *beach* compounds in Figure 9.34, one might reasonably expect *beach ball*, *beach blanket*, *beach buggy*, *beach bum*, and *beach club* to justify entry into the database as compounds, on the grounds

1	Radisson Hotel is the top option, with	beach	access, pool and other amenities.
2	The park has a playground and an improved	beach	area.
3	He caught the	beach	ball in a game of toss.
4	It was found near a	beach	bar that was being built.
5	Rollerblades, tennis racket, guitar,	beach	blanket, sketch book . . .
6	They drove around in Charles's	beach	buggy.
7	Without her, he'd be nothing more than a	beach	bum right now.
8	Price includes dinner, bed and breakfast -	beach	charges are not included.
9	Sewage spills caused four	beach	closures at Marina del Rey
10	one with 1,600 hotel rooms, two	beach	clubs, a water park . . .
11	. . . may develop into more than just a	beach	community . . .
12	They have been in demand at seaside	beach	displays.
13	. . . the lifestyle of California	beach	dwellers.
14	Hurricane Felix caused	beach	erosion from Florida to New York.
15	. . . a noon-to-midnight	beach	event, long talked about . . .

Fig 9.34 Concordances showing *beach* modifying another noun

that they are all 'functional compounds' (as defined in §6.2.2.3); they may well figure in the dictionary drawn from the database, depending on the perceived needs of that dictionary's users. In a bilingual dictionary, it is likely that most, if not all, of these (like most functional compounds) will have L2 equivalents which are not exactly one-to-one translations of the L1 item. Some of the other *beach* compounds mean no more than the sum of their parts: *beach access* ('access to the beach', cf. *hotel access, hospital access*), *beach area* ('area including the beach', cf. *lawn area*, *car park area*), *beach charges* ('charges for using the beach', cf. *car park charges, swimming pool charges*), *beach closures*, *beach displays*, *beach erosion*, and *beach event*. That leaves *beach bar*, *beach community*, and *beach dweller* – worth entering as compounds? The alternative is to include *bar*, *community*, and *dweller* as collocates (cf. §9.2.7 for more information on collocates) in the 'modifier' LU of the headword *beach*. (Many nouns are frequently found, like *beach*, modifying other nouns, and for these the grammatical category 'modifier' exists alongside 'noun', 'verb', 'adjective', and so on.) It's usually better to opt for the modifier section of the first noun's entry (as in Figure 9.35). The dictionary editor always has the option of 'promoting' the MWE to full compound status, or omitting it altogether, depending on the actual dictionary being written. Two ways of entering noun + noun MWEs into the database are shown in Figure 9.35 (within a *modifier* LU, with the second noun recorded as a *collocate* of the keyword) and Figure 9.36 (as a stand-alone LU).

HEADWORD	**bargain**
LU #	**1**
WORDCLASS	noun
MEANING	an agreement between people to do certain things
EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**2**
WORDCLASS	noun
MEANING	something bought for less than usual price
EXAMPLE	That second-hand adjustable table was a real bargain.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**3**
WORDCLASS	**modifier**
MEANING	bought for less than usual price
EXAMPLE	A five-day bargain return for car and two adults costs £84.
EXAMPLE	Millions of us get email messages announcing bargain airfares.
COLLOC	**basement**
EXAMPLE	It was a cross between an early Roman slave market and Selfridge's **bargain basement**.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**4**
WORDCLASS	verb
(*etc. etc.*)	

Fig 9.35 Noun + noun MWE within 'modifier' LU

➔ When in doubt enter a noun + noun MWE into the database in the *modifier* section of the first noun.

Phrasal verbs Here the options are:

(1) Enter the phrasal verb as a separate LU under the verb, i.e. *carry forward* as LU subentry in the *carry* entry.
(2) Enter it as a headword in its own right, i.e. *carry forward* as *carry forward* main entry.

The main problems raised by phrasal verbs are discussed in §6.2.2.4: how to handle phrasal verbs which consist of a motion verb plus a particle, and how to handle two- and three-part phrasal verbs. The Style Guide must give clear instructions on these points.

Support verb constructions Here the options are:

(1) Enter the MWE under the noun, i.e. *make a decision* in the *decision* entry.

HEADWORD	**bargain**
LU #	**1**
WORDCLASS	noun
MEANING	an agreement between people to do certain things
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**2**
WORDCLASS	noun
MEANING	something bought for less than usual price
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**3**
WORDCLASS	modifier
MEANING	bought for less than usual price
EXAMPLE	A five-day bargain return for car and two adults costs £84.
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**4**
WORDCLASS	verb
(*etc. etc.*)	
HEADWORD	**bargain**
LU #	**5**
MWE	**into the bargain**
MWE-TYPE	**idiom**
MEANING	in addition, on top of everything else
EXAMPLE	In consideration of this, not only would they shelve the files in question, but pay him a monthly stipend into the bargain.
HEADWORD	**bargain**
LU #	**6**
MWE	**bargain basement**
MWE-TYPE	**compound**
MEANING	lowest floor in a department store where goods are sold cheaply
EXAMPLE	It was a cross between an early Roman slave market and Selfridge's **bargain basement**.
(*etc. etc.*)	

Fig 9.36 Noun + noun MWE as a stand-alone LU

(2) Enter it as a separate LU under the noun, i.e. *make a decision* as LU subentry in the *decision* entry.
(3) Enter it as a headword in its own right, i.e. in a separate *make a decision* entry (highly unlikely).

The methods shown in Figures 9.32, 9.33, 9.35, and 9.36 for recording idioms, collocations, and compounds, using the data fields MWE and MWE-TYPE are all available for recording phrasal verbs and support verb constructions.

9.2.7 Corpus collocates of the LU headword

You shall know a word by the company it keeps. These prescient words of the linguist J. R. Firth[16] neatly sum up the relationship of 'collocates', of words that appear in each other's company more often than chance can explain. The collocates of the headword are of immense importance to a description of that word. They contribute to the naturalness of the contexts that the dictionary user (especially someone trying to write in a foreign language) will produce for that word. Their presence gives a clue to human and computer alike about which of several possible senses of the word is intended.

The term comes from corpus linguistics, but its definition is not stable. In its most general sense, two words are *collocates* of one another when they co-occur in a specific window of corpus text, which may be any arbitrary number of words, or a sentence, depending on the use to be made of the information. In this volume we use the word *collocates* to refer to words which:

- co-occur with one another with a frequency greater than chance, and
- stand in a major grammatical relationship to the headword of the entry being compiled.

Some examples of such collocates are:

- a noun object of a headword verb (as for instance the nouns *relationship*, *bonds*, and *alliance* for the verb *forge*)
- a noun subject of a headword verb (as for instance *sun* in the case of the verbs *set* and *rise*)
- a noun modifying a headword noun (*maiden* in the case of *speech*)
- an adjective modifying a noun headword (*empty* in the case of *promise*)
- an adverb modifying a verb or an adjective headword (*categorically* for *deny*, *seriously* for *injured*), and so on.

9.2.7.1 *Collocates in the corpus* Collocates worth recording are identified on the basis of significant frequency in the corpus contexts of the headword: that is the task of the corpus query software. They are easily found by the Sketch Engine (cf. §4.3.1), which produces 'word sketches' based on co-occurrence statistics of words standing in specific grammatical relationships

[16] In *A Synopsis of Linguistic Theory, 1930–1955*, Oxford: Blackwell (1957).

to the headword, such as those exemplified in the partial word sketch for the noun *bargain* shown in Figure 9.37, where every word is a collocate of *bargain*.

object_of	**264**	**2.7**	**a_modifier**	**251**	**2.0**
strike	61	43.38	hard	23	25.99
drive	26	27.56	real	20	23.43
get	27	16.38	best	14	19.31
seal	5	14.82	good	19	18.01
make	26	13.6	bad	8	15.31
find	8	7.81	better	8	14.4
modifies	**221**	**0.9**	**n_modifier**	**115**	**1.1**
basement	22	38.62	plea	26	40.62
hunter	22	37.23	wage	6	16.8
price	54	33.65	credit	6	14.68
bookshop	11	26.73	sale	5	10.47

Fig 9.37 Part of the Word Sketch for the noun *bargain*

Figure 9.37 illustrates four grammatical relationships. In the first we see that *bargain* occurs in the BNC 264 times as the object of a verb (the most significant verbs are listed in the table below), and the 'salience score'[17] of 2.7 tells us that this is fairly typical verb behaviour. However *strike* is a very important collocate of *bargain* – the higher the salience score, the more significant the collocate. Similarly important collocates seen in Figure 9.37 are the adjectives ('a_modifier') *hard* etc. modifying *bargain*, the nouns *basement* etc. when modified by *bargain* and the nouns *plea* etc. when modifying *bargain*. These words are typical of collocates to be recorded in the database.

9.2.7.2 *Collocates in the database* Storing collocates in the database involves the use of three data fields: COLLOCATE (the main one), and COLLOCATE-TYPE and GRAMMAR: the use of these two will be explained later in this section. Since good corpus query software will provide the type of grammatical relationship ('object_of' etc.) and the actual salience scores,

[17] Adam Kilgarriff, whose Sketch Engine produces this data, explains this as follows (personal communication): 'The salience score is the product of the MI (mutual information score) and the log of joint frequency. We have found that this provides a better match for lexicographical salience, as judged by professional lexicographers, than MI alone or other measures which have been proposed.'

it is usually enough to record the facts briefly, as in the partial entry for the noun *bargain* (LU-1) shown in Figure 9.38.

HEADWORD	**bargain**
LU #	1
WORDCLASS	noun
EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
EXAMPLE	A credit agreement could be re-opened, if the court thought just, on the grounds that the bargain was extortionate, on the debtor's application to the High Court, a county court or a sheriff court.
COLLOCATE	**strike**
EXAMPLE	Within minutes, Sykes had **struck** a bargain, never stopping for a moment to ask where or how the fish had been caught.
EXAMPLE	Buyer and seller **strike** a bargain with each individual purchase.
COLLOCATE	**hard**
EXAMPLE	She says it was because the EC drove such a **hard** bargain on fish in 1972 that public opinion turned against membership.
EXAMPLE	Don't become despondent just because it seems that your employer is keen to drive a **hard** bargain.

Fig 9.38 Part of the entry for *bargain* (LU-1)

9.2.7.3 *Itemizers and collectives* There are two types of nouns which often appear as collocates of noun headwords, both with an important function in the language. These are 'itemizers' and 'collectives', and they should be recorded systematically. Knowing what itemizers naturally occur with any specific mass noun can be of great assistance to someone writing in a foreign language. A good dictionary will guide its users on this point.

Itemizers English has a rich store of *itemizers*, words that are used to refer to parts of a substance denoted by a mass noun. Well-known itemizers like *spoonful* and *slice* and the idiomatic '*rasher* of bacon' are regularly recorded, but the use of many common nouns as itemizers is more rarely noted. A good example is the noun *speck*, with its wide range of co-occurring mass nouns, in bold type in Figure 9.39. These mass-noun collocates should be recorded in the lexical entry for *speck*, as shown in Figure 9.40.

This type of information is equally useful in the entry for the mass noun itself, and a certain amount of duplication is advisable here. Figure 9.41 shows the itemizers (in bold) found in the BNC in the context of the headword *yellow*. The entry for *yellow* will record the most significant of these itemizers in the way shown in Figure 9.42, where the all-purpose GRAMMAR field (a rather flexible name) is used to hold the fact that the set of collocates to follow are itemizers.

no single	speck	of **dust** dared show itself
there isn't a	speck	of **dust** anywhere
he brushed a	speck	of **dust** from his sleeve
a	speck	of **dirt** never hurt anyone
A	speck	of **grit** danced over the lower corner
he came off the field without a	speck	of **mud** on his shorts
a rosette of small pink petals with a	speck	of **green** at its heart
not a	speck	of **talcum powder** on any surface
the bishop wiped a	speck	of **sauce** from his chin
I climbed high, to a tiny blue	speck	of **water**, unfished for decades
there was a	speck	of **yellow** on the horizon

Fig 9.39 The noun *speck* used as an itemizer

HEADWORD	**speck**
LU #	1
WORDCLASS	noun
MEANING	tiny spot, small particle
GRAMMAR	**itemizer**
CONSTRUCTION	PP-of NP
COLLOCATE	**dust**
EXAMPLE	He brushed a speck of **dust** from his sleeve as he waited for the silence he required.
EXAMPLE	She hung it over the line outside where she beat every speck of **dust** from it before replacing it over her freshly scrubbed quarry tiles.
COLLOCATE	**dirt**
EXAMPLE	His sympathy was probably worth about as much as the dirt beneath his fingernails – if there had been a speck of **dirt** present beneath those immaculately manicured items.
EXAMPLE	There was not a crumb or a speck of **dirt** to be found anywhere on the scrubbed and polished floors.

Fig 9.40 Part of the entry for *speck* showing itemizer uses

There was a **trail** of	yellow	in the purpling sky
A variegated laurel, a **splash** of	yellow	on shiny leaves
There was a **speck** of	yellow	on the horizon
There is just a **touch** of	yellow	along the very top of the dorsal fin
… with a triangular **patch** of	yellow	like a painted sun
a striking **flash** of	yellow	drew the crowds
a **sheet** of	yellow	where the mustard was in bloom

Fig 9.41 Itemizers found in the context of the noun *yellow*

HEADWORD	**yellow**
LU #	2
WORDCLASS	noun
MEANING	the colour as a noun
EXAMPLE	The colour can also vary from yellow to brown.
EXAMPLE	Whereas greens and blues are cool and make flowers seem to recede, reds and yellows are warm colours that attract attention and stand forward from paler flowers.
GRAMMAR	**itemizer**
CONSTRUCTION	PP-of NP
COLLOCATE	**splash**
EXAMPLE	A variegated laurel, a **splash** of yellow on shiny leaves, was losing a three-way debate with an overbearing pyracanthus and a woody wallful of ivy.
EXAMPLE	Her cottage walls stood sturdy and strong, with its sun **splash** of yellow on the front door.
COLLOCATE	**patch**
EXAMPLE	Towards the edge of the wood, where the ground became open and sloped down to an old fence and a brambly ditch beyond, only a few fading **patches** of pale yellow still showed among the oak-tree roots.
EXAMPLE	When it is serious, spots and **patches** of yellow and red appear as the green chlorophyll is not made, or breaks down.

Fig 9.42 Itemizers recorded in the entry for the noun *yellow*

Collectives A relationship similar to that of an itemizer and its mass noun exists between a 'collective noun' and the plural noun denoting the entities that it groups. Idiomatic collectives such as *a pride of lions* or *a school of fish* figure in vocabulary lessons, and will of course be recorded in the database provided they are attested in the corpus. But it is the more generic collective nouns that interest the lexicographer. The nouns that collocate with the collective noun *horde*, and their translations, will be of great value to the dictionary user, and the *horde* entry will include information about these collocates. Often the 'targets' of collectives group into 'lexical sets' (particular semantic classes of nouns), and the COLLOCATE-TYPE field is used to note that fact. Part of the entry is shown in Figure 9.43, where the function of that field is illustrated: it provides a useful way of showing a particular corpus pattern.

9.2.8 Corpus patterns

Corpus patterns[18] cannot be produced by simply thinking about how your headword is used. They cannot be known intuitively, but are visible only

[18] This phenomenon is similar to what Hoey (2005: 43–44) calls 'colligation' (§8.5.2.3), though broader in scope.

HEADWORD	**horde**
LU #	1
WORDCLASS	noun
MEANING	big group, large number of
EXAMPLE	Riding up they are warned away because the castle is under siege by a savage horde.
EXAMPLE	He had planned to go back to London to pick up his bags, but didn't want to have to deal with the media horde.
GRAMMAR	**collective**
CONSTRUCTION	PP-of NP
COLLOCATE-TYPE	**people**
COLLOCATE	**visitor, admirer, reporter, parent, paparazzi, listener . . .**
EXAMPLE	During the summer months hordes of **visitors** regularly congregate there to eat and drink at their leisure on the paved terrace between the mellow sandstone walls of the inn itself and the river's edge.
EXAMPLE	Hordes of **admirers** screech that they would be only too willing to take her home and look after her for a while.
COLLOCATE-TYPE	**insects**
COLLOCATE	**mosquito, beetle, fly, wasp . . .**
EXAMPLE	The three Frenchmen had just settled into the trench as I contemplated the hordes of **mosquitoes** that were now descending on the area.
EXAMPLE	Experts were divided on the cause, with some blaming the hordes of pollen **beetles** that descended from fields of oil seed rape on to garden crops, others believing the hot, dry weather caused a change in the development of the reproductive system of the plants.
COLLOCATE-TYPE	**animals**
COLLOCATE	**cat, snail, slug, rat . . .**
EXAMPLE	As every gardener knows who has tried to protect his vegetables against marauding hordes of **snails** and **slugs**, they have been remarkably successful in making the transition from sea to land.
	The development works taking place in Market Street could be unsettling hordes of **rats** who had built up a network inside and to the rear of the many old buildings.

Fig 9.43 Collective use recorded in the entry for the noun *horde*

in corpus text, where we can study how large numbers of people use their native language. They consist of any marked 'preference' which the headword displays in terms of verb tense, mood, or aspect; noun number, negative contexts, and so on. For example, corpus data shows that more than 50 per cent of the instances of the verb *dispute* are in the passive, and that the MWE *into the bargain* always occurs at the end of a clause or sentence (as seen in Figure 9.29). These are corpus patterns and should be recorded in the database: they show the headword at its most 'natural', and contribute to a better understanding of its behaviour (and hence to the choice of the most appropriate TL equivalent in bilingual dictionaries, and more useful information in monolingual learners' dictionaries).

The noun *speck* provides a good illustration of a special kind of corpus patterning, related to negative contexts. It's a word that people intuitively reach for when they want to talk about a tiny amount of something unpleasant, or something in the wrong place. Of the eighty-seven instances of *speck(s) of* in the BNC, most of them are used in a context that implies that the speck of mud or dirt etc. was not welcome (cf. Figure 9.39): the word *speck* clearly carries a slightly negative connotation. This type of corpus patterning has been called 'semantic prosody', a relatively new concept.[19] Essentially, the corpus shows us how the choice of one word over another of similar meaning can set up expectations about 'what is coming next'. This is something which language-learners can use in order to produce more natural-sounding text in a language not their own, but it is not something of which language-users are consciously aware.

Proof of this is to be found in the recent renaming and re-renaming of the UK postal service, known for many years (probably since its origin in the seventeenth century) as the Royal Mail. In 2001 it was renamed 'Consignia', on the grounds (to quote the PR man who thought the name up) that 'the name change was to enable the organization to compete in an international marketplace into which the Post Office and Royal Mail brands could not stretch'. This is clear evidence that for the people involved the word *consign* carried a positive connotation. For them the meaning of the verb *consign* was no more than can be found in any dictionary: for instance, from the *Shorter Oxford Dictionary 5th Edition*:

> Deliver or transmit (goods) for sale etc. or custody; send (goods) by carrier, rail etc., (*to*).

Corpus evidence tells us more than that, as may be seen from Figure 9.44. In the 195 instances of *consign to* in the BNC, there were almost no 'neutral' destinations. Six per cent of the intended destinations were *the dustbin* (and 50 per cent of these *the dustbin of history*), *oblivion* came next at 5 per cent, then *hell* at 4 per cent, while *the scrap heap* and *the museum* figured in 3 per cent of the citations; at 1.5 per cent we find *memory*, *obscurity*, *perdition*, *the past*, *the rubbish heap*, and *the wastepaper basket*. Perhaps not surprisingly, Consignia fared badly, so badly in fact that the following year its name reverted to Royal Mail.

[19] It was first articulated in Louw (1993), and subsequently developed by (among others) Sinclair (1996), Stubbs (1996, 2001), and Hoey (2005). For a sceptical voice on this issue, see Whitsitt (2005).

dresses. long	consigned	to the back of the wardrobe, returned . . .
these letters Viola	consigned	to the wastepaper basket
it's time the old style of letter was	consigned	to the history books.
their vocalizations are	consigned	to the category of mere noise
the old adage . . . has been	consigned	to the dustbin as a forgotten tradition
the assorted nasties	consigned	to the dustbin of history
these were disastrous and should be	consigned	to the rubbish heap
to say the videodisc should be	consigned	to the scrap heap of technological history
you will be	consigned	straight to hell
if the Thatcher years were not to be	consigned	to oblivion
scenes like this can be	consigned	to the past
he thought he was dead,	consigned	to perdition and gloom until the end of time
Gary had indeed been	consigned	to Satan by his grandmother

Fig 9.44 Concordance lines for *consign*

Semantic prosody is still far from being formalized enough to appear as a specific data type in a lexical database. However, it is too useful to ignore, and the lexical entry for *consign* should contain a section like that shown in Figure 9.45, to alert translators and dictionary editors to this aspect of the word's behaviour.

HEADWORD	**consign**
LU #	1
WORDCLASS	verb
MEANING	send, deliver, put into the care of someone
CONSTRUCTION	NP PP-to
CORPUS-PATTERN	**semantic prosody: object of prep *to*; always negative, implies getting rid of something unpleasant; things are consigned to nasty places**
EXAMPLE	The old adage that the Irish are at their most dangerous when they are at their lowest ebb has been consigned to the **dustbin** as a forgotten tradition.
EXAMPLE	Most of the people responsible for Labour's policies acknowledge that those which may have been relevant in the 1980s were disastrous and should be consigned to the **rubbish heap**.

Fig 9.45 Corpus patterning in the entry for *consign*

The CORPUS-PATTERN data field may be inserted at any point in the LU entry. In Figure 9.45 its appearance at the top of the entry, before any examples, shows that it applies to the whole LU. The same is true of the construction in that entry, where the *to* prepositional phrase is obligatory.

9.2.9 Linguistic labels

Linguistic labels are recorded in the database using specific data fields, drawn from those discussed in §6.4.1.4 and §7.2.8. These are DOMAIN, REGION, DIALECT, REGISTER, STYLE, TIME, SLANG AND JARGON,

ATTITUDE, OFFENSIVE TERMS, and MEANING TYPE. These labels may be attached to any kind of lexicographic item, a whole lemma, or an LU, or a multiword expression, or simply a single example phrase within an LU. They are normally inserted after the item they refer to. The usual practice is to label every appropriate item in the database, where they are particularly useful, in that they offer a way of automating lists of non-standard vocabulary items for the purpose of systematizing their treatment. Domain labels are especially important for the systematic handling of specialist vocabulary.

However, every type of information in any of the labels is essential to have on hand when it comes to translating database material. If a good target-language equivalent of a word or usage is to be found, translators must know whether the headword belongs to a particular domain; whether it is a

HEADWORD	**ape**
LU #	1
WORDCLASS	noun
DOMAIN LABEL	**zoology**
MEANING	large monkey
EXAMPLE	It was her research that showed how close we are in evolutionary terms to the apes: how they communicate with each other, use tools and so on.
EXAMPLE	His paper on the orang-utan is especially welcome for collating so much biological information on this fascinating ape.
HEADWORD	**ape**
LU #	2
WORDCLASS	noun
REGISTER LABEL	**informal**
ATTITUDE LABEL	**pejorative**
MEANING	rude or stupid person (or both!)
EXAMPLE	I chained his bicycle to an ornamental fence outside the apartment block, and persuaded the uniformed ape on security that I was not a terrorist.
EXAMPLE	What on earth are you doing, grinning like an ape with your eyes shut?
HEADWORD	**ape**
LU #	3
MWE	to go ape
MWE-TYPE	idiom
REGISTER LABEL	**very informal**
MEANING	to lose or almost lose control (from excitement, anger etc.)
EXAMPLE	He kissed her on the lips, and the crowd went ape, especially when Cicely draped a hand over the dapper old gent's shoulder and kissed him back.
EXAMPLE	He'd gone ape, lost his temper completely, just about assaulted a pupil.

Fig 9.46 Labels in the database

formal or informal word, or unmarked on that scale; whether it is a modern usage, or obsolescent; whether its import is pejorative or appreciative. The most appropriate TL equivalent will share these properties, as far as possible. Figure 9.46 shows the way labels are used in the partial database entry for *ape*.

9.2.10 Cross-references

If you are using a dictionary writing system (DWS), you will probably have to insert very few cross-references, since all DWS software has extensive cross-referencing functionality (including a final checking routine validating all cross-references). There is still a role for manual cross-references. A CROSS-REFERENCE field may be inserted at any appropriate point in the database entry. In most dictionary writing systems, a cross-reference to the word *lame* inserted in the *duck* entry will send a signal to the editor of the *lame* entry that the idiomatic compound *lame duck* should be handled there.

Cross-references in a database or dictionary most commonly serve one of the following purposes:

- They link a headword to its appearance elsewhere within a MWE, e.g. the entries for *cat* and *dog* might contain a cross-reference to the appropriate *rain* LU.
- They link single-word headwords such as *right* and *cream* to compounds of which these are the second element (*all right*, *ice cream*), wherever these compounds appear in the database (as headwords in their own right, or within the entries for *all* and *ice*).
- They link single-word headwords such as *wrap*, *hair*, or *colour* to hyphenated compounds such as *gift-wrapped*, *black-haired*, or *peach-coloured*, whether these appear as headwords or within entries elsewhere in the database.

9.2.11 Comments

The COMMENT field is exactly what its name implies. It allows the database editors, and anyone subsequently working with the database, to leave a note for other editors. Comments are often used to note insights for which there is no formal space, or to explain some unresolved issues with the entry, or to warn of unexpected problems. Comments can be inserted anywhere in an entry. They will never be published.

9.3 Using template entries in database building

Template entries are fully explained in §4.5. If the template classes are correctly chosen, and the entry structure and contents of each template carefully planned, their contribution to the analysis process is of the greatest significance, both in maintaining consistency throughout the database and in cutting the time it takes lexicographers to write the entries.

The actual choice of templates to be written at the start of a database project should cover both the classes of words for which templates are needed in the monolingual dictionary (cf. §10.1.3) and those needed by the bilingual team (cf. §12.1.3).

Exercises

Exercise 1: Building a database entry

Exercise 1a

For this exercise you need to be able to query a corpus, either in English or in another language that you want to work in.

- Choose a verb headword which, in your dictionary, has two or three senses only.
- Now, for each of these senses . . .
  - Make a subcorpus containing the concordances for that dictionary sense. If you are using a large corpus, we suggest that for each sense you select a sample of no more than 100 concordance lines.
  - Work systematically through §9.2 of this chapter. From the concordances, record all the facts you can find that are relevant to your headword. You could do that in a 'table' format within your word-processing or spreadsheet program.
  - For each fact recorded, choose two or three example sentences from the corpus.
- How many of these facts actually appeared in your dictionary?

Exercise 1b

Do the same for a noun headword.

Exercise 1c

Do the same for an adjective headword.

Exercise 1d

Do the same for an adverb headword.

Exercise 2: Creating a template entry

Choose a clear semantic category (such as fruit, colours, birds, trees, games and sports, or forms of transport). Then, in a dictionary of your choice, check 10 entries for words belonging to that category:

- How consistently are the different category-members treated? Identify clear cases of inconsistency.
- What kinds of information, in a template for the category, would have improved these entries?

Now use your corpus to identify features shared by words belonging to your category, and build a basic template which would support entry-writing.

Reading

Recommended reading

Apresjan 2002; Atkins, Fillmore, and Johnson 2003; Hanks 2000a, 2002, 2004a.

Further reading on related topics

Atkins 1993, 1995; Atkins and Grundy 2006; Atkins, Kegl, and Levin 1988; Atkins, Levin, and Song 1997; Atkins, Rundell, and Sato 2003; Biber, Conrad, and Reppen 1998 (Part I, chapter 2); Church and Hanks 1990; Cowie 1998; Cruse 1986, 1990; Fillmore 1992, 1995, 1997, 2002; Fillmore and Atkins 1994, 2000; Fontenelle 1996, 2002; Geeraerts 1990; Hanks 1988, 1990, 1993, 1998, 2004b; Hoey 2005; Hunston 2007; Landau 2001: 217–342; Levin 1993; Lewandowska-Tomaszczyk 1990; Louw 1993; Mel'čuk 1988, 1996; Mel'čuk and Polguère 1995; Ruppenhofer, Baker, and Fillmore 2002; Sinclair 1996; Stubbs 1996, 2001; Taylor 1990; Vandeloise 1990; Whitsitt 2005.

How words work with other words: Benson 1990; Čermak 2006; Coffey 2006; Cowie 1981, 1994, 1999a; Cowie and Howarth 1996; Fontenelle 1992, 1996; Hanks 2004b; Hanks, Urbschat, and Gehweiler 2006; Hausmann 1989, 1991; Heid 1994, 1998; Kilgarriff 2006b; Mel'čuk 1988; Moon 1988, 1992, 1996, 1998; Rundell and Stock 1992; Siepmann 2005, 2006; van der Meer 1998.

Regular Polysemy Apresjan 1973; Copestake and Briscoe 1995; Nunberg and Zaenen 1992; Ostler and Atkins 1992.

PART III

Compiling the entry

This page intentionally left blank

Introduction to Part III

This is the moment of truth. Everything in this volume has been leading up to this point. In Chapters 8 and 9 we described a methodology for building a database (and translating it if appropriate). Whether you have followed this route or simply applied the age-old method of drafting and redrafting, you are now ready for the last stage in the process: creating final, publishable dictionary entries. You now have a threefold task:

- *to select* what you need from your database of facts about the headword;
- *to present* them in such a way as to be most helpful to your typical user; and
- *to achieve consistency* by following the Style Guide at every point.

At this point, it's worth bearing in mind two basic truths.

- The first is that if someone with enough knowledge and ability uses the dictionary carefully, and yet consistently gets things wrong, that is our fault, not theirs.
- The second is, in the words of the great Dr. Johnson, 'Every other authour may aspire to praise; the lexicographer can only hope to escape reproach, and even this negative recompense has been yet granted to very few.' (*Preface* to *Dictionary of the English Language*, 1755).

Although the various tasks to be completed are in some respects similar, the different needs of monolingual and bilingual dictionaries mean that the pathways diverge here. There is some overlap in topic but not much in detail

between our discussion of preliminaries in Chapter 10 and in Chapter 12. The last three chapters deal separately with compiling monolingual entries (Chapter 10), inserting translations into the database as a preliminary to writing bilingual entries (Chapter 11) and compiling bilingual entries (Chapter 12).

10

Building the monolingual entry

In this chapter we guide you through the process of compiling entries for a monolingual dictionary. Our starting point is a database, which has been populated during the 'analysis' stage, following the methodology outlined in Chapters 8 and 9. Each lemma in the database comes with a structured inventory of corpus-derived facts, and it is from these that the final dictionary entries will be distilled. In Chapter 8, we explained the criteria by which lemmas are divided into LUs, and it is these LUs that form the basis for 'dictionary senses'. Some LUs, of course, may be treated in the final dictionary as multiword expressions (idioms, phrasal verbs, and so on: §7.2.7.1, §9.2.6) or as run-ons (§7.2.10.2), rather than as 'regular' senses. For every LU, the database provides the following kinds of information:

- a rough characterization of its meaning, which will be transformed into a dictionary definition during the present stage
- a detailed record of its combinatorial behaviour, including:
  - syntactic patterns (§9.2.5)

  - MWEs in which it participates (§9.2.6)
  - lexical collocations (§9.2.7)
  - corpus patterns (§9.2.8)
- an indication of any stylistic, regional, subject-field, or other features that require a linguistic label (§9.2.9)
- one or more examples from the corpus to illustrate each individual fact which the database records (§9.2.4).

The next stage ('synthesis') entails transforming a generic set of database records into a finished entry for a specific dictionary. This involves a process of selection and presentation: selection of facts relevant to this particular dictionary, and presentation of the material in a form appropriate to this particular group of users. With a carefully and systematically populated database, you already have all the information you need to create finished entries. The syntactic, collocational, and sociolinguistic data is logged (and supported by example sentences), so you won't – as a rule – need to go back to the corpus. For many kinds of dictionary – especially those designed for learners – selecting good examples is a challenging operation, and we return to this later (§10.8). But the biggest *new* task at the entry-building stage is writing definitions. After word sense disambiguation, definition-writing is the most difficult aspect of the monolingual lexicographer's job, and a substantial part of this chapter is devoted to the theoretical and practical issues relating to this task (§10.4–§10.7). First, though, we will go back to the beginning of the process, and look at the resources you will need in order to do the job successfully.

10.1 Preliminaries: resources for entry-building

In addition to the database itself, three other resources come into play at this stage:

- the user profile
- the Style Guide
- template entries

We will briefly consider how each of these impacts on the entry-building process.

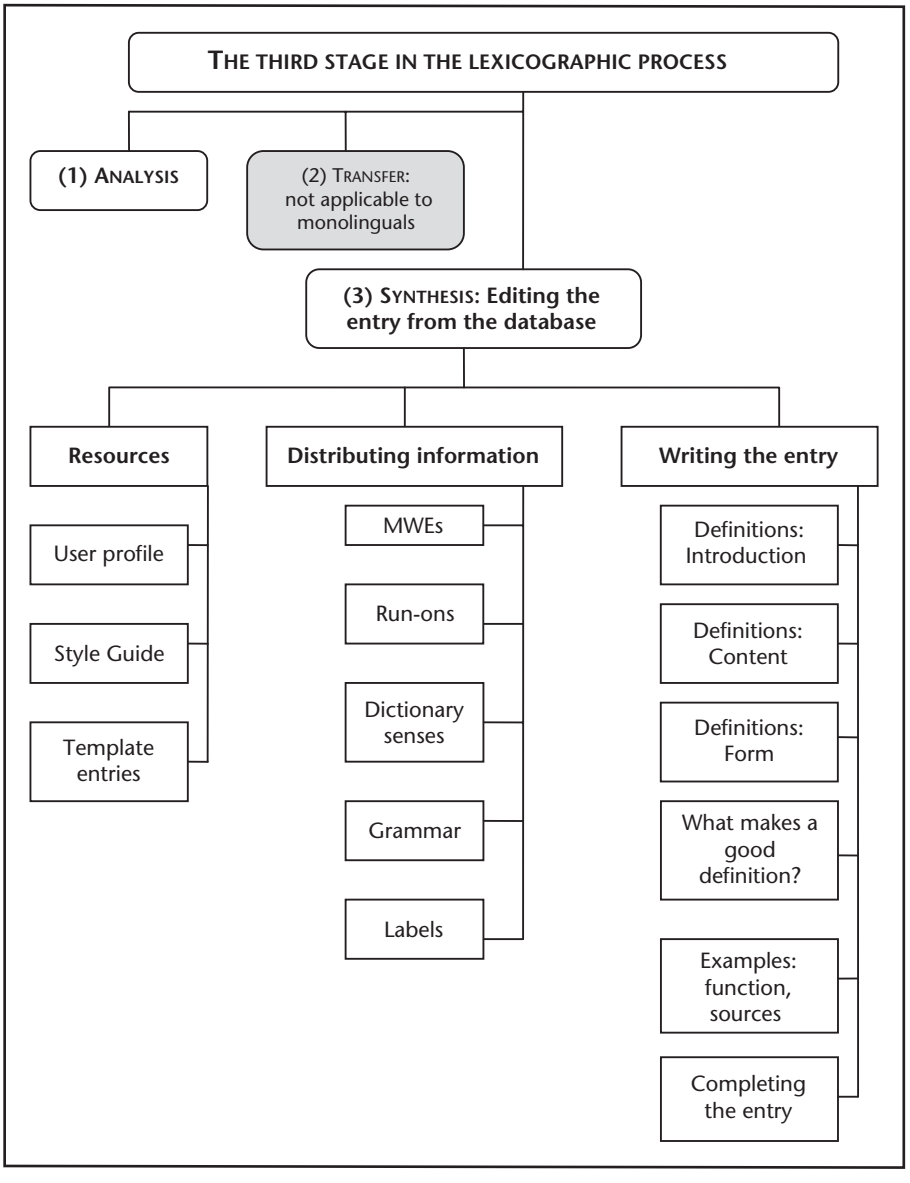


Fig 10.1 Contents of this chapter

10.1.1 The user profile

The user profile (§2.3.1) critically affects the *selection* and *presentation* of information. If we think of the information in a dictionary as a subset of all the facts recorded in the database, it's obvious that a 12-year-old in the early

stages of secondary education is going to need a different configuration of facts from (say) a professional adult writer. And once we have selected the information appropriate to each category of user, the way that information is presented will be largely determined by what we know about the user's skills and knowledge (or lack thereof). As we saw earlier, a well-defined user profile will help us make the right decisions about *content*, affecting areas such as:

- Headword selection: for example, does our user need vocabulary items that are dated, literary, or highly technical?
- Sense selection: similar questions apply.
- Granularity of senses: does our user need a finely split description of a word's different uses, or will a broadbrush treatment be more helpful (§8.1.3)?
- Granularity of labels: the inventory of labels used in a large, unabridged volume may be quite extensive (for example, covering specific subject-fields like *anatomy* and *physiology*), whereas in a lower-level dictionary a smaller set of broad labels (such as *medical*) may be more appropriate.
- Grammatical and syntactic information: native speakers don't generally need to be told that *knowledge* is an uncountable noun (and can't be pluralized), or that *prevent* is typically used in the pattern *prevent sb from doing sth* (rather than **prevent sb to do sth*); but if the user is a language-learner, this is essential information.
- Examples: some types of dictionary contain very few examples, others make extensive use of them, while others again (think of Johnson or the *OED*) use only attributed citations (§§10.8.1–10.8.2). Which of these options we choose will depend on what we know about the user's needs.

Similarly, the *presentation* of information should be guided by an understanding of the user's reference skills, knowledge of the world, and linguistic competence. This can make a big difference in areas such as:

- The dictionary's metalanguage and conventions: will the user understand abbreviations like *colloq*. or *dial*.? Can we assume they know the International Phonetic Alphabet? Will the user be familiar with lexicographic conventions like the specialized use of brackets in definitions?[1]

[1] A good example of metalanguage differences is the way pedagogical dictionaries describe recurrent word combinations. In its first (2002) edition *MED* introduced

- The language used in definitions: can we be confident users will understand the words we use to frame our definitions? Will they 'correctly' interpret conventional defining formulae like 'any of various types of X'?
- If the user needs grammatical information, what form should it take, and to what depth should it go? Can we expect the user to understand transitivity or countability, or even basic grammatical categories like subject and object?

amendment 1 [C] a change made to a law or agreement **1a** [C] one of the changes that has been made to the US constitution **1b** [U] the process of changing a law or arrangement **2** [C] a change made in a document or plan
MED-1 (2002)

amendment [C] a change made to a law, agreement, or document
Macmillan Essential Dictionary 2003

Fig 10.2 Two entries for *amendment* from the same database

Figure 10.2 illustrates the impact of user profiling, by comparing entries for the same word in two different dictionaries derived from a single database. The left-hand entry is from a dictionary aimed at advanced learners, the right-hand one from a dictionary for learners at intermediate level. The entries are, in other words, designed for different kinds of user. The most obvious difference is that the second entry is far shorter – necessarily so because the dictionary it appears in is about 60 per cent smaller than its sister publication. Lower-level dictionaries are almost always smaller than those for more advanced users, and this reflects users' preferences and requirements: novice language-learners feel daunted by big, serious-looking dictionaries, and – because of the kinds of task they typically perform and the kinds of text they encounter – they don't need as wide a range of vocabulary. The first entry here has two main senses and a further two subsenses. Some of this information is simply omitted in the shorter entry. This is done either on grounds of rarity (the uncountable use makes up

what were (internally) called 'collocation boxes', a type of usage note listing common collocates of the headword. These were headed 'Words frequently used with X'. The second edition of *CALD* (2005) does something similar, with the heading 'Words that go with X'. In the *COBUILD Advanced Dictionary of American English* (2007), boxes of this type are labelled 'Word Partnership'. But in its 2nd edition (2007), *MED* opts for the more technical heading 'Collocation', in response to market research which indicated that the term was now widely understood by potential users.

fewer than 10 per cent of corpus instances), or on the assumption that if the user comes across a more specialized use, its meaning can be deduced from the *general* definition (this deals with the 'US constitution' sense). The second main sense in the larger dictionary refers to a change in a 'document', and this point is now included in the single definition in the lower-level dictionary.

This discussion illustrates, in microcosm, the importance of a carefully developed user profile. Every decision we make, in selecting facts from the database and presenting them in the finished entry, will be influenced by our understanding of who the user is, what they need their dictionary to do for them, and what skills and knowledge they bring to the task of consulting it.

10.1.2 The Style Guide

The Style Guide, as we saw earlier (§4.4), is a set of instructions which provides detailed guidelines for handling every aspect of the microstructure. These guidelines reflect general policy decisions made at the outset of the project – and those decisions, in turn, reflect our understanding of the needs and capabilities of the intended user. The Style Guide affects both content and presentation. It will explain, for example, the criteria for deciding whether to treat a word-form as a run-on (cf. §7.2.10.2), the mechanisms that can be used for showing variants or cross-references, and the range of definition types that is allowable. Variations among different dictionaries may be subtle, but from the user's point of view they are significant. In Figure 10.3, we compare the way examples are treated in two dictionaries: *OALD-7* (a learners' dictionary) and *MWC-11* (a dictionary for adult native speakers). This illustrates some of the policy differences between the two dictionaries, and gives an idea of the fine-grained issues that a Style Guide has to address.

The examples here embody a raft of individual policy decisions, each of which will have been influenced by what is known about the dictionary's intended user – and each of which we must take account of as we embark on the task of converting the raw material in the database to finished dictionary entries. As for example sentences, so for every other part of the microstructure: the Style Guide reflects all these policy decisions, and tells you how to apply them. It will tell you how to deal with all the entry components we discussed in Chapter 7.

The Style Guide's principal function is to make the dictionary consistent, no matter how many editors are on the team or how long the dictionary

Example	Source	Type	Comment
She fell off a ladder and broke her arm.	*OALD-7*	full sentence	shown in italics, start with capitals, end with full stop
<the bushes will ~ his fall>	*MWC-11*	full sentence	enclosed in angle brackets, not italic, no capitals or full stops, headword replaced by a tilde (~)
the breakdown of law and order	*OALD-7*	non-sentence, complete noun phrase	also found in *MWC-11*
a coffee/lunch/tea break	*OALD-7*	noun phrase with slashed alternatives	not used in *MWC-11*, where alternatives are handled by separate examples: <abort a project> <abort a spaceflight>
He was breaking the speed limit (=travelling faster than the law allows one).	*OALD-7*	example with explanatory 'gloss'	glosses not used in *MWC-11*
<broke his watch>	*MWC-11*	verb phrase without subject	not used in *OALD-7* – verbs always have a subject

Fig 10.3 Aspects of example policy in two dictionaries

takes to compile. This has benefits for lexicographers and dictionary users alike: a well-thought-through set of editorial policies which reflect a coherent ethos will be easier for the editorial team to assimilate, while users will quickly learn the best way to find what they are looking for. By the time the editorial team is ready to start entry-writing, senior editors will have written a hundred or more sample entries, covering all wordclasses and addressing most of the known problems for monolingual dictionaries, and on the basis of this operation the Style Guide will be developed. Some policy decisions will be built into the dictionary writing system (§4.3.2), so

that you can choose from a set of options (in the case of wordclass markers or grammar codes, for example), rather than hunt down what you need in the Style Guide itself. A well-planned Style Guide will answer most of the questions the editorial team will ask of it, but it doesn't remain set in stone. It evolves as new and unforeseen issues arise during the course of the compilation stage, and is rarely in complete and final form until quite late in a project.

10.1.3 Template entries

The Style Guide incorporates the 'rules' for dealing with each individual entry component. But, as we showed earlier (§4.5), the lexicon includes some entire *categories* of word whose members have so much in common with one another that it makes sense to follow a standard model when compiling entries for them. These standard models are what we call 'templates', and a template is a kind of skeleton entry which you flesh out with information from the database. Templates can be written for many kinds of lexical set, and they have the dual benefit of:

- streamlining the entry-writing process (since half the work has already been done for you)
- ensuring that entries belonging to lexical sets are handled systematically, and that relevant information isn't randomly omitted.

We have seen, for example, that some sets of words exhibit a form of 'regular polysemy' (§5.2.4, §8.3.5), so a template for 'Trees' should remind lexicographers to consider the uncountable 'wood-from-this-tree' use (*doors of solid oak*, *a table made of pine*). The approach you take will vary according to the type of dictionary being compiled, and this affects not just the content of templates but even whether a given category needs a template at all. In the development of the (pedagogical and monolingual) *MED*, for example, a template was used for entries referring to people's jobs and occupations. If you look at the entries in *MED* for words such as *plumber*, *electrician*, *glazier*, or *mechanic*, you'll see that they all begin with the defining phrase 'someone whose job is to . . . '. A bilingual dictionary will have quite different concerns (§12.1.3) – most obviously, it won't need templates giving recommended defining formulae. But for a monolingual dictionary, templates deal primarily with the question of how best to define members of

Defining features	Allowable options
gravity	'serious', 'minor'
genus expression	'illness', 'medical condition', 'disease', OTHER
location	'affecting your . . . ', 'which affects . . . '
person	'which (mainly) affects . . . '
cause	'caused by . . . '
symptoms	'that makes you . . . ', 'that makes it difficult for you to . . . ', 'in which . . . ', OTHER
full form	'X is an abbreviation/short form of . . . '

Fig 10.4 Extract from a template for 'Illnesses and medical conditions' for a monolingual dictionary

the set, and they may additionally provide guidance on the types of example sentence or grammatical information that may be needed.

Figure 10.4 shows part of the template for 'Illnesses and medical conditions' used in *MED*. The features in **bold** are obligatory components: any definition of an illness must start by saying it is 'an illness', 'a medical condition', or something similar. The template gives advice about which of these genus expressions (see §10.5.1) to select. The definition must also include information about what the symptoms are. All the other features are optional, and will be invoked as appropriate. Thus for example the definition of *pneumonia* includes information about its 'gravity' and 'location':

a serious illness affecting your lungs that makes it difficult for you to breathe

The definition of *laryngitis* mentions location ('affecting your throat and larynx') but not gravity, while *malaria* refers to gravity ('serious') and to cause ('caused by being bitten by a mosquito . . . '). The template also gives a checklist of questions to ask, advice about examples, and a set of model entries for various kinds of illness.

Most dictionaries are written over an extended period (often several years) and by large teams of editors (often geographically dispersed). So the potential for variation – in entries for words of very similar type – is significant. The use of templates will help keep this to a minimum, and should speed up the editorial process too.

➔ By exploiting systematicity in the language, we introduce systematicity into the finished product *and* into the task of creating it.

10.2 Distributing information: MWEs, run-ons, and senses

With these resources at our disposal, we're ready to approach the job of extracting a final dictionary entry from the database. But before we get to the central tasks of writing definitions (§§10.4–10.7) and selecting appropriate example sentences (§10.8), we will have to decide whether an item in the database should be handled as a multiword expression, a run-on, or a 'regular' dictionary sense. And for each of these components, the 'rules' for locating, ordering, and describing them will vary from dictionary to dictionary. In this section, we look at the kinds of decision you will need to make.

10.2.1 Multiword expressions (MWEs)

Consider the following expressions that include the word *head*:

- *at the head of* (=at the front of)
- *from head to foot* (=all over your body)
- *come to a head* (=reach a climax)
- *head off* (=prevent)
- *head and shoulders* (=by a significant margin)
- *head start* (=an advantage).

Dictionaries exhibit enormous variation in the ways they treat MWEs like these. Figure 10.5 compares the policies of *MW-3* (a big unabridged dictionary for native speakers) and *LDOCE-4* (an advanced learners' dictionary), showing how each book handles idioms, collocations, phrasal verbs, and compounds, and where these different types of MWE appear in each dictionary's microstructure and macrostructure.

The system used in a given dictionary will have been devised with the user in mind and may also reflect a 'house style' employed by the publisher across a range of products. You need to have a clear grasp of what your dictionary's

MWE	*MW-3*	*LDOCE-4*
at the head of	not explicitly described, but covered by the definition of sense 11a ('the leading element of a military column or a procession')	as collocation at sense 5 ('the front or most important position'), with example
from head to foot	not explicitly described, assumed to be deducible from the core meanings of *head* and *foot*	as collocation at sense 1 ('the top part of your body. . . '), with an explanatory gloss: '(=over your whole body)', no example
come to a head	not explicitly described, but covered by the definition of sense 20b ('culminating part of action or of tension') – linked semantically to 20a ('the part of a boil, pimple, or abscess at which it is likely to break')	as a full sense (sense 9) with definition and example, between sense 8 (the starting point of a river) and sense 10 (the flowering top of a plant)
head off	as a full headword in its own alphabetic place	as a phrasal verb, located with other phrasal verbs, following all the 'regular' senses of the verb *head*
head and shoulders	as a full headword in its own alphabetic place (after *headachy* and before *headband*)	as a full sense (sense 29) with definition and example
head start	as a full headword with two senses	as a full headword with two senses, between *headstand* and *headstone*

Fig 10.5 Handling MWEs: variation in dictionary styles

policies are, and (ideally) an understanding of why it does things the way it does.[2]

But before you even get to the point of applying these policies, there is the question of which units of language qualify for being treated as MWEs in the first place. In many (probably most) dictionaries, you will find an MWE entry like the following:

> **salt** . . . –PHRASES . . . **take something with a pinch** (or **grain) of salt** regard something as exaggerated; believe only part of something (*ODE-2* 2003)
>
> **salt**[1] . . . **3 take something with a pinch/grain of salt** *informal* to not completely believe what someone tells you, because you know that they do not always tell the truth (*LDOCE-4* 2003)

We infer from these entries that this is a fixed phrase, with one variable element, and that *pinch* is the more usual 'itemizer' than *grain*. Corpus evidence supports this interpretation – but only up to a point. *Pinch* does indeed outnumber *grain* (by a factor of about three to two). But there is more variation in real text than either entry implies. Both itemizers can be modified and/or pluralized, so we find instances of people taking something:

> *with a slight pinch of salt*
> *with an optimistic pinch of salt*
> *with some hefty pinches of salt*
> *with more than a grain of salt*
> *with a disparaging grain of salt*

Furthermore, there is occasional variation in the itemizer itself, so we find expressions like:

> *with a large fistful of salt*
> *with a cellarful of salt*
> *with a healthy dose of salt*
> *with a huge lump of salt*

Should we conclude, then, that the entries above are misleading? Another dictionary offers an alternative approach:

> [1]**salt** . . . **4d:** corrective allowance: RESERVE, SKEPTICISM – often used in the phrase *take with a grain of salt* (*MW-3* 1961)

The approach here is less restrictive, and this enables the dictionary to account for every conceivable instantiation of this meaning in text. A

[2] On the placement of MWEs, see also Bogaards (1990; 1996: 285ff).

speaker may say 'I'll take that with a grain of salt', 'I'll take that with several large handfuls of salt' or (as happens in a tiny number of cases) 'I'll *treat* that with a pinch of salt' – and in every case, the definition 'works'. And one could argue that finding a definition which maps onto all possible instances of this phrase is an appropriate strategy for a big unabridged dictionary like *Webster's Third*. But for most types of dictionary user, this is a less helpful model than the ones we showed above. It clearly fails to support the encoding function – using the definition as a basis for producing one's own text (§10.4.2.2) – because it appears to authorize sentences like:

> **It would be wise to exercise a degree of salt.*
> **There was widespread public salt about the official version of events.*

But even from a decoding point of view, there are strong arguments in favour of the approach used in *ODE* and *LDOCE*. The great majority of uses take precisely the form shown in these entries and, although variations in the itemizer can and do occur, they make up a tiny minority of all occurrences (at best 5 per cent). In this sense, treating this usage as an MWE is truer to the data and more likely to help the user find the 'right' meaning of *salt* with minimum effort. Since this meaning is almost always expressed as a fixed phrase, any variations can be easily construed as creative exploitations of a very stable norm.[3]

10.2.2 Run-ons

Run-ons (undefined derived forms, typically located at the end of a main entry) have long been used in dictionaries as a device for achieving broader coverage at a low cost in terms of space. But as we saw earlier (§7.2.10.2), run-ons are not without their problems. A good Style Guide will set out criteria for admitting words as run-ons, and will indicate which suffixes are allowable. Meanwhile, frequency data from the corpus will help us decide whether a given form is worth including in the dictionary at all. Word-formation rules allow language-users to generate almost endless numbers of derived forms, but – as always – the dictionary's currency is 'the

[3] An equally bizarre (and misleading) example of this phenomenon can be found in *LDOCE-1* (1978), where sense 6 of the noun *spot* is defined as 'an area of mind or feeling'. All becomes clear when we read the example that follows: *I have a soft spot for my old school*. But *have a soft spot for* is an even more fixed expression than *take with a pinch of salt*. What other kinds of 'spot' could you have in this meaning?

probable, not the possible'.[4] This is especially worth bearing in mind when working with an electronic dictionary. There is a temptation – once we are liberated from space constraints – to include all sorts of information simply because we can. But there is never a good case for including something which we know to be of no practical value to our target user.

10.2.3 Dictionary senses

With MWEs and run-ons accounted for, we now turn our attention to dictionary senses. In Chapter 8, we described an approach for identifying distinct 'lexical units' (LUs) in words that exhibit polysemy (see esp. §8.5, §8.6.3). These are the building blocks of the database, and it is from these LUs that we will derive the inventory of senses for each of the headwords in a particular dictionary. As we will see in Chapter 12, the difference between database and final dictionary entry in a bilingual dictionary can be dramatic: a word in the source language may have a large number of LUs with very different meanings – but they might all be translated by a single target-language equivalent (see §12.3.1 and Figure 12.9). Nothing quite this extreme applies in the case of monolinguals, but – as we showed earlier (Figure 10.2) – the number and type of senses that make it through to the final entry will depend on the kind of dictionary we are writing. The entries in smaller dictionaries (whether for learners or native speakers) generally have fewer senses, and fewer kinds of sense-division, than those in larger volumes, so the Style Guide needs to advise on how and when to merge database senses, which LUs can be omitted, and whether subsenses are allowable. The order in which the senses appear in the entry will then need to be considered; the main options for determining sense order, and the principles underlying them, are outlined in Chapter 7 (§7.3.2, §7.3.3). With a first draft of the sense-divisions in place, you now have the outline of a final entry, and from this point the focus will be less on the headword as a whole than on each individual sense. As these are fleshed out, you may find yourself re-visiting the entry structure, for example by merging two putative

[4] Thus, among beekeepers, the phenomenon of *queenlessness* is regularly discussed, but for a general-purpose dictionary, the word is far too infrequent to be worth including even as a run-on.

senses which turn out to have so much in common that they are best treated as one. But that comes later.

10.3 Systems for handling grammar and labelling

In the database, the grammatical and sociolinguistic features of each LU are described in detail (see §9.2.5 on grammar, §9.2.9 on linguistic labels). At the point of composing the final entry, you will sometimes find that all the LUs of a headword share the same features. For example, the verb *abandon* has several distinct meanings, but in all of them it functions as a transitive verb – so the system your Style Guide recommends for noting transitivity will in this case be applied to the entry as a whole. Similarly the adjective *pissed* has two senses ('drunk' and 'annoyed') but both are informal. In the latter case, however, the two meanings are further distinguished in terms of *region*: when it means 'drunk', *pissed* is characteristic of British English, whereas for American speakers the word means 'annoyed'. The final entry thus has labels at two different levels:

- at headword level, the label *informal* applies to the whole entry
- at sense level, the labels *British* or *American* apply to one or other sense.

As we convert the information in the database into a final entry, it's important to distinguish between features which apply to the entry as a whole, and features that belong to specific senses. In this section, we discuss the issues you will need to be aware of when dealing with grammar and labels.

10.3.1 Grammar

A well-designed and well-populated database will include detailed grammatical information for each LU of each lemma. How much of this information finds its way into the final entry, and in what form, will depend on the type of dictionary and the type of user. The user profile will give you a steer on what the user needs to know about a word's grammatical behaviour, and on how much prior knowledge of grammar can be assumed. All of this feeds into the policy decisions on grammar that are embodied in the Style Guide. The Style Guide will outline the dictionary's general approach to grammar,

list the categories, codes, or other systems used for describing grammatical behaviour, and explain the circumstances in which each of these elements should come into play.

In all kinds of monolingual dictionary, basic grammar is supplied in the form of wordclass markers (§7.2.6.1). Dictionaries for native speakers rarely go much further than this. Most give some indication of a verb's transitivity (typically using markers like *vt* or *tr.v.*), though it is questionable whether the average user even notices these labels, still less understands their meaning. Among dictionaries for native speakers, the *Oxford Dictionary of English* (*ODE* 2003) is exceptional in providing a more detailed level of grammatical information. Thus the word *mere* (simply labelled as an 'adjective' in most dictionaries of this type) is additionally described as 'attrib' (attributive) in *ODE*, and other adjectives whose use is restricted in some way attract the labels 'predicative' (*asleep*) and 'postpositive' (*galore*). The transitivity of verbs is signalled not by the usual *v.i.* and *v.t.* codes, but by the more transparent labels [no obj.] and [with obj.]. And where a verb takes an obligatory adverbial, this is noted too:

> **barge** *verb* **1** [no obj., with adverbial of direction] move forcefully or roughly: *we can't just* ***barge into*** *a private garden*
> **reside** *verb* [no obj., with adverbial of place] **1** have one's home in a particular place . . .

The notation for nouns distinguishes 'mass' and 'count' varieties, and also notes cases where a noun is used as a 'modifier' (like *bedside* in *a bedside table*). Finally, adverbs are sometimes subcategorized as 'sentence adverbs' (like *regrettably*) or 'submodifiers' (like *comparatively*). Though the dictionary's Introduction simply explains that grammar 'has begun to enjoy greater prominence', one could speculate that *ODE*'s decision to provide more explicit grammatical information reflects (at least partly) the rise of English as a lingua franca, and the dictionary's likely use by proficient speakers of other languages.

In taking this approach, *ODE* is moving into territory traditionally occupied by the monolingual learners' dictionary (MLD). Since Harold Palmer's pioneering *Grammar of English Words* (London: Longmans, Green, 1938), 'the provision of detailed syntactic information has been fundamental to the MLD tradition' (Rundell 1998: 329). *ODE* is grappling here with the same issue that has long preoccupied pedagogical lexicographers: how to devise a system that combines descriptive power with accessibility. *ODE*'s [with obj.] is presumably seen as more transparent than '*v.t.*', but terms like

'attrib.' and 'postpositive' are unlikely to be familiar to many users. In the field of learners' dictionaries, there has been a steady trend – supported by a good deal of user research[5] – away from formal (but undeniably powerful) inventories of syntactic codes towards more user-friendly ways of describing grammatical categories and accounting for syntactic preferences. The main MLDs available at the end of the 1970s (*OALD-3* and *LDOCE-1*) both used elaborate (but mutually incompatible) coding systems which enabled lexicographers to account for almost every conceivable form of grammatical behaviour, as shown in Figure 10.6.

promise[2] **1** [T1,3,5a,b;V3;D1,5a;IØ] to make a promise to do or give (something) or that (something) will be done: *Do you promise secrecy? . . .*

LDOCE-1 (1978)

promise[2] *vt, vi* **1** [VP6A, 7A, 9, 11, 12A, 13A, 17] make a promise (**1**) to: *They ~d an immediate reply . . .*

OALD-3 (1974)

Fig 10.6 Coded grammar systems in older MLDs

But both schemes were abandoned when it became clear that most users simply ignored the codes because they looked too difficult. In contemporary MLDs, grammatical information is generally conveyed through wordclass markers and a small set of basic codes (like [T] for 'transitive' and [U] for 'uncountable'), elaborated by some permutation of:

- 'pattern illustrations': abbreviated strings like **remember to do sth**, **remember doing sth**
- example sentences: these typically follow the pattern illustration they exemplify; in some dictionaries the example itself is the only indicator of syntactic behaviour, with relevant parts of the sentence highlighted: *Did you **remember to lock** the door?*
- definitions: as we see later (§10.6.3), a word's syntactic preference can be encoded into the wording of some types of definition.

Figure 10.7 shows how a contemporary MLD handles the grammar of *promise*. In this version, the opaque codes of earlier editions have been replaced by 'pattern illustrations', and – crucially – each of these is linked to the example sentence which illustrates it. The lists of codes used in the 1970s have no connection with the examples they relate to, and the codes

[5] A good recent example is *User-friendliness of verb syntax in pedagogical dictionaries of English*, Anna Dziemianko, Lexicographica Series Maior 130, Tübingen: Niemeyer (2006).

are shown in a fixed order (thus the *OALD* codes go from the lower to the higher numbers). But in the version in Figure 10.7, the order of the patterns reflects their relative frequency in the corpus.

promise[1] **1** [I,T] to tell someone that you will definitely do or provide something or that something will happen: *Last night the headmaster promised a full investigation.* | **promise to do sth** *She's promised to do all she can to help.* | **promise (that)** *Hurry up – we promised we wouldn't be late.* | **promise sb (that)** *You promised me the car would be ready on Monday.* | *'Promise me you won't do anything stupid.' 'I promise.'* | **promise sth to sb** *I've promised that book to Ian, I'm afraid.* | **promise sb sth** *The company promised us a bonus this year.*

LDOCE-4 (2003)

Fig 10.7 A contemporary approach to grammar in a learners' dictionary

It is true that systems like this can't always match the delicacy and completeness of earlier coding schemes, but for most users the trade-off (in terms of greater accessibility) is worth the price. The Style Guide will explain how valency patterns and other grammatical information in the database should be expressed in the final entry. But it's also important to be aware that not everything in the database will necessarily be shown in the dictionary at all. In a learners' dictionary, for example, infrequent instances of grammatical behaviour can be left unaccounted for. One of the great benefits of the abundant corpus data now available is that we can distinguish between those constructions that are *possible* and those that actually occur with reasonable regularity. The entries in Figure 10.6 (from the pre-corpus editions of *LDOCE* and *OALD*) both include a code for the construction *promise someone to do something* (respectively, V3 and VP17), but the data shows that this pattern occurs only rarely in text, so – even if this information is logged in the database – it can be safely ignored when we come to write the final entry.

10.3.2 Labels

Linguistic labels are discussed in Chapter 7 (§7.2.8), where we explain

- the various categories of label
- the principles underlying their application in the dictionary
- the scope of their application.

The role of labels in the compilation of the database is outlined in Chapter 9 (§9.2.9). Our advice there was to devise a fine-grained inventory of labels that will allow you to identify any variation from 'default' or neutral values in a word's style, register, regional characteristics, currency, or pragmatic force. It's a good principle to record this kind of information in the database whenever there is anything useful to say, and this point applies especially to labels denoting the 'domain' (or subject-field) in which a word is typically used. Systematic application of a comprehensive set of domain labels and other label types makes it easy to generate specialized wordlists. With a list of (say) every database item in the domain of 'music', we can extract terms to be checked by specialists for accuracy and by editors for consistency. A resource like this could also form the basis for a dictionary of musical terms. And as we noted earlier (§7.2.8.1), domain labels in lexical databases can be used to support automatic word sense disambiguation (cf. §8.5.1.1).

The function of labels in the final dictionary entry, however, is different in significant ways, and you need to be clear about the labelling policy of your specific dictionary before getting into the meat of the entry. Consider the following words:

piano, composer, symphony, embouchure, diatonic

In a database using domain labelling, these would all attract the label 'music', and this has the advantages outlined above. But human users don't need to be told that *piano* (or *composer* or *symphony*) are words from the subject-field of music (stating the obvious won't endear you to your users), and only the last two terms would have a domain label in the dictionary. As always, style policies will differ according to the type and size of dictionary: an unabridged dictionary (or an electronic version of a print dictionary) may use a more fine-grained set of labels than a concise one. So it's important to have a good understanding of where the boundaries are drawn between 'database-only' labels and the labels that will appear in the dictionary itself.

Dictionary labels typically appear as single words (like 'Astronomy' or 'dated') or as abbreviations (like 'N.Amer.' or 'colloq.'). But other strategies are occasionally employed and are worth thinking about. In the first (1987) edition of *COBUILD*, labels were replaced by more discursive explanations attached to the end of the definition, like this one for *boffin*:

A **boffin** is a scientist; an informal word used in British English.

This was a well-motivated attempt to overcome users' known tendency to skip over conventional labels. How effectively it worked is a moot point (a naïve user might construe the second half of the sentence, after the semicolon, as a separate definition), but this approach was abandoned in later editions. On somewhat similar lines, the *Longman Language Activator* (1993) puts information of this kind at the beginning of the definition, as in this entry for *dosh*:

an informal British word meaning money

Here we see a policy tailored to the specific function of the dictionary. One of the *Activator*'s key goals is to make explicit the differences between close synonyms. This often calls for quite subtle semantic disambiguation, but in the case of *dosh* the semantics aren't an issue. What distinguishes *dosh* from other members of the set of words about money is the fact that it is marked for register and region – so this information is shown right at the outset, in the hope that even the most careless user won't be able to ignore it.

The choice of the 'right' label may also be affected by social change. Shocking as it seems now, the word *half-caste* had no label at all in *OALD-3* (1974), and in *LDOCE-1* (1978) it only attracted the rather tentative marker *sometimes derog*. Compare this with its unambiguously negative treatment 25 years later (Figure 10.8).

half-caste *n* [C] *taboo* a very offensive word for someone whose parents are of different races. Do not use this word.
LDOCE-4 (2003)

Fig 10.8 Labelling a 'taboo' word in a learners' dictionary

It is important to be aware, too, of any disparities in use or currency across different speech communities. In the US, for example, *apparel* is a register-neutral word for 'clothing', used mainly in the retail sector. But in contemporary British English, it is rarely used and it has a distinctly formal or literary flavour. Different labels will be needed, depending on the target user.

But as this example suggests, conventional labels are at best a blunt instrument: categories like 'formal' and 'literary' are umbrella terms that conceal a good deal of variation. The word *purchase* has a more formal ring than *buy*, and would sound pompous if used in ordinary conversation.

But the data suggests that in certain situations (for example when talking about buying 'major' items like land, companies, or military hardware) it is a perfectly natural word to choose. A *formal* label may not be much help here. Or again, people in the medical profession, when referring to surgical operations, tend to prefer the word *procedure* – but does this make it an item of medical terminology? Or an instance of euphemism? The corpus can help us to a degree. If the data shows that a particular word (*slaying* for example) appears predominantly in US newspapers, then we can apply labels like *N.Amer, journalism* with some confidence; and there are now corpus-query systems that alert the lexicographer to cases where a word's distribution in the corpus is in some way anomalous. But in general, labelling is an area of lexicography where there is more work to be done, in terms of how the information is presented (so that users actually notice it), and in terms of the quality and delicacy of the information we provide about the kinds of situation in which a word is typically used. The electronic medium offers exciting opportunities here, and one can imagine a hierarchy of options (which the user can display or suppress) from simple broadbrush labels like *informal* to more detailed descriptions – backed by frequency data – of typical contexts of use.

10.4 Definitions: introduction

Explaining what words mean is the central function of a monolingual dictionary. It is also, as Johnson observed, one of the most contentious aspects of the lexicographer's work.

> That part of my work on which I expect malignity most frequently to fasten, is the Explanation; in which I cannot hope to satisfy those, who are perhaps not inclined to be pleased, since I have not always been able to satisfy myself
>
> (Johnson, *Preface*, 1755)

The raw materials we will be working with are already logged in the database, and they include:

- a provisional division of the lemma into LUs, or potential 'dictionary senses' (Chapter 8)
- a rough characterization of the meaning of each LU, or how it contributes to the overall sense of any text it forms part of

- one or more examples from the corpus, showing how the LU is typically used and the kinds of context it usually appears in
- information about the LU's register, collocational behaviour, syntactic and colligational preferences, pragmatic features, and so on, with each fact typically supported by at least one example sentence.

All these resources will come into play as we embark on the challenging task of creating definitions for a dictionary entry. What the definition says, and how it says it, will be heavily influenced by the Style Guide of the particular dictionary you are writing, and the policies which the Style Guide encodes will in turn be influenced by what is known about the target user (see §10.1.1).

10.4.1 Initial thoughts: you can't help noticing . . .

Different dictionaries often define the same concept in very different ways. For example, see Figure 10.9.

cattle *pl.n.* **1.** Any of various chiefly domesticated mammals of the genus *Bos*, including cows, steers, bulls, and oxen, often raised for meat and dairy products
AHD-4 (2000)

cattle PLURAL N **Cattle** are cows and bulls
COBUILD Student's Dictionary (1990)

cattle noun pl cows, bulls and oxen that we breed for meat, milk and leather; they are also used to pull a plough etc.
Heinemann International Students' Dictionary (1991)

Fig 10.9 Three definitions of *cattle*

A naïve observer might wonder why three definitions of the same thing should be so dissimilar. To explain this, we need to think about all the variables that come into play when people write definitions. Each of the dictionaries cited here is designed for a different group of users. The third entry, for example, comes from a dictionary intended for use by schoolchildren in anglophone countries of Africa, and this helps to explain the focus on the products of cattle and on their use as draught animals. The second definition – aimed at learners of English with a low level of proficiency – is obviously the most simple, but it is well adapted to the needs and skills of its target users. The first differs from the third not so much in the

information it supplies as in the language used to supply it ('any of various chiefly domesticated mammals . . . '). What all this shows is that definitions differ in response to what we know about their users, and that the two key parameters are:

- content: the information which the definition includes
- form: the words and structures used for conveying this information.

In the sections that follow, we address the issues relating to both these aspects of defining, and we establish guidelines that will help you make the right decisions about content (§10.5) and form (§10.6). Over the centuries, the question of what words mean and how to define them has stimulated a good deal of theoretical speculation, and relevant contributions will be discussed when appropriate. Yet surprisingly little has been written about the relationship between the definition and the needs and skills of those who will use it. The user's perspective is our invariable starting point, so we will begin by addressing two fundamental questions.

10.4.2 Function: what are definitions for?

The term 'definition' is a misnomer. It implies that a word's meaning can be precisely (and 'definitively') isolated and pinned down. Johnson preferred the term 'explanation' (and so, significantly, did the first corpus-based dictionary of English).[6] This is a more realistic description of what lexicographers actually do, but for the purposes of this discussion, we will stick with the more familiar term.

It could be argued that definitions exist in order to catalogue the meanings in a language, and this is perhaps their chief function in a serious historical dictionary. But their practical purpose is to resolve the communicative needs of dictionary users. It's helpful to characterize these needs in terms of:

- reference, or 'decoding': the user goes to the definition because s/he has encountered an unfamiliar word or expression and needs to know what it means
- productive, or 'encoding': the user wants to write or say something, and this involves encoding the meaning that is in his or her head, in a way that is natural, appropriate, and effective.

[6] The Introduction to the first edition of *COBUILD* (1987) includes a section headed 'Explanations of meaning and use', and the word 'definition' is nowhere to be seen.

Each of these requirements affects the content and form of the definition, so – before we discuss those two aspects in more detail – it will be worth getting a fix on what the decoding and encoding functions entail.

10.4.2.1 *Definitions for decoding* The user's decoding needs can often be satisfied by quite minimal information. For a low-level language-learner who reads about 'a field of *cattle*', a simple explanation like 'Cattle are cows and bulls' will usually be quite adequate. At a somewhat higher level, imagine a reader who encounters the following passage in a novel, and doesn't know what *exiguous* means:

He indicated that Miss Danziger should sit while he dispensed a potion. She gulped it down, paid the *exiguous* dispensing fee, and left the premises.

(Elisabeth Russell Taylor, *Tomorrow*, 1991)

The reader's goal is modest: s/he doesn't need to find out everything there is to know about *exiguous*, but simply to understand what the writer is saying in this particular passage. A definition such as:

very small in size or amount (*ODE-2* 2003)

supplies the necessary information, and does it without making heavy demands on the reader. This is a good illustration of Bolinger's observation that definitions exist 'to help people grasp meanings, and for this purpose their main task is to supply a series of hints and associations that will relate the unknown to something known' (Bolinger 1965: 572). In this case, the process is straightforward: many users would be unfamiliar with a rare word like *exiguous*, but they can all be expected to know what 'small in size or amount' means.

10.4.2.2 *Definitions for encoding* But suppose our reader, having learned a little about this new word, decides to use *exiguous* – to turn it from a passive vocabulary item to an active one. In the original context the word is applied to a sum of money. But what else could be described as *exiguous*: a house? a person? a problem? *Exiguous* has turned up here in a piece of fictional narrative, but how appropriate a choice would it be in other contexts, such as a conversation with a friend, a business report, or an academic paper? The entry in *ODE* provides a couple of other clues: the word is labelled *formal* (which at least rules out using it in casual conversation) and is supported by an example:

my exiguous musical resources.

This does not tell us a great deal. Charles Fillmore (2003: 268) has discussed the application of dictionary definitions to decoding and encoding, and his reflections are worth quoting in full:

> If I find a dictionary that tells us (as many do) that **carrion** is the rotting meat of a dead animal . . . , I will, on learning of some species of reptiles, birds or insects that live on carrion, be equipped to understand what it is that they eat.

In other words, the definition fulfils the decoding function in the context Fillmore imagines, and a quick look at the eighty-six occurrences of *carrion* in the BNC suggests that this definition would almost always be adequate for decoding. However, Fillmore continues:

> If I want to be able to use the word productively, and in appropriate contexts, I need to know more than that. The definition does not inform me that I can't legitimately use the word **carrion** to refer to meat that had been left out of the refrigerator while the family was vacationing, nor can I use it to refer to dead animal parts that I accidentally stepped on while walking in the woods. **Carrion** is the word used of the food of scavengers.

It will already be clear that successful encoding is a more challenging task than understanding a word in context. To use a word or expression productively, you need to know a great deal about it – not only the kinds of contextual detail Fillmore alludes to here, but also:

- its precise semantic features: when and why, for example, would you use *steadfast* rather than *resolute* or *dogged*?
- its collocational and selectional preferences: what things do people typically *carry out*, *perform*, or *conduct*, given that their meanings are very similar? What kinds of thing are typically described as *exiguous*?
- its sociolinguistic features (in terms of register, regional distribution, and so on), which may determine whether a word like *brainy* or *smart* is an appropriate alternative to *intelligent*
- its pragmatic and connotative features: the expressions *it's a piece of cake* and *it's not rocket science* both imply that something isn't (or shouldn't be) difficult to do, but there is a world of difference in how they are used in text. And the same applies to many other words and expressions which not only convey denotative meanings but also tell us something about the speaker's attitude (think of the semantically equivalent pair *svelte* and *skinny*, for example).

Decoding is usually an *ad hoc* operation: you look up an unfamiliar word, the dictionary helps you understand it in the context you found it in, and you get back to the task in hand. You may never see the unfamiliar word again, and (unless you encounter it several more times) it probably won't gain much of a foothold in your mental lexicon. But encoding is a very different kind of skill, and one that can't be applied successfully unless you have access to the diverse types of information we identified above – either by having internalized them already, or by finding them in your dictionary. Not surprisingly, then (to quote Fillmore again), 'the encoding function is in general not accomplished in ordinary dictionaries . . . and often not well achieved in dictionaries made for second-language learners'.

10.4.2.3 *Encoding and decoding: incompatible goals?* It begins to look as if there is a conflict between the encoding and decoding functions. The best kind of definition for decoding will in most cases be a fairly short and underspecified one (like those for *exiguous* and *carrion*, above): definitions like these answer the question at hand while making minimum demands on the user. (If the definition itself is easy to process, the decoding user's main problem will be identifying the 'right' sense when the target word has several to choose from.) But definitions that are well-adapted for reference use (by being short, easy-to-follow, and therefore rather general) are by their nature unlikely to be adequate for encoding. Conversely, of course, a definition which supplies the level of detail needed for encoding might seem unnecessarily complex for the user who is just looking for a 'quick fix' to a reference query. This raises the question (which discussions of defining have rarely addressed) of how far a dictionary should be expected to cater for users' encoding needs, and indeed whether the two functions can be adequately catered for by the same definition. The problem is especially acute in dictionaries for language-learners, whose users need a great deal of support if they are going to produce accurate and natural-sounding text.[7] There are exciting opportunities here in the electronic medium; one can envisage different styles of entry geared to different user functions. For the time being, it is fair to say that dictionary definitions generally cater quite well for their users' decoding needs, but that the task of encoding requires

[7] The *Longman Language Activator* (1993) is the only monolingual learners' dictionary of English explicitly designed for the encoding user; its entries go further than most in supplying productively relevant information (and in carefully disambiguating close synonyms) – but arguably this makes them less suitable for decoding purposes.

access to so many types of information that it is not reasonable to expect a mainstream dictionary (even a learners' dictionary) to be adequate in most cases.

10.4.3 Usability: who are definitions for?

Consider the definition in Figure 10.10.

> **catkin** *n* a spicate inflorescence (as of the willow, birch, or oak) bearing scaly bracts and unisexual usu. apetalous flowers
> *MWC-11* (2003)

Fig 10.10 A definition of *catkin* from a dictionary for adult native speakers

This is from a dictionary aimed mainly at the adult native speaker; a definition like this would obviously be unsuitable for anyone with limited competence in English. But its style and content raise a number of questions about the kind of user a definition like this might be designed for:

- Is the user a specialist, or an 'average' reader? The definition would be useful for someone with a reasonable knowledge of botany, but for the layperson its technical character is likely to be problematic.
- How familiar is the user with the lexicographic conventions employed here? For example, the expression 'as of' (in the formula 'as of the willow . . .') and the abbreviation 'usu.' are far from typical of general English discourse; while this meaning of 'bearing' ('bearing scaly bracts') is a marginal use.[8] Will this present obstacles to understanding?

A definition may score well in terms of the adequacy and technical accuracy of its content. But this can't be the only criterion by which we judge its effectiveness. The user is (or should be) the central actor here, and whatever information the definition sets out to supply must take account of the user's prior knowledge, linguistic competence, and understanding of reference conventions. It is difficult to legislate on matters of definition content because, even when the information supplied works well for the decoding user, it may prove less than adequate for encoding purposes. We will come

[8] When *bearing* is a verb form, about 25 per cent of occurrences appear in the expression 'bearing in mind'; instances of 'child bearing' (or 'bearing children'), 'weight bearing', and 'bearing the name (of)' are also common – but in the sense used here it is rare.

back to content issues shortly (§10.5). But regardless of content, we are already in a position to lay down some basic requirements with regard to *form* – the words and structures used for conveying information.

➔ Definitions must be intelligible, and intelligibility requires – as a minimum – that:

- The language used should be appropriate to the linguistic skills, and presumed technical knowledge, of the user.
- If the definition includes words that are polysemous (and most definitions do), they should not be used in senses which are marginal or atypical.
- The user shouldn't have to consult *another* definition in order to understand the one s/he is looking up (this won't always be feasible, but it's a desirable objective).
- The wording and structures of the definition should conform as far as possible to 'normal' prose, and should not oblige the user to learn a set of lexicographic conventions (especially conventions that are unique to one particular dictionary).

It hardly needs saying that this definition for *catkin* performs badly on all these measures. Words like 'apetalous' and 'unisexual' are likely to be unfamiliar, but at least an educated adult reader should be able to guess what they mean. Not so for 'bract', 'inflorescence', or 'spicate', which are highly technical terms (and rare, too: 'spicate', for example, has no hits at all on the BNC). To understand the definition, the non-specialist user would have to look up other words (and this sometimes raises more questions than it answers). And as we noted earlier, the definition employs lexicographic conventions which can't be assumed to be familiar to the average user. The definition is concise and (we assume) accurate – but these virtues do not override the need for intelligibility. Whatever its other merits, if a definition cannot readily be understood by its intended user, it has failed. Or to put it another way, if the user can't understand the definition, the fault is not the user's but the dictionary's. By contrast, Figure 10.11 gives another definition for the same word.

catkin noun a downy, hanging flowering spike of trees such as willow and hazel, pollinated by the wind (*ODE-2*)

Fig 10.11 Another definition of *catkin*, from a different dictionary of the same type

In this case, the definition looks like regular English prose, and none of the words it uses is likely to cause problems for an educated reader. It could be argued that the first definition is more technically precise (compare 'flowering spike' with 'spicate inflorescence'), but this brings us back to the dictionary's function. A quick look at the sources in which *catkin* occurs in the BNC shows that most are either 'imaginative' texts (fictional descriptions and the like) or texts about gardening (and occasionally birdwatching) aimed at a lay readership, whether in books, magazines, or newspapers. In other words, this is where the average reader is most likely to come across the word *catkin*. It will also, no doubt, appear in textbooks about botany and ecology (which are scarce in the BNC), but definitions in general-purpose dictionaries are not usually geared to the needs of subject-specialists.

To conclude this introduction to the principles of defining, it will be useful to set out some basic requirements for a good definition. Intelligibility should be regarded as a given. What else should a definition achieve?

- As a minimum, the definition should supply enough information to enable the user to understand the word in the context in which s/he has encountered it.
- It should also enable the user to interpret the word successfully in any new context (so that the word enters the user's passive vocabulary).
- Ideally, it should enable him or her to use the word, correctly and appropriately, in a new context (so that the word enters the user's active vocabulary).

10.5 Definitions: content

The definer's first decision is: what should I say about this word (or to be more precise, about this lexical unit, or LU)? From every possible observation that could be made about the ways in which a given LU contributes to the meaning of its context, which will be of most value to the user? There is no single right answer to this question. The short definition of *cattle* we saw above ('Cattle are cows and bulls') is simple in two ways: it is expressed in simple language, but it also provides simple information – the minimum necessary for understanding the concept. But this definition is designed for learners of English at a low level of proficiency, so – as a

guide to understanding the kinds of text this user is likely to come across – it should do its job perfectly well. For other kinds of users and uses, we will need a different collection of facts about *cattle*. So how do we know which facts to select?

Readers who have got this far won't be surprised to learn that the chief factors in this selection process are the type of dictionary, and the needs and expectations of its users. In this section we discuss the issues relating to definition content, and show how a clear understanding of the target user will help us decide on the kinds of information our definition should include. But it will be useful, first of all, to set this discussion in the context of theoretical ideas about defining and the changes these ideas have undergone in recent years.

10.5.1 Content – the traditional model

In Chapter 8 (§8.3.1) we saw how, in classical semantics, the process of identifying senses was underpinned by the notion that a discrete meaning can be identified through a unique set of 'necessary and sufficient conditions'. These in turn would – in this traditional model – supply the content of each sense's definition. A definition composed in this way typically consists of two elements:

- a superordinate word or expression, which locates the item being defined in the right semantic category
- additional information which indicates what makes this item unique and in what ways it differs from other members of the same category (its cohyponyms).

These two elements are usually referred to, respectively, as the 'genus' (or 'genus expression') and the 'differentia' (or 'differentiae' if there are several distinguishing features). Thus a *convertible* is defined in *ODE-2* (2003) as

> car with a folding or detachable roof

'Car' is the genus, and the differentia – which distinguishes a *convertible* from a *saloon*, *estate car*, or *people carrier* – is the postmodifying expression 'with a folding or detachable roof'. Following the same process, the genus here ('car') is itself defined by another genus ('vehicle'), and then differentiated from other members of the 'vehicle' category by its own distinguishing features (cf. the discussion of hyponymy in §5.2.1).

This is an effective defining strategy in many cases, and we saw earlier (§10.1.3) that template entries for monolingual dictionaries almost always specify one or more 'approved' genus expressions. The approach has obvious similarities with the Linnaean taxonomies used for classifying and identifying plants, animals, and other living things. Not surprisingly, definitions of items like these work especially well within the genus-and-differentia model. But many other types of word can be successfully defined using the same strategy. These include not only most kinds of noun (especially those referring to concrete objects), but also many classes of verb, including verbs of motion (thus *trudge*, *tiptoe*, and *stroll* can all be defined using the genus 'walk'), verbs of making or creating (*reproduce*, *photocopy*, and *forge* can all be defined with the genus 'copy', which is itself a hyponym of 'make' or create'), and several others.

When this approach is applied in a definition, you need to take care in selecting the most appropriate genus expression. Sometimes there is more than one possible candidate. The definitions in *ODE* and *MWC* for the intransitive use of *negotiate* have more or less the same content, but the choice of genus determines the definition's primary focus. The *ODE* definition focuses on the *goal* of negotiating, while the one in *MWC* foregrounds the *process* (the genus expression is underlined):

> to <u>try to reach an agreement or compromise</u> by discussion (*ODE-2* 2003)
>
> to <u>confer with another</u> so as to arrive at the settlement of some matter (*MWC-11* 2003)

Both definitions work well, and there is no obviously preferable genus. But there are other cases where a dictionary chooses the wrong semantic component to focus on, as for example in this definition of *whistle* (in the sense of 'bullets whistling past my head'):

> produce a high-pitched sound by moving rapidly through the air or a narrow opening (*ODE-2* 2003)

Clearly the verb includes elements of both sound and movement, but which is more important? A look at the data helps to steer us in the right direction. The verb is typically followed by a particle. The particle occasionally indicates location (as in 'the wind whistled in the trees'), but there is a strong preference for particles showing direction:

A mortar bomb burst . . . sending shrapnel whistling ***through*** *the trees and thudding into the walls of the little cottage.*

Chen, always the skilful extrovert, responded with a 'cartwheel' as the ball whistled ***past*** *him.*
Winds of a hundred miles an hour or more roared and whistled ***round*** *the isolated house. As the door swung back behind her, an icy draught whistled* ***into*** *the room from the black corridor beyond.*

With particles like these being favoured in most cases, it is clear that we are dealing with a verb of motion (cf. Atkins and Levin 1988). A definition which foregrounds that semantic element provides a better basis for interpreting sentences like these – and this means selecting a motion verb as the genus, as in this definition:

> to move, go, pass etc. with a whizzing sound, as a bullet, the wind etc. (*Random House Dictionary of the English Language*, Unabridged, 1967)

Provided the genus matches what the data tells us about the word's behaviour, the genus-and-differentia format will often be an effective defining strategy. But it is important to recognize that large areas of the lexicon don't fit this taxonomic model, and cannot sensibly be explained in these terms. This applies not only to most adverbs (and to practically all of the rarer wordclasses), but also to a majority of adjectives: as we saw earlier (§5.2.1), few adjectives belong to hierarchies of hyponymy, so finding appropriate genus expressions is often impossible. (For more on adjectives, see §10.6.4.1 below.) So don't to try to force definitions into this mould simply in order to achieve an illusion of 'consistency'.

10.5.2 Problems with the traditional model

If the genus-and-differentia model is sometimes unworkable, the traditional goal of identifying an LU's 'necessary and sufficient conditions' is even more questionable. It is well adapted for describing concepts whose boundaries are distinct and whose essential attributes are easily identified, invariable, and not in dispute – words such as *passport*, *tsarina*, *encrypt*, *intubate*, *heterosexual*, and *instantly*. But the lexicon is full of words encoding concepts which don't conveniently resolve themselves into a precise and finite set of features (cf. §8.3.1). Misguided efforts to apply the traditional model in all cases can lead to definitions which try to account for every conceivable instantiation of the concept being defined. This tends to have one of two outcomes:

- long and detailed definitions, which attempt to list every possible feature or every imaginable instance
- short and vague definitions, which avoid specifying any features that might exclude valid members of the category being defined.

As an example of the first problem, it is difficult to beat the now famous definition of *door* in *Webster's Third International* (1961).[9] In eleven densely packed lines of text, the definition goes to extraordinary lengths to describe every entity that could conceivably be called a door, with information about its structure, its various functions, mechanisms for opening and closing, and the varieties of space it provides access to. But a definition which attempts to match all possible exemplars is doomed to fail, because one can never predict every entity that might at some point be described as a 'door'. The second tendency is illustrated by this definition of the nominal use of *absolute*:

> Something that is absolute (*AHD-3* 1994)

The best that can be said of this is that it rules nothing out: any occurrence in text of *absolute* as a noun is, undeniably, fully covered by the definition. But by failing to give any indication of how the noun *absolute* is normally used, the definer is letting the user down. Defenders of this approach might argue that the definition is preceded by a full description of the many adjectival uses of *absolute*; all the user has to do is search the various senses of the adjective to find one that maps onto its use as a noun. But both these attempts at 'total accountability' – either by trying to say everything, or by avoiding saying anything – make unreasonable demands on the user. Worse than this, they fly in the face of what we know about how words encode meanings. The next section explains why, and suggests alternatives to these two approaches.

10.5.3 Prototypes and defining

The goal of analysing an LU into its 'essential constituents' has come under pressure from two convergent developments. On the one hand, cognitive science has shown that language-users develop and internalize a 'prototypical' version of a given entity or concept, and that they interpret

[9] Discussed by (among others) Hanks (1979).

individual instances by reference to this prototype (§8.3.1).[10] On the other hand, we have learned from corpus linguistics that meanings often have quite fuzzy boundaries, and that interpreting individual language events will sometimes require us to 'stretch' these boundaries a little (§8.2.3). Both developments draw us away from the (spurious) certainties implied by a traditional approach to 'defining', and take us into a messier, more relative world, in which we create 'explanations' that will enable users to interpret all but the most marginal uses of a word. Instead of trying to isolate necessary and sufficient conditions, we aim – by analysing many individual instances of our word in text – for a typification which will show the user 'what is normally the case rather than what is necessarily the case' (Hanks 1987: 118). Two examples will show how this approach works in practice. First, another definition of the noun *absolute*:

> a value or principle that is regarded as universally valid or which may be viewed without relation to other things (*ODE-2* 2003)

In this case, the data clearly shows that – when used as a noun – *absolute* almost always occurs in philosophical contexts, so it is perfectly reasonable to narrow the focus of the definition in this way. Other uses are so marginal that we don't need to try to account for them. This definition of *bus*, by contrast, shows how it is possible to convey an idea of the prototypical bus while leaving open the possibility of variations from the norm:

> a large motor vehicle carrying passengers by road, typically one serving the public on a fixed route and for a fare (*ODE-2* 2003)

The first half of the definition contains the most essential information about what a *bus* is: it's large (so it's not a *car* or *taxi*), it goes by road (so it's not a *train*), and it carries passengers (so it's not a *truck*). The second half explains the features of a *typical* bus: it operates as a public service on a fixed route and charges its passengers a fare. We know that, in the real world, there are also things like school buses, hotel buses that pick up guests from airports, buses that take airline passengers from terminal to plane, and buses for transporting employees to or around a large business campus. Buses like these are not for the general public, and don't usually charge a fare. But the definition doesn't exclude them, and it provides enough information to enable users to interpret such minor variations from the prototype.

[10] Interestingly, the *MW-3*'s notorious definition of *door* is accompanied by an illustration showing an absolutely prototypical door.

Prototype theory encourages an approach to defining which recognizes the inherent variability and lack of precision in human communication. Claims that our definitions specify – in an 'authoritative' way – the essential characteristics of a given LU are likely to prove unsustainable in the face of observable language data. So we should settle for the less ambitious but more realistic goal of abstracting, from a mass of individual instances, the central and recurrent semantic features of a word or LU and, when appropriate, providing additional information that will help users to identify prototypical members of the category.

10.5.4 Content and the user: does more mean better?

The dramatic contrast between the pair of entries in Figure 10.12 (taken from different editions of the same learners' dictionary) is instructive.

raft[2] *v* **1** [X9] to carry (something) on a raft (somewhere): *raft the stores over to the island* **2** [X9] to send (wood) in the form of a raft (somewhere): *raft the logs down the river* **3** [T1] to cross (water) on a raft: *They rafted the lake*. **4** [L9] to travel (somewhere) on a raft: *They rafted down the river to New Orleans*.

LDOCE-1 (1978)

raft[2] *v* [I,T] to travel by raft or carry things by raft

LDOCE-3 (1995)

Fig 10.12 Two entries for the verb *raft*

Strictly speaking, the two entries exhibit 'lumping' and 'splitting' approaches to the senses of *raft* (cf. §8.1.3), but the differences also reflect decisions about content. The shorter entry accounts for the same uses as the longer one, but with far less detail. It's unusual for a later edition to say so much *less* about a word than an earlier one, but the arrival of corpus data (between the two editions) will have influenced this adjustment to the original version. We know that people's 'active' vocabulary – the words and meanings they use productively – is generally far smaller than their passive vocabulary.[11] A corollary of this is that there is a rough correlation between a word's frequency and its likelihood of being used productively (for encoding: §10.4.2.2). Corpus data tells us that *raft* is a very rare verb (occurring

[11] See for example I. S. P. Nation, *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press (2001), esp. chapters 2, 5.

less than once per 10 million words), so it's reasonable to conclude that advanced learners will rarely if ever need to encode it. A short, simple description is therefore perfectly adequate.

➔ The shorter entry is not only serviceable and 'fit for purpose': it is also *better*, because it is better adapted to the needs of its intended user.

The first entry is both complete and accurate, but its complexity is inappropriate and the space it requires can't be justified in terms of user need. Remember, too, that the longer the entry, the greater the processing demands it imposes on the user (in this case a language-learner). The moral here is that it is important to distinguish between information which is true, and information which is relevant. Landau (2001: 170) quotes Richard Robinson as saying that 'A lexical definition could nearly always be truer by being longer', but Landau also makes clear, rightly, that a definition's 'truth-value' is not the only factor that determines its success. We should always start from the kind of comprehensive description that a well-populated database will include, but a good definition is one that meets its user's needs without providing more information than is necessary.

10.5.5 Defining by synonym: pros and cons

In many dictionaries, definitions often consist of one or more synonyms, and on the face of it this looks like an economical way of conveying content. In the entry for *keen* in Figure 10.13 (from a dictionary for adult native speakers), each of the four senses shown is defined solely through synonyms.

keen . . . **4.** Sharp; vivid; strong: *"His entire body hungered for keen sensation, something exciting"* (Richard Wright). **5.** Intense; piercing: *a keen wind*. **6.** Pungent; acrid: *A keen smell of skunk was left behind.* **7.a.** Ardent; enthusiastic: *a keen chess player*
AHD-4 (2000)

Fig 10.13 Synonym definitions for *keen*

Sense 6 is probably the most successful of these definitions, because the two synonyms offered are themselves (more or less) monosemous. The message is this: one of the less central meanings of *keen* equates to the most central (or only) meaning of *pungent* or *acrid*. Provided the user knows these two words (a reasonable assumption here), the definition will at least

be adequate for decoding. But in most cases this is an unsatisfactory way of defining. Sense 4, for example, is hopelessly ambiguous: *sharp* and *strong* are highly polysemous words, and the meanings referred to here are far from central. The strategy is slightly more effective when a synonym is supported by a note about selectional restrictions (§8.5.2.2), thus reducing the risk of ambiguity. The entry for *keen* in *Chambers 21st Century Dictionary* (1999) includes statements like these:

> **2** said of competition or rivalry, etc.: fierce **3** said of the wind: bitter **4** said of a blade, etc.: sharp

But there is a fundamental objection to defining by synonym, namely, that no two words are exactly alike. True synonymy (what Cruse calls 'absolute synonymy') entails complete interchangeability in every possible context of use. There are plenty of words and phrases which have the same meaning as *die* (from *expire* at one end of the register scale, to the unmentionable idiom involving buckets at the other), but none is equally appropriate in all situations. Hence, 'absolute synonyms are vanishingly rare, and do not form a significant feature of natural vocabularies' (Cruse 2004: 155).

➔ Using a synonym definition is only really acceptable when the *definiendum* and the synonym are semantically identical, and any points of difference are in the area of register, regional distribution, speaker attitude, and the like.

We mentioned the word *dosh* in the context of labels (§10.3.2), and there is no reason not to define it simply by its synonym 'money' – provided we also supply relevant information about its use. It is reasonable, too, to adopt this approach in the case of technical and non-technical pairs (like *patella/kneebone*, *myocardial infarction/heart attack*).[12] But apart from situations like these, using synonyms as the sole indicator of meaning is not an acceptable substitute for serious semantic analysis. Synonyms can have a useful complementary role, when supporting a 'full' definition, as in one of the senses of *sharp* in *MWC-11* (2003):

> clear in outline or detail: DISTINCT

(as in *a sharp image*). In this case a user who has processed the definition may feel reassured to find a familiar word that consolidates what s/he has learned. But defining solely by synonym – though a tempting option

[12] In similar vein, Zgusta notes that 'obsolete, dialectal, colloquial, vulgar etc. lexical units can be explained in this way, provided there is a connotatively neutral synonym' (1971: 262).

because these are the easiest kinds of definition to write – should be avoided except in the specific situations described above.[13]

10.5.6 Extralinguistic messages: pragmatics, sensitivities, connotation

If someone asks you a question that you can't answer, you may give a neutral response (*I don't know*) or a more emphatic one (*I have no idea*). But if you say *How should I know?*, you are doing more than just explaining that you have too little information to give a helpful answer. This expression encodes not only a meaning but an attitude: it conveys a sense of irritation, and suggests that the questioner ought to realize that you are the wrong person to ask. While expressing meanings, speakers also routinely communicate their feelings and opinions through the words and phrases they select. And corpus data shows that speaker attitude tends to be conventionally lexicalized in a fairly limited number of frequently occurring words and phrases. This is the domain of pragmatics, an aspect of socio-cultural competence, and it poses interesting challenges for the lexicographer. Pragmatics is a big field,[14] but our focus here will be on those lexical items in which the feelings of the speaker or writer (conveyed with varying degrees of directness) are at least as important as the semantic content. We will also, in this section, look at other cases where a word or phrase encodes 'extralinguistic' information: where the message it conveys is not fully retrievable from its semantic content alone.

10.5.6.1 *Handling speaker attitude* The word *typical* – when used to indicate that something is characteristic of the way someone usually behaves, or of the way things usually happen – can have a positive or negative 'spin' according to context. For example:

How typical of him to turn it back on her and make it seem as if she was at fault.
She says he was doing what he wanted to do. It was typical of him. He was a soldier and he wouldn't have wanted to die any other way.

[13] Not surprisingly, definitions in the collaborative *Wiktionary* project – an offshoot of *Wikipedia* – depend heavily on synonyms, as in this definition of the first meaning of *absorb*: 'To include so that it no longer has separate existence; to swallow up; to engulf; to overwhelm; to cause to disappear as if by swallowing up; to use up; to incorporate; to assimilate'.

[14] The classic text is Stephen C. Levinson's *Pragmatics*. Cambridge: Cambridge University Press (1983).

But as we saw earlier (§8.4), a specific permutation of syntactic and colligational features can sometimes encode a more specialized use. The examples above illustrate a common pattern used with this adjective: *it is typical + of sb* (+ *to so sth*). But when *typical* appears without a following preposition, at the end of a clause or sentence (or on its own, in the form of an interjection), and sometimes also with modifiers like *just*, *absolutely*, or *so*, it invariably expresses criticism or exasperation:

'Now isn't that just typical?' Polly's mouth curled. 'Blaming me for your own inadequacy'.

Allan was due to run in Zurich but he didn't appear, which was absolutely typical.

This is so typical. Every time you have problems you run off travelling without facing the situation.

In similar ways, we can choose to express a meaning using words or expressions which also convey attitudes such as admiration (e.g. *tireless*, *understated*), amusement or irony (*princely sum*, *bright spark*), and contempt (*nerd*, *busybody*, *ingratiating*). Very often, the effects are counterintuitive: thus *mischievous* and *rogue* may be terms of endearment, *pious* and *do-gooder* (what could be wrong with 'doing good'?) are generally disparaging, and if you preface a remark with *with the greatest respect*, you are probably signalling profound (and not very respectful) disagreement.

To deal with items like this in the dictionary, we have a range of strategies at our disposal. Sometimes a simple style label will be adequate. The inventory of labels used in *ODE-2* (2003) includes *humorous*, *derogatory*, and *euphemistic*, and other dictionaries have similar descriptors:

> **tireless** adj *showing approval* working very hard without stopping: *a tireless worker* |*I am very grateful for your tireless efforts.* (*MED-2* 2007)

In other cases, you can phrase the definition in a way that makes the speaker's attitude clear (so that no label is needed). This definition of *do-gooder*, for example, begins with a neutral description, then adds information about how such behaviour is perceived by others:

> **do-gooder** *n colloq.* A person who actively tries to help other people, *esp.* one regarded as unrealistic or officious (*Shorter Oxford English Dictionary*, 2nd Edition 1993)

You can also use what we described earlier as a 'pragmatic force gloss' (§7.2.3.3), a common device for conveying a word's pragmatic message. This typically takes the form of a phrase like 'used for showing . . . ' or 'used when

you . . . ', and it can appear either on its own or tacked on to the end of a standard definition:

> **come off it** (used when you think that what someone has said is definitely not true, and that they probably do not believe it either) (*Longman Language Activator* 1993)
> **do-gooder** someone who helps people who are in bad situations, but who is annoying because their help is not needed – used to show disapproval (*LDOCE-4* 2003)

Where the dictionary user is from a different culture, you may sometimes need to use a range of strategies to explain the full socio-cultural significance of a word or phrase. The way English speakers use *bourgeois* (in its evaluative rather than classifying sense) is a case in point:

> **bourgeois** adj **1** *showing disapproval* typical of middle-class people and their attitudes. This word often shows that you dislike people like this, especially because you think they are too interested in money and possessions and in being socially respected (*MED-2* 2007)

This may look like overkill, but this is an entry designed for learners of English, and the concept it lexicalizes could well be unfamiliar to people from different cultures.

The 'full-sentence' style of definition (which we discuss more fully below: §10.6.3.1) offers another way of dealing with items of this type. As this entry for *typical* illustrates, it allows you to place the main focus of the definition where it belongs – on the feelings of the speaker who selects this word:

> **typical . . . 3** If you say that something is **typical** of a person, situation, or thing, you are criticizing them or complaining about them, and saying that they are just as bad or disappointing as you expected them to be. (*COBUILD-5* 2006)

Perhaps the most difficult challenge is knowing when it is appropriate to deploy one or more of these strategies. Two criteria will help to clarify things:

(1) The surface meaning is at odds with the intended message: there is no need to use a *disapproving* label with words like *lousy*, *idiotic*, or *disgusting*; the definition tells you all you need to know. But in many of the cases mentioned above (such as *do-gooder*, *princely sum*, *rogue*, or *with respect*), a literal reading gives no clue as to the intended message, so it makes sense to explicate this further.

(2) Recurrence: as always, we are looking for language events that are frequent and well dispersed. It's a fact of life that almost anything we say can, in context, convey a pragmatic message. Expressions like 'That's really helpful' or 'What a great movie!' could easily be ironic in some circumstances, but as lexicographers we need to focus on what speakers and writers do regularly. And, as the case of *typical* showed, a significant shift in usage is often reflected in more specialized patterning.

10.5.6.2 *Sensitivities: insulting or offensive language* Part of a word's 'meaning' is the effect it has on the listener, whether intended or not. A dictionary owes it to its users to give a clear account of the sensitivities that attach to a given word or expression, and the need is especially acute in the case of dictionaries for learners. Users may have different socio-cultural norms from those of the speech community they aspire to communicate with, and a good definition is one that will help them avoid embarrassment. Special caution is needed in the case of words referring to:

- ethnic or racial origin
- disability
- sexual orientation
- age
- gender.

The same set of strategies used for explaining a word's pragmatic message (labels and pragmatic glosses, as well as the wording of the definition itself) is available for dealing with insulting or offensive language – as we saw in the entry for *half-caste* above (§10.3.2). And if the definition warns the user (implicitly or otherwise) to avoid the word being defined, it is helpful to supply a more appropriate alternative:

> **Siamese twin** *old-fashioned* one of two people born with their bodies joined together. This is now considered offensive and it is more polite to say **conjoined twin**. (*MED-2* 2007)

There is a delicate balance to be struck between helpful warnings and excessive caution. As Landau notes, dictionaries were once rather slow to recognize the sensitivities attaching to some words (cf. the unlabelled entry for *half-caste* in the 1974 edition of *OALD*: §10.3.2), but some now risk looking overzealous. For example:

> **crone** an offensive term that deliberately insults a woman's age, appearance, and temperament (*offensive)* (*Encarta World English Dictionary* 1999)

Here the tone of censure (with two mentions of 'offensive' and one of 'insult') so overwhelms the semantic content that one is left wondering: 'Yes, but what does it *mean*?'[15] When the same dictionary applies the *offensive* tag to words like *screwed up*, *klutz*, and *crazy*, there is a danger of the label becoming devalued. As Landau wryly observes, this rather po-faced approach 'views the language as a fortified castle of virtue, and every battlement is equipped with a cannon loaded with warnings' (2001: 234). It is important to recognize, though, that there may be variations across a speech community in the 'thresholds' for what is deemed offensive. Thus, an epithet or expletive regarded as (at worst) mildly offensive in Australia may be virtually taboo in the US, so it is useful to get advice from members of the relevant group. Remember, too, that this is an area where sensitivities change over time and the language reflects this: the word *coloured*, for instance, was widely used in the first half of the twentieth century, as a 'polite' alternative to *negro* (or worse), but in contemporary English it is seen as demeaning. One needs to tread carefully here, and in pedagogical dictionaries, at least, it is better to err on the side of caution rather than expose users to the risk of causing offence and embarrassment.

10.5.6.3 *Connotation and cultural associations* In *Webster's Third*, the word *champagne* is defined like this:

> a white sparkling wine that undergoes one fermentation in a cask and a second fermentation in a bottle, the latter generating carbon dioxide that makes the wine sparkle

Accurate and informative though this is (there is enough information here to enable you to make your own champagne), the definition tells us nothing about the word's connotations. Two papers on this subject (Bullon 1990 and Stock 1992) refer to champagne and the extralinguistic notions it evokes: luxury, hedonism, and celebrations. Champagne may not be the best illustration of the problem (its cultural associations are more or less universal), but the point is an important one, especially for writers

[15] *Crone* is unlabelled in the two great historical dictionaries of English, *OED* and *Webster's Third*.

of pedagogical dictionaries. Many words refer to objects and institutions which are loaded with cultural associations. These associations form part of the native speaker's knowledge about the word, but they are 'culture-bound and cannot be conveyed by means of a standard dictionary definition' (Bullon 1990: 27).[16] Both authors make a good case for including connotation as an element in definitions, and some pedagogical dictionaries have gone quite far along this road.[17] This is a well-motivated idea, though difficult to apply in practice since so many vocabulary items carry cultural associations of some kind. Bullon, for example, mentions *darts*: a simple definition of the game doesn't account for its strong associations, for British people, with pubs and with 'fat men who drink a lot' (ibid.: 28). But how far can one go? This kind of cultural knowledge would be invaluable for a visitor studying or working in the UK. But the majority of non-natives learn English because it is a lingua franca (not because of a deep interest in British culture), and most of their communication will be with other non-natives. The best advice we can give is to be alert to any word whose connotative features are so central to its 'meaning' that the word cannot truly be understood without reference to these associations. In cases like this, it is worth adding a sentence to the basic definition, as in this entry for *caviar*:

> fish eggs eaten as food, usually spread on bread. In many countries caviar is considered to be a special and expensive food, eaten mainly by rich people. (*MED-2* 2007)

10.5.7 'Editorializing' and the myth of neutrality

Lexicographers, like historians, are expected to be 'neutral' recorders of facts – but this isn't as straightforward as it sounds (in either discipline). Departures from lexicographic neutrality are characterized as 'editorializing', which is seen as a reprehensible tendency. In this context, people

[16] Strictly speaking, it is less a question of being a non-native speaker than of not belonging to a particular speech community: thus for speakers of British or Indian English, the full resonance of American words like *homecoming queen* or *Greeks* (referring to members of fraternities and sororities) is just as hard to retrieve as it is for non-native speakers.

[17] Most notably the *Longman Dictionary of English Language and Culture*, whose main objective is to explain the cultural associations of words to non-native learners.

often mention entries from Johnson's *Dictionary* that flout the embargo on editorializing, like his well-known definition of *patron*:

> One who countenances, supports or protects. Commonly a wretch who supports with insolence, and is paid with flattery.

Johnson tends to be indulged as an exception to the general rule, and we can all agree that using a dictionary to pursue personal vendettas isn't a good idea. But there are plenty of genuine grey areas, where the choices available to the definer are not between subjectivity and impartiality, but rather between two different forms of partiality. The situation is especially sensitive in the areas of belief systems, whether religious or political, and the structures and institutions they support.

Consider the following definitions of *apartheid* (which were both written during the apartheid era):

> the keeping separate of races of different colours in one country, esp. of Europeans and non-Europeans in South Africa
>
> (*LDOCE-1* 1978)

> (in South Africa) the system established by government of keeping different races separate, esp. so as to give advantage to white people
>
> (*LDOCE-2* 1987)

On the face of it, the second definition looks more tendentious, because it imputes a political agenda (advantaging the white community), while the first appears simply to state the facts of the case. But the counter-argument is that, by failing to say anything about the motives behind apartheid, the earlier definition implicitly endorses those who advocate this system. The reality is that it is sometimes impossible to avoid taking a stance, and we need to be honest enough to admit this. It is a commonplace that 'one person's freedom-fighter is another person's terrorist', but language is a powerful political weapon, and the appropriation by certain elements of words like *terrorism* and *extremist* puts lexicographers in an invidious position. Impartiality is a good aspiration, but it's important to recognize that a dictionary will inevitably reflect the values of the culture from which it springs. Think of the idea of 'freedom of speech': it is suppressed in many countries, and openly rejected by some religious groups (or should that be 'extremists'?). But if you look at English corpus lines for *freedom of speech*, you find that people 'fight for freedom of speech', they work to 'uphold' it, they resist 'encroachments' or 'attacks' on it, and they 'exercise'

it 'fearlessly'. It doesn't take long to realize that freedom of speech is assumed, in English-speaking cultures, to be an unambiguous good, and any definition will inevitably reflect this position.

As Swanepoel observes (2006: 1272), all languages include words denoting concepts which 'can only be defined relative to some other larger meaning construct', and he considers ways in which dictionaries 'encode the concepts of relative existence and relative truth'. One of his examples is this definition of *sangoma*:

> An African witch doctor, usu. a woman, often claiming supernatural powers of divination
>
> (*Dictionary of South African English* 1996)

Here the sangoma's powers are 'claimed' rather than actual, but it is easy to see that we are entering dangerous territory. Few readers would have a problem with a definition of the *tooth fairy* which described it as 'an imaginary being'. But if we were to follow Richard Dawkins and define *Allah* in the same way, all hell would break loose (whether or not we believed in 'hell'). Fillmore (2003: 278ff.) discusses the pitfalls that lie in wait for anyone defining religious terms. He compares (for example) these two definitions of *reincarnation*:

> the belief that on the death of the body the soul transmigrates to or is born again in another body
>
> (*CED-1* 1979)

> Rebirth of the soul in another body (*AHD-4* 2000)

The first definition describes reincarnation as a 'belief', the second straightforwardly equates it with 'rebirth of the soul' – just as a definition might equate a car with a 'small road vehicle'. Thus each entry (consciously or otherwise) adopts a particular stance with regard to the concept being defined. But as Fillmore makes clear, there is no 'neutral position':

> If atheists read a definition of **God** as 'the Supreme being who created and maintains the universe', they could complain that the producers of the dictionary are using language that presupposes something they find objectionable.

Similarly, definitions of *creation science*, *climate change*, or *astrology* are unlikely to achieve complete neutrality.[18] And as with words that give

[18] The collaborative 'Urban Dictionary' project, to which anyone can submit their own definitions, provides some striking illustrations of the problem. See for

offence (§10.5.6.2), we need to be alert to the fact that the epistemological framework can vary over time as well as place. Veisbergs (2002: 659) quotes an extraordinary definition of *imperialism* from the 1948 edition of Hornby's *Learner's Dictionary of Current English* (the ancestor of the *OALD*): it carries none of the negative connotations that word has in contemporary discourse, but even at the time when it was written, it is unlikely to have had much resonance with the liberation movements then emerging in Africa. It should be clear that 'neutrality' isn't always possible, so it is important first to be aware of the belief system in which we are operating (and its possible impact on the 'stance' of some definitions), and secondly to react to changes in the real world as they occur.

10.5.8 Definition content: conclusions

We return later (§10.7) to the bigger question of 'what makes a good definition'. But before we move on to discuss the *form* of definitions, it is worth drawing some interim conclusions on issues of content:

➔ Don't attempt to account in your definition for every conceivable instantiation of the LU.

This means avoiding the extremes of:

- telling the reader everything you know about the concept being defined; lengthy definitions make unreasonable demands on the user (§10.5.4), and it is your responsibility to distinguish what is marginal from what is central
- making the definition so vague and underspecified that, even if it does match every corpus instance of the LU, it fails the reader by not saying anything useful (§10.5.2).

➔ The search for 'necessary and sufficient conditions' (or 'criterial' features) is worth making in the first instance, and will support a good definition in many cases.

But the definition must reflect the variability evidenced in the data, and you will often find that a 'prototype' approach does this more effectively; a good strategy here is to use a basic definition to describe what is invariably

example the 'rival', no-holds-barred definitions of *intelligent design* or *red-stater*: www.urbandictionary.com .

true, then add a clause beginning 'typically' or 'especially' to describe the prototypical case (cf. the definition of *bus* in §10.5.3).

➔ Make sure your definition accounts, when appropriate, for the kinds of extralinguistic features we described in §10.5.6.

➔ Above all, consider what your user *needs to know*.

Selecting facts for a definition is not a 'scientific' enterprise: the content of a definition will vary (quite properly) from one dictionary to another. A scholarly dictionary may tell us that there are numerous species of *shark*, that they vary in length from 25cm to 10 metres, and that many are entirely harmless. But in a dictionary designed for children or language-learners, the definition will rightly focus on 'prototypical' sharks, which are large and often dangerous animals. In cases like this, technical accuracy and completeness are not the key goals. Rather, the definition needs to 'enable the dictionary user to identify the concept in question, that is, to retrieve it from his/her own conceptual memory' (Geeraerts 1990: 196).

Dr. Johnson grappled with the problem of knowing what to say about words, and his conclusions – as always – are worth quoting:

> It is not to be expected, that with the explanation of the one [*baronet*] the herald should be satisfied, or the philosopher with that of the other [*barometer*]; but . . . it will be required by common readers, that the explications should be sufficient for common use
>
> (Johnson, *Plan*, 1747)

10.6 Definitions: form

Once you have a clear idea of your definition's content – the information you want it to convey – your next task is to decide on its form. The form of a definition is the language used for encoding its content, and this includes grammatical structures as well as words. We made the point earlier (§10.4.3) that, regardless of the adequacy and accuracy of their content, definitions can never be effective if they are unintelligible to the intended user. Intelligibility entails making the right choices about the language you use, and in this section we discuss the factor that influence these choices. But it is important to be aware, too, that there is a considerable body of principles and conventions that inform the process of writing definitions. These have developed over the past two centuries or so, and many of them are still widely used, so this is an appropriate point to review them.

Box 10.1 Defining in dictionaries: a brief history

In the earliest English dictionaries – from Cawdrey's *Table Alphabeticall* (1604) to Johnson's *Dictionary of the English Language* (1755) – defining styles had not yet been standardized and were quite heterogeneous. It is not unusual to find definitions which don't match the wordclass of the headword, or which take the form of complete sentences. For example:

> **homonimie** when divers things are signified by one word (Cawdrey 1604)
> **TRANSCENDENTAL** *Curves* [in the *higher Geometry*] are such as cannot be defined by Algebraical Equations, or . . .
>
> (Nathan Bailey *Dictionarium Brittanicum* 1721)

In this early period, dictionaries made little claim to 'authority', and, for all the ambition that motivated his original *Plan of a Dictionary* (1747), Johnson ended up with a realistic appreciation of the limits of lexicography, and he saw his task as a practical one.

As time went on, a consensus developed about what the scope and function of a dictionary should be. Dictionaries now aimed (unlike Cawdrey, for example) to cover the whole of the lexicon, not just a subset, and (following Trench's characterization of the lexicographer as 'an historian, not a critic': §3.3.2.2), lexicographers increasingly saw themselves as descriptive linguists, rather than prescriptive 'authorities'. (This didn't stop dictionary users ascribing 'authority' to their dictionaries, however, nor dictionary publishers from claiming it.) At the same time, lexicographers sought to bring a degree of system and consistency to their definitions, while maximizing the value of the limited space available. These two goals combined to spawn a repertoire of defining styles, many of them still widely used. But in the words and structures they employed, these conventions increasingly departed from the norms of general discourse, leading to what some commentators call 'lexicographese'.

From the 1960s, we see a gradual move away from conventional models towards styles that are both more explicit and more user-friendly. Two new entrants to the market, the *American Heritage Dictionary* (whose first edition appeared in 1969) and the *Collins English Dictionary* (first published in 1979) made serious efforts to 'open up' definitions and explain meanings in more accessible ways. These two definitions of *decadent* nicely illustrate the differences between traditional and evolving approaches:

> **decadent 1** marked by decay or decline (*MWC-8* 1980)
> **decadent 1** characterized by decay or decline, as in being self-indulgent or morally corrupt (*CED-1* 1979)
>
> (*cont.*)

Box 10.1 (*Continued*)

The second definition avoids the conventional formula 'marked by' (hardly a central use of this verb), opting instead for the more explicit 'characterized by'. But it also provides useful hints about the ways in which the 'decay or decline' referred to here usually (or prototypically) manifests itself. Meanwhile, monolingual learners' dictionaries (MLDs), which had originally adopted traditional defining practices to a surprising degree (Rundell 1988: 130–132), were beginning to develop distinctive styles of their own, rejecting 'lexicographese' and aspiring instead to a style of defining which, as far as possible, resembled ordinary prose; a key development here was the adoption of 'full-sentence definitions' (which we discuss later: §10.6.3.1). These innovations in defining practices were driven partly by the need to make definitions accessible to inexpert users, and partly by a desire to situate meanings in their typical contexts. To some extent, these newer defining styles fed back into mainstream dictionaries (as innovations in MLDs influenced practice in the wider lexicographic community), so that the more extreme manifestations of lexicographese have been toned down in most contemporary English dictionaries for native speakers.

10.6.1 Received wisdom: the principles and conventions of defining

It is worth knowing about the principles of defining in the English lexicographic tradition, not only because some of them still apply today but also because, *mutatis mutandis*, similar principles underpin lexicographic practice in other parts of the world. Many of the defining conventions which these principles gave rise to have been challenged or even – in some areas of lexicography – abandoned altogether during the last few decades. But an understanding of the principles that motivate the conventions gives us a benchmark against which to evaluate emerging alternatives to 'standard' defining styles.[19]

10.6.1.1 *Traditional principles of defining*

(1) Use simpler words than the word being defined: though clearly a useful guideline, this cannot – as must be obvious – apply in every case. The irreducible core of the language – the words we use for describing the fundamentals of human existence and our interaction with the

[19] Zgusta (1971: 257ff.) and Landau (2001: 157ff.) both provide useful discussions of the principles and conventions of traditional defining.

world – consists of a smallish number of high-frequency items which cannot usually be defined in terms simpler than the words themselves.

(2) Avoid 'circularity': roughly speaking, circularity is what happens when you define Y as X and X as Y. Here is an egregious example:

> **allow** *v*. **1** to let; permit
> **let** *v*. **1** to allow; permit
> **permit** *v*. **1** to allow; let
> (*Newbury House Dictionary of American English*, 4th edn 2004)

This may seem too obvious a principle even to need stating but, as we have just seen, some concepts are practically irreducible (and perhaps 'allow' is one of them). Consequently, a determination to avoid circularity at all costs may lead to definitions that are needlessly difficult. The following pair describe two such universal (and universally familiar) concepts:

> **father** a male parent of a child or animal
> **parent** a person's father or mother
> (*OALD-7* 2005)

Although these definitions are circular, most MLDs define *father* in much the same way, and this is a reasonable line to take: users understand the concept perfectly well – they simply don't know how it is encoded in English. A resolute effort to define without circularity seems to underlie the two definitions that follow, which nicely illustrate the limitations of this principle:

> **father** a man in relation to a child or children born from an ovum that he has fertilized (*OALD-5* 1995)
>
> **father** Your **father** is the man who made your mother pregnant with you (*COBUILD-2* 1995)

The first is far too difficult for its intended users; the second becomes more alarming the longer you look at it. The reliably contrarian Anna Wierzbicka believes that circularity is always avoidable, and that lexicographers are 'deceiving themselves' when they justify it as 'something that may bother theoretical semanticists but . . . will never bother the ordinary reader' (Wierzbicka 1993: 61). She follows a complex trail around no fewer than twelve entries to demonstrate that the entire set is ultimately built on the sands of circularity (ibid.: 63–64). Her logic can't be faulted, but her conclusion – that this makes them useless as definitions – does not necessarily follow. Most

ordinary people are relaxed about the fact that definitions – whether in dictionaries or face-to-face discourse – rely on reference to other words, and that the process sometimes entails a degree of circularity. But unless the circularity is flagrant and deliberately obfuscatory, it is hard to imagine the average user even being aware of the problem, still less being as fazed by it as Wierzbicka appears to be.

(3) Definitions should be 'substitutable': the idea here is that a definition should be written in such a way that it can be substituted for the definiendum in any context in which it appears. Thus, if *tenable* is defined as:

> capable of being defended against attack (*MW-3* 1961)

the sentence 'Their position was no longer tenable' can be reformulated with the definition substituted for the word *tenable*: 'Their position was no longer *capable of being defended against attack*'. Like so many of these principles and conventions, it is fine when it works well, and it works well here. Defining 'substitutably' can also be a useful training exercise, but its value for dictionary users is by no means clear (do non-lexicographers really interpret definitions in this mechanical way?), and any policy that required all definitions to be forced into this mould would be pointlessly restrictive.[20]

(4) Aim for maximum economy: once the dictionary progressed from specialized glossary to complete inventory of the vocabulary of a language, space was always at a premium. From the eighteenth century until the (very recent) arrival of electronic delivery media (which to some extent liberate us from the constraints of the printed page), a great deal of ingenuity has been applied to the goal of cramming as much information as possible into a finite space.[21] This imperative has had a major influence on dictionary metalanguage in general and defining practices in particular (think, for example, of the use of abbreviations like 'esp' and 'usu'). The three entries below typify a strategy found in many dictionaries: a full explanation of the key

[20] See also Hanks (1987: 119), who notes 'the awkwardnesses in the phrasing' of definitions that this often leads to.

[21] See for example the entry from the *Concise Oxford Dictionary* in Chapter 2 (Figure 2.3): this eight-line extract records at least fifteen separate facts about the word *bag*.

concept is provided only once, and related entries refer back (or forward) to this definition. In this case, a user who genuinely didn't know what the verb *bribe* meant would have to go first to *bribery* and then to *bribe*-noun in order to find out:

> **[1]bribe** *vt* to induce or influence by or as if by bribery
> **[2]bribe** *n* money or favor given or promised to a person in a position of trust to influence his judgment or conduct
> **bribery** *n* the act or practice of giving or taking a bribe
>
> (*MWC-8* 1980)

To purists like Philip Gove (Editor of the Merriam-Webster dictionaries during the 1950s and 1960s), this approach is not only economical but desirable, since it guards against introducing concepts into the definition of one word which may be absent from a related word. But such 'dangers' tend to be overrated by those who treat dictionary users as if they are Lexicographers. As all user research shows, it is never a good idea to require users to go to a second (or third) entry to find the information they are looking for in a first. Commenting on Gove's prescriptions, Landau bemoans the fact that the dictionaries he edited sometimes 'sacrificed intelligibility to a purity of style bordering on lunacy'.[22]

10.6.1.2 *Conventions of defining* The conventions described in this section are used – to some degree or other – in almost every type of monolingual English dictionary, but (as we shall see) none of them is indispensable.

(1) The genus-and-differentia model: in classical defining theory, a word or sense is described in terms of its superordinate, or 'genus expression' (which indicates the broad semantic category the word belongs to), and of its additional features, or 'differentiae' (whose function is to distinguish the current meaning from other category members). So for example:

> **surgeon** *n* a doctor who does operations in a hospital
>
> (*OALD-7* 2005)

'Doctor' is the genus (a surgeon is a type of doctor), and 'who does operations in a hospital' is the differentia (the feature that

[22] This splendid observation appears in the first edition of Landau's *Dictionaries* (1984: 127) but is curiously absent from the second edition.

distinguishes *surgeon* from cohyponyms like *anaesthetist*, *paediatrician*, and *general practitioner*). This approach (which goes back to Aristotle, via Linnaeus) rests on the notions that words belong to taxonomies, and that their meanings are reducible to a set of essential conditions. This is an effective defining strategy in many cases but, as we saw earlier (§10.5.1, §10.5.2), there are large swathes of the lexicon where the model simply doesn't fit the facts of the language. Many words and meanings cannot sensibly be explained in these terms, because the way natural languages are organized does not always correspond so conveniently to this taxonomic model.

(2) The 'lexicographic' use of parentheses: in traditional defining, parentheses have two specialized functions:

- to indicate a word's 'selectional restrictions' (a verb's usual range of subjects or objects, or an adjective's typical complements: §8.5.2.2)
- to encode in a single defining phrase two or more possible readings; different readings are activated when the information in parentheses is suppressed, or when the parentheses themselves are ignored.

The following entries show how the system works:

> **assassinate** (. . .) *tr.v.* **1.** To murder (a prominent person) by surprise attack, as for political reasons. (*AHD-4* 2000)

The information in parentheses tells us that the usual object of *assassinate* is 'a prominent person'.

> **send** *v*. . . . **8** (of a (person using a) radio apparatus) to transmit (*LDOCE-1* 1978)

Here, the parentheses specify the normal range of subjects when the verb is used in this meaning. What they tell us is that the sender may be either 'a radio apparatus' or 'a person using a radio apparatus'.

> **shatter** *v*. to (cause something to) break suddenly into very small pieces (*CALD-2* 2005)

Here, the parentheses indicate two possible readings, showing that *shatter* is an ergative verb in this meaning. If the words in parentheses are suppressed, *shatter* is intransitive and means 'to break suddenly into very small pieces'. But if the parentheses themselves are simply deleted, it is a transitive verb meaning 'to cause something to break suddenly into very small pieces'. In some types of dictionary, this convention still flourishes, but its use has been abandoned in most

learners' dictionaries. Some of the functions performed by parentheses are now handled in other ways (see §10.6.3), but their use as an indicator of transitivity has simply been dropped, as the following entry shows:

> **assassinate** . . . to murder an important or famous person, especially for political reasons (*OALD-7* 2005)

The argument that precision is thereby lost is easily countered: there is no evidence that users of learners' dictionaries ever understood (or even noticed) this idiosyncratic convention (one wonders how many users of other dictionaries really understand it either), and the risk of a learner misconstruing the definition to produce an 'incorrect' sentence is in most cases minimal.

(3) Formulaic defining components: over time (cf. Box 10.1), a range of defining formulae has evolved which – like parentheses – enable lexicographers to account for contextual variability within a single, concise defining statement. Here are some examples:

> **wolf** any of various large dog-like mammals . . . (*MW-3* 1961)
> **abstinence** the action or practice of abstaining (*OED-2* 1989)
> **absurdity** the state or quality of being absurd (*OED-2* 1989)
> **pedantic** of, relating to, or being a pedant (*MWC-11* 2005)

Each word in the formula has a precise function. Thus the expression 'of a pedant' maps onto uses like these:

Correct punctuation is neither an irrelevant luxury nor a pedantic affectation.
His slightly pedantic manner isn't perhaps quite what's wanted for the part.

And 'being a pedant' defines the word in contexts like these:

Punctilious and at times pedantic, he could appear abrupt and unfriendly.
To his priests, McQuaid was increasingly a remote and pedantic disciplinarian.

This approach to defining is undeniably economical, enabling lexicographers to account for a wide range of uses in very few words. It can be argued, too, that formulae like these can impart a degree of systematicity that is missing from the more ad hoc approaches found in early dictionaries. For the lexicographer, the usefulness – and meaningfulness – of these conventional formulae is not in doubt. The more important question, however, is whether they mean much to the ordinary user.

10.6.1.3 *Received wisdom: some conclusions* Taken together, these conventions provide lexicographers with a repertoire of strategies for describing meanings and usage with precision and economy. They are still used extensively in many types of dictionary, and their cultural influence is clear. Whenever newspaper columnists or advertisers produce mock definitions, they invariably make them look like entries from the most conventional dictionaries around – as if aping the style conferred credibility on their efforts. The conventions we have discussed can be seen as a kind of formal language, giving the dictionary an air of rigour and 'authority'. But is this a good thing? Formal languages are well adapted for use in mathematics, computer science, and logic, but explaining the messy business of human communication is another matter. As Bolinger observed, 'The orderliness and apparent system in a dictionary are more the result of our instinct to be orderly than of any towering need for system based on the subject matter' (1965: 572). Whatever the internal coherence of the 'system' applied in dictionaries, the only useful criterion by which it should be judged is its value for the dictionary user. What matters – and this is critical – is not the writer's intention but the reader's interpretation.

10.6.2 Ordering information: 'form' and 'function'

Most human artefacts (a vast category) and many objects in the natural world have both a 'form' (what they look like, what they are made of) and a 'function' (what they are used for). For many kinds of definition, this is a useful distinction to keep in mind. The definer has to decide which is more important: if the definition consists, in Bolinger's immortal phrase, of 'hints and associations that will relate the unknown to something known' (1965: 572), then you need to ask yourself whether form or function is more likely to be effective in this role. Consider, for example, a *windmill* (of the traditional type): in industrialized economies, windmills are rarely used for their original function of crushing grain, but they are familiar structures with a distinctive form. In this case, a definition that prioritizes form looks an appropriate strategy:

a building with sails or vanes that turn in the wind and generate power to grind corn into flour

(*ODE-2* 2003)

But what about a *watering can*? A dictionary designed for non-native learners of Spanish defines it (or rather, defines its Spanish equivalent, *regadera*) like this:

recipiente con un tubo acabado en una boca ancha con muchos agujeros pequeños que se usa regar, generalmente plantos

(*VOX Diccionario para la enseñanza de la lengua española* 1995)

A rough translation reads:

a container with a tube that ends in a wide mouth with many small holes which is used for watering things, usually plants

Ask yourself this: how much of the definition do I have to read before I can identify the definiendum? In this case, rather a lot: the first thirteen words describe what the thing looks like. We would have grasped the concept much faster if function had preceded form, as in this definition:

a container used for pouring water on plants, with a handle and a long spout

(*MED-2* 2007)

Two additional points are worth making. First, most dictionary look-ups are made to resolve a specific communicative problem. Once the user has found the required information, s/he can close the dictionary and return to the task in hand. In other words, users don't necessarily need to read the whole definition. Where form and function are shown in the optimal order (as in the second definition of *watering can*), we increase the chances of users being able to 'log off' before the end of the definition (which is hardly an option in the first definition). The second point is that form is a less reliable indicator, since the shape and construction of things tends to vary but their function is usually stable.

10.6.3 Alternatives to conventional definitions

Since the 1970s there has been a greater readiness, in most English dictionaries, to accommodate the kind of user who can't be assumed to be familiar with arcane lexicographic conventions. Dictionaries like *CED*, *AHD*, and *ODE* have made serious efforts to adapt traditional defining principles to a style of language that approximates more closely to 'normal' prose. But the story doesn't end here. This drive for greater accessibility has

combined with insights from corpora and from linguistic theory to give rise to a number of new defining strategies. The most significant of these is the 'full-sentence definition' (or 'FSD'), and in this section we assess the merits of this approach and of two other recent developments in the techniques of defining.

10.6.3.1 *Full-sentence definitions* As the name implies, full-sentence definitions (FSDs) present defining information in the form of a complete sentence in which the *definiendum* is embedded. The style was devised during the genesis of the first *COBUILD* dictionary (1987), and is now widely used in dictionaries for learners.[23] Here are some examples:

> **exorbitant** exorbitant prices, rents, charges etc are very much higher than they should be and you think they are unfair (*Longman Essential Activator* 1997)
> **expire** When something such as a contract, deadline, or visa **expires**, it comes to an end or is no longer valid (*COBUILD-5* 2006)
> **apprentice** if someone is apprenticed to another person, they are employed by that person to learn the type of work that they do (*MED-2* 2007)

In formulations of this type, the 'left-hand side' exemplifies usage, while the 'right-hand side' supplies the definition. Thus the first half of the entry for *exorbitant* tells us the kinds of thing this adjective typically describes ('prices, rents, charges etc'), then the second half tells us what it means. The case for FSDs rests on the notion that, for many words, 'the characteristic co-text is part of the meaning, and so is relevant to the definition of the item' (Sinclair 2004: 5). Thus the entry for *expire* not only shows that it is intransitive but also provides helpful information about its selectional restrictions (the kinds of thing that are typically said to 'expire'). And as for *apprentice*, the wording indicates that this verb has a strong tendency to occur in the passive and be followed by 'to'. The FSD approach allows us to embed these colligational and collocational preferences in the definition itself, giving learners a fuller picture of how the word is normally used. FSDs are also well adapted for conveying the kinds of extralinguistic information we discussed earlier (§10.5.6), as shown in the following entry for *old school tie* – a fiendishly difficult concept to define using conventional styles:

[23] FSDs are discussed in detail in Hanks (1987) and Rundell (2006).

When people talk about the **old school tie**, they are referring to the situation in which people who attended the same public school use their positions of influence to help each other

(*COBUILD-5* 2006)

For many kinds of headword, this can be an effective strategy, and the FSD is a valuable addition to the definer's repertoire. It works especially well for words whose selectional and syntactic preferences are easy to identify and fairly limited. This makes it a much better way of handling a word like *temerity* (which appears in the pattern 'have the temerity to do something' in 75 per cent of cases) than a traditional definition could ever be. Moreover, it fits nicely with the corpus-driven ethos of focusing on the way words typically behave in text, on 'stating what is normally the case rather than what is necessarily the case' (Hanks 1987: 118).

Problems can arise, however, if the contextual information in the left-hand side is not carefully selected. *COBUILD*'s definition for the noun *insight*, for example, starts with the phrase 'If you gain **insight** or an **insight** . . . '. By analogy with what happens in the definitions for *exorbitant*, *expire*, and *apprentice* (above), the user will deduce from this that:

- the noun can be countable or uncountable ('insight or an insight')
- it often appears as the object of a verb
- and when it does, the verb is usually *gain* (or something similar).

The first two deductions match the evidence pretty well, but the third is wildly wrong. Corpus data shows that, while *gain* + *insight* is a fairly common combination, the noun is much more likely to appear as the object of *give*, *offer*, or *provide*: events and observations 'provide insights' far more often than people 'gain' them. The definition framework has obliged the lexicographer to situate the word in a context, but the context that has been selected is overrestricted and thus gives an inaccurate picture of how *insight* is normally used.

A more serious criticism is that the genuine benefits of the FSD approach are squandered when the style is applied wholesale. The fact is that many words in the language can appear in a wide range of contexts, and in such cases a full-sentence format is unhelpfully restrictive. Where the contextual possibilities are less limited than they are for, say, *temerity* or *expire*, we can end up either with entries that mislead the user (like the one for *insight*), or – worse still – with pointlessly wordy definitions like these:

If you say that someone or something is **fortunate**, you mean that they are lucky
Something that is **inspirational** provides you with inspiration

(both *COBUILD-5* 2006)

Since *fortunate* is a relatively unrestricted adjective (it can apply to people, events, or situations, and it has no marked preference for attributive or predicative position), the definition ends up saying very little but taking a long time to do so. For learners of English, this has two big drawbacks. First, longer definitions increase the processing load for unskilled users, making their lives harder than they need to be.[24] Secondly, if the FSD approach is used – as in the *COBUILD* dictionaries – for *every* entry in the dictionary, the cumulative effect of all these longer definitions is to reduce the space available for other material, leaving the dictionary with significantly fewer headwords (cf. Rundell 2006: 327 for details).

➔ Look carefully at the corpus data and decide whether your LU is sufficiently restricted in its contextual behaviour to benefit from a full-sentence approach.

The FSD style works well for many adjectives and for intransitive verbs whose selectional restrictions are limited and predictable, and it can also be helpful in the case of items with marked colligational preferences (such as transitive verbs occurring mainly in the passive, or nouns that are almost always pluralized). But applying it to words whose contextual range is broad (like *advice*, *house*, *easy*, or *kill*) simply leads to bloated definitions whose length isn't justified by any obvious benefits for users.

10.6.3.2 *'When' definitions* In a still more recent development, some learners' dictionaries have experimented with a style of definition that begins with 'when' but (unlike the FSD) consists of a single clause and has no main verb. For example:

discussion when people talk about something and tell each other their ideas or opinions (*CALD-2* 2005)
peace when there is no war (*Longman Essential Activator* 1997)
balancing act when you are trying to please two or more people or groups who all want different things (*LDOCE-4* 2003)

[24] And the longer the definition, the greater the risk of problems with anaphora resolution, as in cases like this: 'If something **necessitates** an event, action, or situation, it makes it necessary'. What do these 'its' refer to?

This approach seems to be used mainly for defining nouns that refer to states or situations, and as we show later (§10.6.4.2), these can pose difficult problems for definers. The 'when-definition' is one way of avoiding the lexicographic formulae that generally introduce definitions for words of this type (cf. §10.6.1.2), and the style somewhat resembles the folk-defining techniques used, for example, by teachers and parents.[25] A typical exchange might go like this: 'What does *discussion* mean?' 'Well, it's when people talk about something and tell each other their ideas or opinions.' But this is a high-risk strategy: the folk-defining technique is well adapted to face-to-face encounters, but in the setting of a dictionary (or indeed any written discourse), a statement starting with 'when' looks like a subordinate clause. This sets up an expectation of a main clause (many FSDs follow exactly this path) – which in this case is not fulfilled. The risks of misinterpretation are especially high when the same word-form can be either a noun or a verb, and the noun is defined in this way. For example:

delay *n* when someone or something has to wait (*LDOCE-4* 2003)

Dziemianko and Lew (2006) report two empirical studies with Polish students: the results are not conclusive, but in one of the studies users had real problems in identifying the wordclass of the item being defined. On the whole, this style is best avoided, at least until we have a clearer idea of how users cope with it.

10.6.3.3 *Short definitions* We saw earlier how some learners' dictionaries provide navigational aids for longer entries, in the form of short, suggestive defining phrases (not complete definitions). These are either listed together in a menu at the top of the entry (§7.2.1.3) or shown separately as 'signposts' before each full definition (§7.2.5.2). One dictionary for native speakers – the *Encarta World English Dictionary* (Bloomsbury 1999) – takes this a step further, supplying 'quick definitions' for every sense of every headword. At the entry for *fork*, for example, the full definitions for the first three senses are preceded by shorter versions saying (respectively) 'UTENSIL FOR EATING', 'GARDEN OR AGRICULTURAL TOOL', and 'DIVIDING POINT IN ROAD OR RIVER'. As well as having a navigational function, these quick definitions provide 'a thumbnail sketch' for 'the user who does not want, or need, the full picture' (*Introduction*). This is an interesting

[25] On folk-defining styles, see Stock 1988: 84–85.

development, especially in the light of our earlier discussion (§10.4.2.3) of the divergent needs of encoding and decoding users. As electronic dictionaries start to exploit the opportunities of the medium more imaginatively, they could well offer users a range of defining options, from truncated definitions like these to versions which provide enough information to support successful encoding. Meanwhile, dictionaries have already started to appear on mobile devices (a trend that looks set to accelerate), and short definitions make optimum use of small screens.

10.6.4 Difficult cases

Some of the defining styles we discussed in the previous section – notably full-sentence definitions – are at least partly motivated by inadequacies in the traditional models. For some categories of word, conventional definitions don't provide an efficient or helpful format, so we need to find other solutions. In this section, we discuss three types of 'difficult case' and suggest ways of dealing with them.

10.6.4.1 *Adjectives* We noticed earlier that adjectives conform less well than nouns and verbs to a taxonomic model in which there are superordinates and hyponyms (§5.2.1). A genus-and-differentia approach will work in some cases, and there is no reason not to apply it when it does:

> **irate** very angry because someone has done something to offend you or upset you (*Longman Language Activator* 1993)

Here the genus ('angry') is differentiated both by an intensifier ('very') and by a causal explanation ('because someone has . . . '). But for most adjectives, this is not really an option. It was the lack of a single 'default' style for defining adjectives that gave rise to the kind of formulae (like 'marked by' and 'of, being, or pertaining to') that make life easy for lexicographers but do little to enlighten the user. Landau (2001: 172) provides a helpful list of more transparent opening gambits, such as 'consisting of', 'capable of', made of', and 'full of', and a good Style Guide should give definers a range of options to choose from. For example, English words beginning with a negative prefix can often be defined using the formula 'difficult or impossible to . . . ':

> **incomprehensible** difficult or impossible to understand (*MED-2* 2007)

And the more informal style of many learners' dictionaries allows for definitions like this:

> **inconsolable** so sad that it is impossible for anyone to comfort you (*LDOCE-4* 2003)

But this is where the full-sentence definition really comes into its own. For many adjectives, the best clue to the meaning lies in the class of things the adjective typically modifies – its selectional restrictions, in other words. Traditional lexicography handles these by means of a bracketed statement at the head of the definition:

> **etiolated 1** (of a plant) pale and drawn out due to a lack of light (*ODE-2* 2003)

But a full-sentence style provides a more elegant solution:

> **innocent 4** An **innocent** question, remark, or comment is not intended to offend or upset people, even if it does (*COBUILD-5* 2006)

FSDs often enable you to produce a definition which avoids the awkwardness of wording that mars so many adjective definitions in more traditional dictionaries. And as well as dealing with selectional restrictions, the FSD is an effective way of defining adjectives that describe permanent character traits as opposed to transient feelings or behaviour. For example:

> **bad-tempered** Someone who is **bad-tempered** is not very cheerful and gets angry easily (*COBUILD-5* 2006)

➔ By selecting the most appropriate strategy from the ones described here, you should be able to avoid unhelpful lexicographic formulae.

10.6.4.2 *Abstract nouns* One of the most intractable problems in monolingual lexicography is how to define abstract nouns without resorting to expressions like 'the act of X-ing' or 'the quality of being X':

> **insistence 1:** the act or an instance of insisting **2:** the quality or state of being insistent (*MWC-11* 2003)

It is unlikely that the average dictionary user derives much illumination from definitions like this, yet these perennial standbys can be found even in dictionaries that explicitly reject lexicographic conventions:

insignificance **Insignificance** is the quality of being insignificant (*COBUILD-5* 2006)

destruction **Destruction** is the act of destroying something or the state of being destroyed (*COBUILD-5* 2006)

The resilience of these formulae isn't surprising: when defining words like these (typically nominalizations of verbs and adjectives), it can often be difficult to find a form of words that genuinely explicates the meaning. It is worth making the effort to abstract meaning from the evidence of usage, and explain it more discursively – as in this entry for *honesty* (defined in many dictionaries as 'the quality of being honest'):

an honest way of talking or behaving, so you tell the truth and do not try to cheat people or hide information from them

(*Longman Essential Activator* 1997)

In other cases, it may be possible to find a more specific genus word: thus *advocacy* can be defined as 'public **support** for something', *isolationism* as 'a **policy** of . . . ', and *accreditation* as 'official **approval** of . . . '. Other useful genus expressions include 'the **ability** to . . . ', 'a **feeling** of . . . ' and 'a **situation** in which . . . '.

➔ You won't always be able to avoid the 'act of/state of/quality of . . . ' formulae, but it is worth trying every other option before you reach this point. Applying these formulae as a sort of automatic, default approach suggests a lack of concern for the needs of users.

10.6.4.3 *Grammatical words and other non-lexical headwords* Even the most diehard traditionalists accept that conventional defining techniques can't be applied to every type of vocabulary item. For example, it must be obvious that explaining the 'existential' use of *there* (in a sentence like *There were soldiers on every street*) can't be achieved with a standard substitutable definition. For 'grammatical words' (discussed above: §6.2.1.1), and related items like abbreviations, interjections, forms of address, and affixes, a different set of strategies comes into play. These need not detain us long. Depending on the type of dictionary and user, some form of 'usage' definition (similar to the ones used for explaining pragmatic messages: §10.5.6.1) is generally employed. Words like this do not really 'mean' anything, so our job is to explain their function in the sentence and the contribution they

make to its overall meaning. A few examples will illustrate the kinds of approach you can use:

> **that** *relative pron.* **I.1.a.** Introducing a clause defining or restricting the antecedent, and thus completing its sense (*OED-2* 1989)
> **hm 1** used for representing the sound you make when you are pausing to think before saying something else **2** used for . . . (*MED-2* 2007)
> **mate 2** Some men use **mate** as a way of addressing other men when they are talking to them (*COBUILD-5* 2006)
> **-ly** suffix forming adverbs from adjectives, chiefly denoting manner or degree (*ODE-2* 2003)
> **BTW** by the way: used in emails and text messages for adding additional information (*MED-2* 2007).

10.6.5 Words used for defining

We made the case earlier (§10.4.3) for seeing intelligibility as a *sine qua non* for a successful definition. In previous sections we have shown how various defining techniques and defining structures can contribute to intelligibility (or in some cases, compromise it). But what about the actual words used in definitions? The notion that definitions should, as Johnson put it, use 'terms less abstruse than that which is to be explained' is generally accepted in principle if not always applied in practice. Certainly, the more user-friendly defining practices introduced by *AHD* and *CED*, and further refined in more recent dictionaries for native speakers, have tended to see accessibility as at least as important as precision. And this has led to a reduction in the use of words in what could be seen as 'dictionary-specific' meanings. Traditionally, dictionary metalanguage has included words which – even if not inherently difficult – are used in rather mannered, old-fashioned, or idiomatic senses; the use of 'strike' (meaning 'to hit') in definitions is a case in point. We mentioned the defining phrase 'the quality of being X' in the previous section: this is proving hard to eliminate from dictionaries, but the argument against it is that this is not a salient meaning of *quality* (which is mainly used for talking about how good something is), but one that is used almost exclusively in dictionaries.[26] As dictionary definitions strive to use natural, unremarkable prose, uses like this are gradually becoming rarer.

[26] *Quality* occurs 18,621 times in the BNC. Though the string 'quality of' is frequent enough (with over 4,000 occurrences), it usually appears in expressions like 'improving the quality of medical care', which invokes the word's 'basic' sense. The precise string

10.6.5.1 *Controlled 'defining vocabularies'* Dictionaries designed for learners need to go a good deal further to ensure intelligibility. Most use a 'defining vocabulary' (DV) in their definitions. A defining vocabulary is a finite list of high-frequency words (typically the most frequent 2,000–3,000 words in the language) which the learner is expected to 'know' sufficiently well to be able to understand any definition in the dictionary. The genesis of these lists goes back at least as far as the 1920s, to work done by Harold Palmer, Michael West, and A. S. Hornby (Rundell 1998: 316–320; Cowie 1999b, chapter 1), and the first dictionary to use a defining vocabulary (in this case of 1,490 words) was West's *New Method English Dictionary* (1935).

Using a DV is not without its problems. There are cases, for example, where a word cries out for the use of another word in its definition. Consider the set *volcano*, *erupt*, *lava*: each item would benefit if its definition could refer to one or both of the others, but the constraints of the DV rule this out. Some systems have been criticized – on the whole fairly – for 'abusing' the controlled list, either by including idiomatic uses (such as phrasal verbs or expressions like 'let slip' and 'take place') or by generating 'new' words from the basic set. Bogaards, for example (1996: 289) notes that the then current edition of *LDOCE* used the word 'independence' in definitions, though its DV list showed only the verb 'depend' and the affixes 'in-' and '-ence' (see also Jansen, Mergeai, and Vanandroye 1987). Even the latest *LDOCE* uses 'birthday' in definitions but lists only 'birth' and 'day' in its DV. But these are not insurmountable problems. Many of the DV-related faults that critics have identified arise from an inexpert application of the policy rather than from inherent problems with the system. Despite the occasional glitch, a good DV that lists all the word-forms which are actually used in definitions (and all the multiword expressions like 'by means of' or 'in charge of') still provides the soundest basis for ensuring intelligible definitions. A possible variation on the theme was proposed many years ago by Janet Whitcut (1988: 53), who suggested a DV with a 'hierarchy of strata': the idea is that the simplest words are themselves defined by other simple words (entailing some degree of circularity), but then a banding system is applied, so that words in band 2 are available for defining those in band 3, and so on. This is theoretically appealing, and we now have the data

'quality of being X' – the meaning used in dictionary definitions – appears only twelve times.

to do the banding properly. Whether the market would accept it is another issue.

10.6.6 Definition form: conclusions

As we have seen in this section, there is a rich inventory of defining styles available to the lexicographer. A good maxim to keep in mind is 'horses for courses': a traditional 'genus-and-differentia' approach works well in many cases (but you need to take care in selecting the best genus expression), while a full-sentence style is highly effective in others. But an uncritical application of either type to the whole lexicon is sure to lead to some bad definitions. A good Style Guide will list and exemplify the allowable options for each wordclass, and provide guidelines indicating which is recommended for which types of meaning. It is then up to you to choose the most appropriate definition framework. If your project uses template entries (§10.1.3), then the style and wording of the definition will be to some extent predetermined. It is true that a dictionary definition is a somewhat specialized form of discourse, but this doesn't excuse definitions which are difficult to read or which depart so far from 'normal' prose that they barely sound like English. Finally, when selecting lexical items to encode the facts and ideas you want to convey, be careful to strike the right balance (which will vary according to the type of user the book is aimed at) between accuracy and accessibility. As Bolinger observes: 'Many things can misrepresent a meaning, including an excess of erudition' (1985: 73).

10.7 What makes a good definition?

Definitions succeed when they get two things right: content and form. The precise configuration will be determined by the needs and skills of the users of the particular dictionary you are working on, but if a definition doesn't provide the information its users require, in a form they can readily digest, it has failed. So for example, a definition consisting only of synonyms may be easy to follow, but in most cases it won't give an adequate account of content (§10.5.5). On the other hand, a definition that provides the necessary content in technically precise language is of no value if it is unintelligible to the users it is aimed at (§10.4.3).

So much for general principles. Here are a few more specific words of advice:

➔ Explain, don't 'define': you have to tell users what people really mean when they use a word.

A common mistake here is to focus on etymology instead of meaning. Consider, for example:

> **abase** to lower oneself/sb in dignity (*OALD-5* 1995)
> **extramural** outside (the walls of) a town or organization (*LDOCE-1* 1978)

In both cases, the definer has fastened on the word's origins (linking 'base' to 'low', and 'mural' to 'walls'), so the definition ends up telling us more about how the word evolved than about how it is currently used to form meanings.

➔ Remember that for many users, the concept being defined may already be familiar.

In such cases, the definition's primary function is to 'enable the dictionary user to identify the concept in question, that is, to retrieve it from his/her own conceptual memory' (Geeraerts 1990: 196).

This applies especially to common concepts being defined in a dictionary for adult learners (who already know, for example, what a *bicycle* or a *parachute* is, but simply don't know the English words for them). A useful principle here is to present the information in stages, stating the most basic points as early in the definition as possible, then elaborating or exemplifying as necessary. (Our discussion on form and function is relevant here: §10.6.2). The user who knows the concept may then be able to identify it quickly, and has the option of 'logging off' before the end of the definition.

➔ A definition should contain no more words than is necessary, consistent with the demands of intelligibility and information-transfer.

Once you have written your definition, it's always a good idea to go back and see whether any words can be removed without compromising naturalness or making the definition less informative. In many cases, 'less is more'. This is especially important as we begin to exploit more fully the opportunities of the electronic medium: the idea of multi-level, multi-purpose definitions is appealing (cf. §10.6.3.3), but the absence of space constraints is not an excuse for wordiness.

➔ Remember that there is an inverse correlation between the time it takes you to write a definition, and the time it takes the user to process it: the more effort we put into this task as lexicographers, the easier we make life for our users.

It is well worth having a look at George Orwell's essay 'Politics and the English Language' (1946), which is easily found on the web. Though Orwell was talking mainly about 'the abuse of language' in political discourse, his central point is that clear thinking requires careful use of language, and almost all his prescriptions have relevance to the art of writing definitions. He notes that some writers are 'haunted by the notion that Latin or Greek words are grander than Saxon ones', and he deplores the tendency to replace perfectly adequate simple words with 'pretentious diction' in the mistaken belief that this confers 'dignity' on the text. His strictures against 'meaningless words' could be applied to this definition for *virginity* from the dictionary.com website: 'the state or condition of being a virgin'. Is there any useful distinction here between 'state' and 'condition'? Orwell's comments on the lazy use of over-familiar phrases could apply to some of the lexicographic formulae we discussed above – which can be a substitute for thinking carefully about what words really mean. He concludes with some practical suggestions which apply as much to definition-writing as to any other form of prose, including: 'never use a long word where a short one will do; if it is possible to cut a word out, always cut it out; never use a passive where you can use the active'. Johnson believed that the best definitions combined 'brevity, fulness, and perspicuity' (*Plan* 1747), and this encapsulates perfectly the qualities we should aim for.

10.8 Examples

Example sentences are a vital component of the kind of database we described in Chapters 8 and 9. Their function in the database is to support and illustrate every linguistic fact recorded there, and to provide editors at the 'synthesis' stage with the raw materials for constructing a dictionary entry (cf. §9.2.4). Space isn't an issue at this point, and database examples will typically be complete sentences taken from the corpus. In the finished dictionary, however, the examples have somewhat different functions, and these vary according to the type and level of dictionary. The use of examples in a bilingual dictionary is discussed in Chapter 12 (§12.3.3). Here we look at

their functions in monolingual dictionaries. Our discussion will also address the issue of where examples should be sourced from, and we conclude with some guidelines for producing good dictionary examples.

10.8.1 The function of examples

Though Johnson was not the first English lexicographer to add illustrative quotations to his entries, his *Dictionary* (1755) was the first to be based on a systematic analysis of language data. Almost every word or meaning he describes is supported by a quotation from one of the numerous sources in his bank of citations (cf. §3.2 above). Johnson's dictionary thus embodies the principle that languages should be described on the basis of objective evidence of their use – and this, in a sense, is the primary function of examples of usage: as a source of data from which lexicographers construct their entries. Attaching examples to definitions is a separate process, and this is what we discuss here.

10.8.1.1 *Attestation* One of the fundamental goals of a historical dictionary is to fix the origins and trace the development of a word, meaning, or phrase. In many cases, a quotation is used, as Johnson says, for 'no other purpose, than that of proving the bare existence of words' (*Preface*). This is what we mean by 'attestation', and it is a major function of examples in dictionaries like the *OED*, because a sentence taken from an authentic text 'demonstrates objectively that a word . . . may be found in the language' (Simpson 2003: 268). Johnson initially tried to make his quotations serve a didactic purpose too (whether by being morally uplifting or stylistically elegant).[27] But he was eventually obliged to admit that 'words must be sought where they are used', and this is the policy followed by contemporary historical dictionaries. It goes without saying that an attributed quotation should never be altered or adapted, though a sentence from the original may be shortened in the dictionary entry, for example by the deletion of a non-central clause.

[27] 'I was desirous that every quotation should be useful to some other end than the illustration of a word.' And again: 'I have studiously endeavoured to collect examples and authorities from the writers before the restoration, whose works I regard as *the wells of English undefiled*.' (*Preface* 1755)

10.8.1.2 *Elucidating meaning* Examples illustrate usage, and are often a helpful complement to the definition. A well-chosen example can also clarify sense distinctions in a polysemous word; indeed, you sometimes find that an entry is almost incomprehensible without its examples.[28] Ideally, definition and example will each be self-sufficient, and a definition which can't be understood without its supporting example is less than optimal. But a dictionary definition is by its nature a rather abstract construct, and there are many cases where the full sense of a difficult concept only becomes clear when you read the example:

> **tantamount** If you say that one thing is **tantamount** to a second, more serious thing, you are emphasizing how bad, unacceptable, or unfortunate the first thing is by comparing it to the second: *What Bracey is saying is tantamount to heresy . . . He said the decision was tantamount to protecting terrorist organisations around the world.* (*COBUILD-3* 2001)

10.8.1.3 *Illustrating contextual features: syntax, collocation, register, etc.* A well-populated database will record the full range of contexts (whether lexical or syntactic) in which a word or meaning typically occurs. And in a dictionary entry compiled on the basis of this information, examples have an important role in illustrating the word's contextual range. This is especially important in dictionaries aimed at learners. Even an apparently straightforward word like *television* can, as Fox points out, be difficult for a learner to use appropriately unless s/he knows, for instance, that televisions are 'turned on' and 'turned off', and that in English 'we "watch" the television rather than "see" or "look at" it' (1987: 137). In most learners' dictionaries (and in some dictionaries for native speakers), it is usual to back up any statement about a word's syntactic behaviour with an example that instantiates the pattern:

> **decide** *verb* [with obj.] come or bring to a resolution in the mind as a result of consideration: [with clause] *she decided that she liked him* [with infinitive] *I've decided to stay on a bit | this business about the letter decided me.* (*ODE-2* 2003)

If an example illustrates – as it should – a typical instance of a word in use, then it will often show the word in one of its frequent collocational pairings. With their special emphasis on phraseology, most learners' dictionaries use examples to provide a full account of a headword's collocational behaviour:

[28] See for example the entry for *keen* in Figure 10.13, §10.5.5.

> **advice** noun [U] an opinion that someone gives you about the best thing to do in a particular situation: *You can always contact your tutor for advice and support… Let me give you some advice… I took his advice and left… We are here to give people advice about health issues… Tenants involved in a dispute with their landlord should seek legal advice.… She's acting on her lawyer's advice… She applied to York University on the advice of her tutor.* (*MED-2* 2007)

The examples in this entry incorporate a huge amount of information on the way *advice* typically combines with other words, including:

- you go to people *for* advice, and you get advice *about* something
- you can *give*, *take*, or *seek* advice
- you do things *on someone's advice* (or you *act on* it)
- *advice* can be modified by adjectives like *legal* (the dictionary lists other adjectives of this type in a separate box)
- *advice* often occurs with *support* in an 'and/or' pairing.

Finally, where an item is marked for style, register, or regional distribution, a good example will show it in its natural setting. A verb like *endeavour* poses few problems from the point of view of meaning: the most important thing the user needs to understand is the 'tone' of the word, which this example helps to convey:

I remained for some time endeavouring to engage Mr Campbell in conversation.
(*Longman Language Activator 1993*)

The headword itself belongs to a rather formal register, and this formality is nicely reflected in the lexis that makes up the rest of the sentence.

10.8.2 Where examples come from

The illustrative quotations in dictionaries like the *OED* typically come from large citation banks, collected over many years and now often complemented by data from diachronic corpora. Examples are 'attributed': historical dictionaries generally provide information about the source and date of the quotation. But in most other kinds of dictionary, attribution is rare, and examples may come in a variety of forms (from short fragments to full sentences) and from a range of sources (authentic texts, the lexicographer's imagination, or some combination of the two).

Pedagogical dictionaries present special challenges. For A. S. Hornby and his immediate successors, it was axiomatic that examples were there

to show learners the way words were typically used in text. For this purpose, invented (rather than authentic) examples were preferred because they allowed the lexicographer to illustrate several points in a single, carefully-contructed phrase or sentence (see e.g. Rundell 1998: 316–317; Cowie 1999b: 134–137). Early learners' dictionaries often used short fragments which made no claim to replicate actual performance, for example:

*a **serious** illness*
*to **introduce** a new law*
***modern** technology/architecture/art*

This consensus was blown apart by the arrival in 1987 of the first fully corpus-based English dictionary (*COBUILD-1*), and a fierce debate ensued. The issue was whether examples should be made up by lexicographers or taken directly from authentic texts. Antagonists supported their position by citing the worst instances of each type: a favourite target for the 'authentic' tendency was this example for *salvage*, from the first (1978) edition of *LDOCE*:

'We'll try to salvage your leg', said the doctor to the trapped man.

Few would defend this clumsy and unnatural sentence (which so obviously violates Grice's maxim of quantity, cf. Cruse 2004: 368), but in fact, pre-corpus editions of the main learners' dictionaries contained many perfectly good examples which looked authentic even if they were not. Members of each camp reported empirical research that appeared to justify their respective positions.[29]

The case for wholly authentic examples rests on the proposition that it would be 'ridiculous to have studied real language in order to find out the facts about the language and then to have abandoned this and concocted fake examples for the dictionary' (Fox 1987: 138). All of which would be fair comment if we were talking about a historical dictionary in which facts about language are supported by attributed quotations. But learners' dictionaries have very different goals, and here intelligibility and helpfulness

[29] For example: Laufer reports a study measuring the relative pedagogical effectiveness of authentic and made-up examples, and concludes that 'lexicographer's examples are more helpful in comprehension of new words than the authentic ones' (1992: 75). But Potter counters with evidence from a *COBUILD* user survey which 'found overwhelming approval among teachers and learners of English for real examples taken directly from a corpus' (1998: 358).

are at least as important as showing words in their natural settings. Consider these two examples, both from the first (1987) edition of *COBUILD*:

> **gravitate** . . . *He gravitated, naturally, to Newmarket.*
>
> **grudge** *(verb)* . . . *Not that she grudged it.*

Both examples are natural, typical, and authentic. But both are quite useless, because they appear in a dictionary for learners but could only be understood by a fluent speaker of English. Why would it be natural for someone to 'gravitate' to Newmarket? You need to know quite a lot about British culture to see what this is about: Newmarket is the centre of the British horse-racing industry, so we infer that 'he' is an aficionado of the sport. As for the second example, most native speakers would be able to reconstruct the kind of context in which it might appear (e.g. *She had spent over $10,000 on the wedding. Not that she grudged it – it was an unforgettable day*.) But the learner is not so well placed. The appeal to authenticity, as a sole guarantor of quality, may be missing the point. Language-learning (as both teachers and students accept) involves all sorts of 'unnatural' uses of language; think, for example, of the use of 'drills', which have a clear pedagogical purpose but don't pretend to be authentic language events. A dictionary example is an inherently unnatural object because it has been removed from the context which would (in real life) surround it – and clarify it. While this does not justify the 'over-contextualized' awkwardness of the notorious 'salvage' example, it is equally true that severely 'decontextualized' examples like the two shown above can only leave users mystified and discouraged. Being too informative gives a false view of how language works, but not being informative enough is just as unhelpful.

In fact this debate was ill-founded for two important reasons. In the first place, it presupposes a simple binary choice between two extreme positions: either you invent examples out of thin air, or you take them direct from a corpus without altering a single syllable. In reality, lexicographers rarely do either. We analyse a great deal of corpus data, identify recurrent patterns, and aim to reflect these in example sentences. Though the ideal example is one taken straight from the corpus with no editorial intervention, it is surprisingly rare – even in today's mega-corpora – to find corpus sentences that fulfil all the criteria for being 'good' examples (we discuss these criteria in the next section). What usually happens is that we find in the corpus a central 'core' consisting of a string of perhaps four to six words which show the headword in a highly typical context. We then make adjustments

to the rest of the sentence as appropriate, which may entail (*inter alia*) lopping off a long coordinate clause, changing a distracting proper name to a pronoun, or simplifying an obscure vocabulary item in a non-central part of the sentence. Thus the notion of a simple choice between 'made-up' and 'authentic' gives a misleading picture of how lexicographers really work.[30] The other unfortunate outcome of this debate is that it gives a spurious impression about what the corpus is for. In all types of dictionary, the primary function of the corpus is as a source of *evidence* rather than as a source of examples. As we have argued throughout this book, everything we say in the dictionary about what words mean and how they behave must be informed by, and faithful to, what the data tells us. But, in a dictionary designed for learners, there is no incompatibility in supporting a corpus-driven description with examples that reflect the recurrent patternings in the corpus within an accessible and intelligible format.

10.8.3 What makes a good example?

The nature of examples will vary according to the type of dictionary and the needs and expectations of its users. But the guidelines we give here apply to most situations; even in historical dictionaries that use attributed quotations, the basic criteria remain valid. These are that examples should be:

- natural and typical
- informative
- intelligible.

In this section we will flesh out these ideas with some practical advice. Many of the problems we identify here are exemplified in the first edition of *COBUILD*. Some commentators had a field day when the dictionary first appeared,[31] and this inevitably (and unfairly) detracts from the project's hugely positive impact on the world of dictionary-making. But *COBUILD*

[30] In fact, only a decade after the first edition of *COBUILD*, the dogmatic commitment to authenticity was already softening. As Potter observes (1998: 357), 'the distinction between real and invented examples . . . has become somewhat blurred', and many examples in the second edition of *COBUILD* edited corpus sentences to reduce their length and 'remove distracting, obscure, or possibly offensive elements'.

[31] Notably Hausmann and Gorbahn, who list dozens of examples which they criticize (convincingly, it must be said) as inadequate (1989: 45–47).

was making up the rules for a new lexicographic paradigm as it went along, and to its credit, many of the shortcomings in the original version were successfully addressed in later editions.

10.8.3.1 *Naturalness and typicality* 'Typicality' is easy enough to recognize: for all but the rarest items, a large corpus will show the contexts, syntactic patterns, collocations, and multiword expressions in which a word is most frequently found, and these represent its typical forms of behaviour. Naturalness is a more intuitive and less objective measure, but – in addition to the features just mentioned – aspects of colligation (such as preferred tense, number, mood, or position in the sentence) contribute to a sense that a text or utterance is 'natural'. 'Recurrence' is important here. The mere fact that something has been found in a corpus is not in itself a good enough reason to include it in a dictionary. We noted earlier (§3.1.2) that individual members of a speech community will sometimes use language in idiosyncratic ways. This example for the verb *sweat* is a case in point:

He was sweating like a bullock (*COBUILD-1* 1987)

Any native speaker would (without consulting a corpus) recognize this as an aberrant, creative usage, and we are not doing learners any favours by recording idiolectal quirks in a dictionary. Nowadays, we can confirm our intuitions objectively, and a quick search on Google shows around 20 examples of this pattern, as against over 66,000 of the 'canonical' expression *sweating like a pig*. But some early corpus enthusiasts privileged data over intuition as a matter of principle – so that even a single occurrence like this was allowed to overrule the collective knowledge of a team of native-speaker lexicographers. Fortunately, dilemmas like this rarely arise now. With today's mega-corpora, we will almost always find abundant evidence for any word or combination we want to describe, and where the data clearly shows a particular usage to be recurrent, there is no case for overturning it on grounds of intuition.

Naturalness is also a function of the amount of context a sentence provides. A besetting problem with many pre-corpus examples was their tendency to over-contextualize – as we saw with the 'salvage' example above (§10.8.2), which provides far more context than is natural. This is a challenging aspect of example-selection, because 'real' examples (being abstracted from their larger context) often run the opposite risk, of providing too little context to be helpful, or being overloaded with mystifying references to

people or things outside the sentence. Hausmann and Gorbahn (1989: 46) note several instances of this problem in *COBUILD-1*, such as this example for *every*:

One woman in every two hundred is a sufferer. (of what?)

We see a similar problem when the expression 'hot at' (=good at) is exemplified like this:

. . . which suggested that we weren't so hot at these things as we used to be
(what suggested it, and what are 'these things'?)

➔ A good example has to get the right balance between too much context and too little.

Finally, a natural example is one that maintains a consistent register. Thus an example for a colloquial usage found mainly in spoken mode shouldn't include more formal words. This example for *the latter* exhibits the opposite problem:

We have to decorate the kitchen and the hall – I'd rather do the latter first.
(*Cambridge International Dictionary of English* 1995)

Here, a rather formal expression is exemplified in a conversational, domestic context, and the result is an example which is easy to follow but completely unnatural.

10.8.3.2 *Informativeness* An informative example is one that complements the definition and helps the user understand it better. Here again, you need to strike the right balance between being so lacking in content as to convey almost no useful information, and giving the learner an extended reading passage. These examples (all from *COBUILD-1*) illustrate both problems:

'bring up the rear': *Jack brought up the rear.*
'crawl' (in the sense of 'grovel'): *Let's see who comes crawling to whom.*
'region': *To have access to the truth and so to pass beyond the region of mere opinion is to take great risks.*

It's important, too, that the information in the example doesn't appear to conflict with what the definition says. If the definition describes a 'common cold' as a minor illness which most people get quite regularly, it is not helpful to add an example saying:

A common cold could kill her. (*COBUILD-1* 1987)

In the right context, this would be a perfectly natural thing to say, but for a learner who has struggled to process the definition, and who believes s/he has grasped the concept, it can only be discouraging to find an example that seems to contradict all this.

A final point here is that the example must have some clear function. For many words in the lexicon, an example can add little of value, and the space it takes up could be used more productively. The entry for *Norwegian* in *COBUILD-3* (2001) has no fewer than four example sentences, but the same largesse is not bestowed on most other nationality words in the dictionary; nor does it need to be. (And in *COBUILD-5* 2006, all four *Norwegian* examples have disappeared.). As Johnson remarked, 'there is more danger of censure from the multiplicity than paucity of examples' (*Preface* 1755).

10.8.3.3 *Intelligibility* We made the point earlier (§10.4.3) that a definition may be accurate and may convey adequate content, but still fails if it is not readily intelligible to the type of user it is aimed at. The same principle applies to examples. We have seen cases above of examples which are natural, typical, and authentic, and which would even be informative if the user could understand them – but if the example is incomprehensible it is of no value. This means we need to avoid gratuitously difficult lexis and structures wherever possible. A user trying to process the idea of someone thinking themselves *above* other people will only be mystified by an example like this:

I had always considered Anthony priggishly above the rest of us. (*COBUILD-1* 1987)

'Priggish' is a rare word which encodes a difficult concept, and it is a pointless distraction here. It is true, of course, that some words we want to exemplify are themselves 'difficult' and typically occur with other difficult words. Here the demand for naturalness requires that we don't distort the facts of the language by surrounding our word with atypically low-level lexis. All of which shows what a complex challenge example-writing represents. But if you use the corpus carefully and get the right balance between the three criteria discussed here (naturalness, informativeness, intelligibility), the examples you produce should bring real benefits for your users.

10.9 Completing the entry

Finally, you've reached the end of your dictionary entry. The various senses of the headword have been teased out (at a level of granularity appropriate to the user-group you are writing for), and have been ordered in the entry in a way that best meets your users' needs. For each LU, you have provided a definition which conveys the information the user will need in order to grasp the concept, and does so with the minimum number of words and in language your user can readily understand. As far as possible, your entry will also cater for those users who want (or need) to use the word productively. This means describing – transparently, and without resorting to codes that have to be learned – the syntactic and lexical environments in which the word typically occurs. If the word is 'marked' for register, regional distribution, or any other sociolinguistic feature, this needs to be recorded too. And all of these aspects should be illustrated in well-chosen examples that faithfully reflect the evidence of the corpus but do so without compromising intelligibility. Once the whole thing has been checked for length (in a well-run project, you will know how much space you have to play with), voilà – your entry is complete.

Exercises

Exercise 1: Definition content

Collect corpus data for the following near-synonyms or their equivalents in your language:

look (at), *stare*, *gaze (at)*, *eye* and *eye up*, *leer*.

Now analyse the data and . . .

- find a genus expression for each word
- list all the meaning components which differentiate each word from the others
- taking account of the need for brevity, identify for each word those meaning components which you see as central (and which must therefore be shown in the definition) and those which could be omitted.

Exercise 2: Form and function in definitions

In two monolingual dictionaries of your choice, look up some entries for artefacts which have a particular function. These could include the following or their equivalents in the language of the dictionary:

chainsaw, *propeller*, *parachute*, *syringe*, *abacus*, *hammer*

Which definitions start by describing the object's *form*, and which foreground its *function*? Are there any cases where reversing the order would improve the definition? Which dictionary, in your view, has the better approach to this issue?

Exercise 3: The function of examples

In one or more dictionaries of English for advanced learners, look at all the examples for the following words:

crime, *advice*, *remember*, *decision*, *kill*

Can you say why each example has been chosen, and what linguistic point(s) it conveys?

Exercise 4: Selecting examples

Re-visit your corpus data for the verbs you analysed in Exercise 1.

On the basis of the data, select – for each verb – three example sentences that would be suitable for use in a monolingual dictionary for adult native speakers. As you do so:

- Identify for each verb two sentences from the corpus which, in your view, are completely unusable. Indicate why you think this.
- For each example you select for the dictionary, explain why you chose it, and indicate whether you modified it or used it exactly as it appeared in the corpus. If you modified any examples, explain what you did and why.

Reading

Recommended reading

Bolinger 1965, 1985; Fillmore 2003; Geeraerts 1990; Hanks 1979, 1987, 1994; Landau 2001: 153–216.

Further reading on related topics

Apresjan 1992; Atkins and Varantola 1997, 1998; Atkins, Rundell, and Sato 2003; Ayto 1988; Bogaards 1990, 1998a,b; Bullon 1990; Cowie 1981, 1999b (esp. 156–162); Fedorova 2004; Fillmore and Atkins 1994; Hanks 1988, 1990, 2000a, 2004b; Heyvaert 1994; Kilgarriff 1997b; Lew 2002, 2004; Lewandowska-Tomaszczyk 1990; Mackintosh 2006; Moon 1992, 1996, 2004; Nakamoto 1998; Norri 1996, 2000; Piotrowski 1988; Robins 1987; Rundell 1988, 1998, 1999; Rundell and Stock 1992; Scholfield 1999; Schutz 2002; Silva 2000; Stock 1988; Taylor 1990; van der Meer 1999, 2000, 2004; Veisbergs 2002; Verlinde, Dancette, and Binon 1998; Vilpula 1995; Walter 1992; Wierzbicka 1993; Zgusta 1971: 257ff.

Grammar in dictionaries: Bogaards 1990: 304–307; Bogaards and van der Kloot 2001; Cowie 1987b; Herbst 1996: 328–335; Lemmens and Wekker 1986.

New types of definition: Fillmore 1989; Hanks 1987; Lew and Dziemianko 2006; Rundell 2006; Sinclair 2004; Stock 1992.

Controlled defining vocabularies: Cowie 1987a: 14–24; Herbst 1996; Jansen, Mergeai, and Vanandroye 1987; Whitcut 1988.

Examples: Drysdale 1987; Fox 1987; Hausmann and Gorbahn 1989: 45–47; Humble 1998; Laufer 1992; Potter 1998; Cowie 1987a (esp. 93–96, 134–137); Rundell 1998: 334–335.

11

The translation stage

11.1 Transfer: Translating the database

'Transfer' is the term we use to describe the second stage of the three-stage lexicographic process. It consists of adding translations to the monolingual, target-language-neutral database described in Chapters 8 and 9. The transfer procedure described here is simply a more formalized version of the process that all bilingual dictionary editors go through at some point. First, in the *analysis* stage, they sketch out in as much detail as possible the full potential of the headword, distinguishing its various LUs and setting down the important facts about each, together with examples (this equates to stage 1, creating the database). Next, in the *transfer* stage, they work through each LU (both word senses and multiword expressions), adding target-language (TL) translations, going forward and back over the entry, and seeing which TL word seems to fit best as the first, or 'direct', translation – in essence, the word that suits most of the contexts before them. Then they decide which of the remaining contexts (those which the direct translation doesn't fit) are important enough to be kept in the entry, and translate the headword in these contexts. It is this transfer stage that we focus on in the present chapter. We deal with the final operation – the *synthesis* stage of a bilingual dictionary, in which dictionary entries are extracted and refined – in Chapter 12.

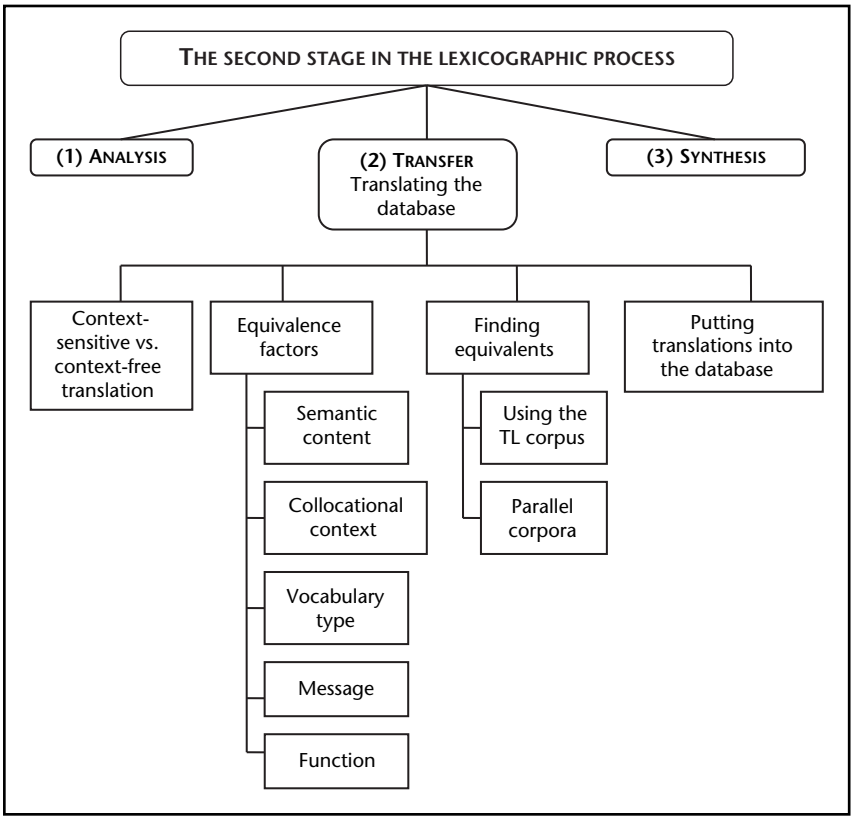


Fig 11.1 Contents of this chapter

Figure 11.1 provides an outline of the contents of this chapter.

This chapter looks at what is involved in the 'transfer' process – finding translations for the database material. There is an important difference between translating for dictionaries and the more familiar discourse- or text-translating. In the case of extended discourse, as opposed to dictionary entries, a good translation will produce language so idiomatic and natural-sounding that the reader may not be aware that it is a translation. It is well known, however, that on analysis the original texts (in dictionary terms, the source language, or SL) and the translation (the target language, or TL) rarely align perfectly, in that the sense of one individual word in the original is not exactly reflected in any corresponding TL word. Rather, the sense of a longer stretch of SL text is rendered in a corresponding piece of TL text.[1]

[1] For example, in two aligned sentences from the Canadian Hansard English and French corpus, we find *a unanimous report containing 18 recommendations* translated

The two types of translating are related of course, and in a way complementary, but lexicographers must never lose sight of the crucial difference between the translation of an English expression *in context* into a foreign language (as in books and documents) and the translation of an English expression *out of context* (as in dictionary entries). The first may be called *context-sensitive* translation, the second *context-free*. The lexicographer starts by producing a great number of translations of the headword in context, finally distilling from these translations the most suitable equivalent to appear as the 'direct translation'[2] of the headword in the entry. By 'most suitable' we really mean 'safest'. The direct translation must be as near context-free as possible. The dictionary will be used by people who have no idea of the meaning of any of the foreign words offered to them as a translation of the headword. Indeed, research (cf. Atkins and Varantola 1998) has shown that many dictionary users simply reach for the first TL word in the entry and use that in whatever context they have in front of them. It's our job to make the result as reasonable as possible. As users get older, or more practised, or simply more wary, they start to read the material we put in italics, or parentheses (or both), and they learn that further into the entry, among the examples and their translations, there are more subtle (context-sensitive) ways of translating the word they are looking up.

The distinction between context-free and context-sensitive translation lies at the heart of the task of putting translations into the database and is even more central to the selection of equivalents to be included in the dictionary entry proper (cf. §12.3.2).

11.2 Equivalence factors

The perfect translation – where an SL word exactly matches a TL word – is rare in general language, except for the names of objects in the real world (natural kind terms, artefacts, places, etc.), and even then it's not always plain sailing. Most of the 50,000 or so words you have to translate

by *un rapport unanime dans lequel on retrouvait 18 recommandations*. The sense of the English is excellently rendered into French, but no direct translation is offered of the single word *contain*. Such an option is not open to the editor of a bilingual entry: if *contain* is the headword, a TL equivalent must be found.

[2] See §7.2.4.1 for what we mean by this term, which refers to one of the components of a bilingual dictionary.

in the course of compiling a standard-size bilingual dictionary are going to present problems, some more daunting than others. The equivalence relationship between a pair of words, SL and TL, varies from exact to very approximate, from perfect to just-adequate, and the skill of the dictionary-writer lies first in selecting the best TL match available, and second in making sure that the SL-speaking, encoding users are aware of the pitfalls that lie in wait for them.

The relationships discussed in this section are between a *lexical unit* (a word or MWE in one of its senses) in the SL, and a lexical unit in the TL. It's a waste of time to try to plot out all the panoply of relationships between one SL *lemma* and all its possible TL equivalents, and vice versa. Translations of SL headwords are offered within an LU, that is, they are translations of the headword in a single sense. This situation avoids irrelevant discussions on, for instance, denotation versus connotation, since in the database a word denoting X and connoting Y will be recorded as having two separate LUs. Translating the headword within an LU means finding the best 'fit' as direct translation, and offering other translations (1) for the headword when it occurs within an example, and (2) for MWEs containing the headword. The factors which play a role in evaluating SL–TL equivalence are:

- semantic content (single words and MWEs)
- collocational context (mainly single words)
- vocabulary type (single words and MWEs)
- message (of phrases, including idioms and sayings)
- function.

The first four of these factors relate to *lexical items* while the last is principally of interest when you're looking for equivalents of *grammatical items* (cf. §6.2.1.1).

➔ Remember that everything we say here about equivalence refers only to the lexical unit (i.e. one usage of the headword or phrase), not to the headword itself.

11.2.1 Semantic content

There is considerable divergence in linguists' use of expressions like 'denotation', 'reference', and 'cognitive meaning', but all of these are included in the term 'semantic content', which designates the 'literal' meaning of

an expression together with its 'connotation' or any figurative meaning that may be associated with it. In lexicography, our aim is to find a TL expression whose semantic content matches as closely as possible that of the SL expression. The more fragmented the match, the less effective the translation. Two words denoting the same object form an exact match of semantic content, thus (in English and French) *tiger* and *tigre*, *London* and *Londres*, and *diamond* and *diamant*. Since these words are not affected by either 'collocational context', 'vocabulary type', or 'message' they are perfect bilingual partners. This is the usual relationship between SL and TL pairs of terms denoting objects in the real world, especially in specialist domains such as mathematics, medicine, and so on, but is rarely found in general language.

More commonly, even when collocational context and vocabulary type are shared, there is only a partial match between the semantic content of the SL headword and that of its TL translation. An example of this is the relationship between English *teacher* and the French *professeur*, the word that most people would think of as its translation. The similarities and differences within this pair are clearly seen in the extracts from their entries in *OHFD-3*, shown in Figure 11.2.

teacher *n* (*in general*) enseignant/-e *m/f*; (*secondary*) professeur *m*; (*primary*) instituteur/-trice *m/f*; (*special needs*) éducateur/-trice *m/f*;…

professeur *nm* (*enseignant*) (*de collège, lycée*) teacher; (*dans l'enseignement supérieur*) teacher, lecturer GB, professor US; (*titulaire*) professor; …

Fig 11.2 Semantic content: partial match

From these entries we see how essential the metalinguistic material is to the encoding user, i.e. the anglophone in the case of *teacher* and the francophone in the case of *professeur*. When dealing with such SL items which – in one single sense – legitimately require more than one translation, it is essential to make the TL details crystal clear when you are putting translations into the database, or choosing TL material for the dictionary entry.

➔ Match the semantic content of the SL and TL items: it's the most important thing of all in translating (except for the message of MWEs, cf. §11.2.4).

11.2.2 Collocational context

Collocational patterns are a powerful force in matching equivalents across languages. A monosemous headword may require quite a number of TL equivalents because of the way the SL and TL items collocate, as for instance in the case of the English headwords seen in Figure 11.3 in the extract from their *CRFD-8* entries, where the collocates of the English words produce quite different French translations, again because of collocational requirements in French. (Note that while this point is illustrated here by dictionary entries, it is just as important to specify these details when translating database items.) The nouns shown in square brackets in the *bunch* entry should be read as 'of flowers, watercress . . . ' etc., while those in the *grow* entry are typical subjects of the headword, and those in the *dark* entry typical of nouns modified by the headword.

bunch *n* [*flowers, watercress, herbs*] bouquet *m*; [*hair*] touffe *f*, houppe *f*; [*bananas*] régime *m*; [*radishes, asparagus*] botte *f*; [*twigs*] poignée *f*, paquet *m*; [*keys*] trousseau *m*; [*ribbons*] nœud *m* . . .

grow *vi* [*plant, hair*] pousser; [*person*] grandir; [*animal*] grandir, grossir; [*tumour*] grossir; [*crystal*] se former . . .

dark *adj* . . . (**c**) *complexion* mat; *skin* foncé; *hair* brun; *eyes* sombre . . .

Fig 11.3 The effect of collocation on the selection of TL equivalents

➔ The more SL collocates you put in, the easier it will be for people to choose from among several unfamiliar translations the one most likely to match the TL collocates.

11.2.3 Vocabulary type[3]

The semantic content of SL and TL items and their collocational needs are the most important factors in finding equivalent pairs, but you should always try to match up SL and TL items along the axis of vocabulary type. It's important to remember that register, style, region, attitude, and other vocabulary types can cause SL–TL mismatch. An informal SL expression should if possible be translated by an informal TL expression, a literary

[3] Vocabulary types are described in §6.4.1.4, and normally give rise to linguistic labels in a dictionary entry (see §7.2.8).

word by a literary word in the other language, a pejorative one by another pejorative one, and so on. When this is impossible, the differences must be specified by database translator and dictionary editor. Sometimes an SL headword will require two TL equivalents because of regional differences in the target language. This is the case for the French noun *trottoir*, translated by *pavement* in British English, and *sidewalk* in American English, and again the TL regions must be specified in both database translation and dictionary entry, as in the *OHFD-3* entry for *trottoir* which reads '**trottoir** *nm* pavement GB, sidewalk US'.

➔ Matching the vocabulary type of SL and TL items helps people to sound more natural in the foreign language.

11.2.4 Message

'Message' denotes the underlying meaning of a phrase, as opposed to what it literally means. The term includes much of what is called elsewhere *pragmatic force*. It is possible to find pairs of SL and TL idioms or sayings which match both in *semantic content* and in *message*, as for instance:

- English *all's well that ends well*
  French *tout est bien qui finit bien*

When semantic content and message diverge, then the latter must prevail. SL idioms and sayings and their TL partners must match in message, and often the semantic content of the expressions may differ totally. Such is usually the case of proverbs and many sayings across languages, as for instance:

- English *birds of a feather flock together*
  French *qui se ressemble s'assemble*, literally: 'people who are like each other congregate together'
- English *can you beat it!* (informal)
  French *faut le faire!* (informal), literally: 'someone has to do it' (ironic)

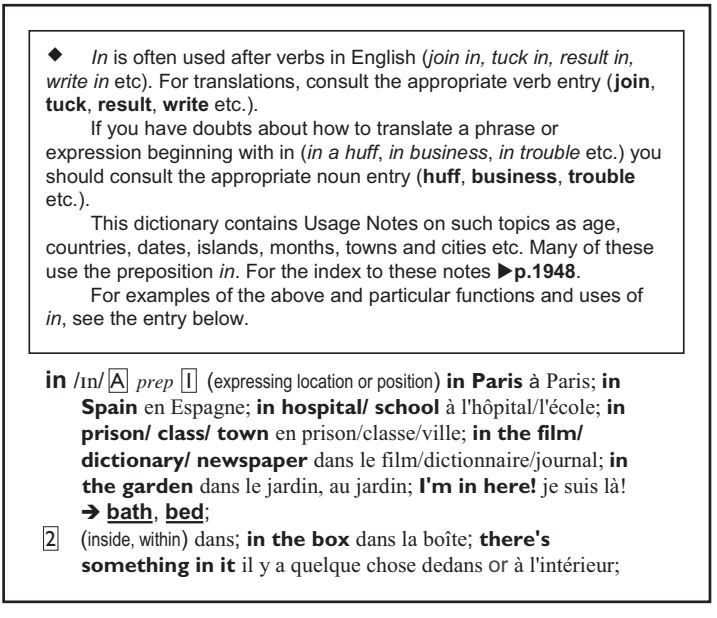
Many SL proverbs and idioms will not find a matching TL partner, and in such cases must be glossed, both in the database and the dictionary, as for instance:

- English *too many cooks spoil the broth*
  French *on n'arrive à rien quand tout le monde s'en mêle*, literally: 'you never achieve anything when everyone pitches in'
- English *to save the day*
  French *sauver la situation*, literally: 'to save the situation'

➔ Only the message really matters when it comes to translating idioms and sayings.

11.2.5 Function

In the case of the grammatical words of the language, also aptly called 'function words', the semantic content is not the whole of the picture. Just as important in a bilingual dictionary is the function of the word (the role it plays in expressing or interpreting the meaning of a phrase or sentence), together with its collocational context.



◆ *In* is often used after verbs in English (*join in, tuck in, result in, write in* etc). For translations, consult the appropriate verb entry (**join**, **tuck**, **result**, **write** etc.).

If you have doubts about how to translate a phrase or expression beginning with in (*in a huff*, *in business*, *in trouble* etc.) you should consult the appropriate noun entry (**huff**, **business**, **trouble** etc.).

This dictionary contains Usage Notes on such topics as age, countries, dates, islands, months, towns and cities etc. Many of these use the preposition *in*. For the index to these notes ▶**p.1948**.

For examples of the above and particular functions and uses of *in*, see the entry below.

in /ɪn/ A *prep* I (expressing location or position) **in Paris** à Paris; **in Spain** en Espagne; **in hospital/ school** à l'hôpital/l'école; **in prison/ class/ town** en prison/classe/ville; **in the film/ dictionary/ newspaper** dans le film/dictionnaire/journal; **in the garden** dans le jardin, au jardin; **I'm in here!** je suis là!
➔ **bath**, **bed**;
2 (inside, within) dans; **in the box** dans la boîte; **there's something in it** il y a quelque chose dedans or à l'intérieur;

Fig 11.4 Function and collocation in a grammatical entry from *OHFD-3*

The beginning of the *OHFD-3* entry for the preposition *in* is a good example of this, cf. Figure 11.4. A summary of some of the principal functions of this word prefaces the entry proper, and even in the first part of the entry there is no overt mention of the semantic content of this preposition. Rather, it opens with a statement of the function of the word in that sense, which is to express 'location or position', and continues by showing how it is translated in specific types of contexts: names of cities and countries, institutions, printed works, and so on. Only in the second LU do we find mention of the word's semantic content 'inside, within'.

The entry shown in Figure 11.4 is typical of a grammatical word entry in a good bilingual dictionary. It is clear from this that the approach to finding

equivalents for function words in the database must be quite different from the way we go about finding them for lexical items. Because these words are so language-specific (for instance, many of their functions are carried out in other languages by cases) we shall not discuss function words and their entries any further in this volume.

11.3 Finding equivalents

Translators always start with some good ideas about how to translate words and phrases, but everyone has moments of doubt. Scanning bilingual dictionaries and checking out one's intuitions with a native speaker of the language that is not your own have traditionally been the way to deal with such doubts. Indeed, until quite recently these were the only options open to bilingual dictionary editors. Now of course the world has changed, and we can use corpus data to widen our translating horizons. The database itself was written on the basis of a monolingual SL corpus. Another monolingual corpus – the TL corpus – comes into its own at the translating stage, as does the bilingual corpus, if you are lucky enough to have one (and the time to use it).

11.3.1 Using the TL corpus

The TL corpus has immense potential for dictionary translators, particularly those without the benefit of a native-speaker informant. It offers a way of finding translations, of checking those you are doubtful about, and of correcting those that are simply wrong.

11.3.1.1 *Finding translations* The SL corpus concordances for *measure* in Figure 11.5 show how important it is to include the phrase *measures aimed at doing*... in the dictionary. A cursory glance at similar lines for *mesure*,[4] sorted on the right context, offer an instant, excellent translation, *mesures destinées à faire*...

11.3.1.2 *Checking translations* The English phrase *for good measure* suggests at once *pour faire bonne mesure*, but even in such apparently straightforward cases it's worth carrying out a routine check. The TL corpus offers many examples of *pour faire bonne mesure* (cf. Figure 11.6), and confirms the match between the two phrases.

[4] The French concordances come from OUP's Oxford French Corpus.

Most of the	measures	aimed at competing in the single market
and that all	measures	aimed at countering misunderstanding
the government announced	measures	aimed at curbing the black market
in relation to the	measures	aimed at implementing the Social Charter
specific transport	measures	aimed at making real cuts in warming gases.
the use of	measures	aimed at modifying pathological processes
the adoption of	measures	aimed at promoting these values
other coercive	measures	aimed at redistributing wealth
a battery of	measures	aimed at reducing speeds on main roads
There are also	measures	aimed at reducing street litter
a £50m package of	measures	aimed at speeding up the postal service.
emergency	measures	aimed at stopping a mass rally
parliament successively passed	measures	aimed at weakening the church.
les	mesures	destinées à empêcher sa propagation
les	mesures	destinées à encourager l'activité des . . .
s'accompagner de	mesures	destinées à maîtriser l'évolution des dépenses.
de prendre de toute urgence	mesures	destinées à mettre fin à . . .
opérations complétées par des	mesures	destinées à permettre le départ des . . .
prendre de nouvelles	mesures	destinées à relancer l'appareil économique.
les	mesures	destinées à s'attaquer aux racines du problème
prendre des	mesures	destinées à sanctionner l'organisateur présumé . . .
les élus ont voté diverses	mesures	destinées à sauvegarder l'environnement.

Fig 11.5 Concordances for *measures aimed at* and *mesures destinées à*

11.3.1.3 *Correcting translations* In many cases, however, there is not such a neat fit. The phrase *in full measure* occurs many times in the SL corpus, in very varied contexts. The word-for-word translation of *full measure* is

Take your licence along for good	measure	The national coach will give you
There are even a few herrings for good	measure,	though they are . . .
the British guns joining in for good	measure	
pudding with a little cocoa powder for good	measure	. Light in texture, it is sublime served
And for good	measure	it should be noted they had long supported
the old system and added a new one for good	measure	
transfer to Syria, adding, for good	measure,	that the Syrian regime is as bad as . . .
Mr Yavlinsky did so, and for good	measure	outlined a programme for radical reform
Shivering, he added two sweaters for good	measure.	
ajoutant pour faire bonne	mesure	qu'une . . .
un sac de plastique . . . pour faire bonne	mesure,	des câbles sont noués entre deux blindés . . .
Et pour faire bonne	mesure,	il décrit une situation apocalyptique. . .
et, pour faire bonne	mesure,	ils ont franchi la ligne presque roue dans roue
Pour faire bonne	mesure,	le commerçant avait ajouté le livre de Charriére
pas plus, pour faire bonne	mesure,	que le film de Michel Deville
Pour faire bonne	mesure,	Renault, comme Peugeot, dispose d'un arsenal de. . .
Pour faire bonne	mesure,	sa carte de séjour, n'a pas . . .
Le moulin et, pour faire bonne	mesure,	treize hectares de terrain autour.

Fig 11.6 Concordances for *good measure* and *bonne mesure*

pleine mesure, which instinct suggests is worth testing, and the TL corpus produces the examples shown in Figure 11.7 (and many more).

he receives it suddenly and in full	measure,	above the groundswell of heckling, at
the young woman can feel it in full	measure.	
shall redeem our pledge, not in full	measure,	but very substantially.
they had all these qualities in full	measure,	and yet . . .
my expectations were met in full	measure,	as I think you may ascertain
and possessing in full	measure	the strong will and harsh determination
It was a sentiment shared in full	measure	by combatants of both sides at Verdun.
. . . live long enough to know in full	measure	the contempt in which you are held
He had in full	measure	the energy of most boys of that age
he had in full	measure	that fear of the unknown which . . .
my sympathy went out in full	measure	to those involved
on peut prendre la pleine	mesure	des enjeux du rapport entre . . .
à prendre la pleine	mesure	du problème
la difficulté à prendre la pleine	mesure	des enjeux écologiques
personne n' avait encore pris la pleine	mesure	de ce pavé de plus de mille pages.
. . . à appliquer dans leur pleine	mesure	les principes de justice
la vraie musique écoutée dans sa pleine	mesure	n'en sonnera que meilleure.
. . . donner leur pleine	mesure	sans se heurter à aucune limite ...
. . . n'ont pas encore donné leur pleine	mesure,	et les arguments politiques
Il acquiesça de la tête, prenant pleine	mesure	des risques encourus.

Fig 11.7 Concordances for *full measure* and *pleine mesure*

It's clear at once that the *pleine mesure* examples are not going to provide a good translation of *in full measure* (*prendre pleine mesure de* would be translated by 'get the measure of, fully grasp'). When you look again at the English contexts in the database the adverbs *pleinement* and *entièrement* come to mind, and the contexts in which these are found make it clear that both are needed. The eventual dictionary entry will have to show typical contexts to help the English speaker choose the right one, such as 'feel, possess, fulfil, contribute *pleinement*; repay *entièrement'*.

A different problem arises in the case of *beyond measure* and its literal translation *outre mesure* (see Figure 11.8). The corpora prove these to be *faux amis*: the French *outre mesure* does have the meaning of 'excessively' but is almost always found in negative contexts, and never equates to *beyond measure*, which is rather the equivalent of adverbs like *extrêmement* or *énormément*. A further look at the English contexts, where *beyond measure* modifies adjectives (including past participle adjectives) as well as verbs, makes it clear that one single TL translation is inadequate, and that the dictionary entry will have to include a formulation like '*change, increase* énormément; *anxious, difficult* extrêmement'.

I've been comforted beyond	measure	by these words.
Shocked and distressed beyond	measure,	her worst forebodings realized . . .
Max was embarrassed beyond	measure	and realised that . . .
That hurt me beyond	measure.	
Exasperated beyond	measure,	he threw down his chequebook . . .
This year Edberg has improved beyond	measure.	
the wealthy Da Gamas, rich beyond	measure	from the spice trade
"Get back to bed!" Irritated beyond	measure	at these events, he . . .
His presence puzzled her beyond	measure.	
That notion disturbed him beyond	measure.	
I am anxious beyond	measure	to be in the country
It would hurt beyond	measure	should I lose her.
Gérard ne s'inquiete pas outre	mesure.	
ils ne semblaient pas l'inquiéter outre	mesure.	
. . . n'a pas impressionné M. Mandela outre	mesure.	
L'opinion belge n'en serait pas étonnée outre	mesure:	elle a l'habitude des compromis
il n'y a pas de quoi s'émouvoir outre	mesure.	
Mais Mitterrand ne semble pas outre	mesure	inquiet sur l'issue de la bataille.
l'affaire ne semble pas intéresser outre	mesure	la police judiciaire.
. . . de sauver la face, sans compromettre outre	mesure	la suite du processus de paix.
. . . qui ne devrait toutefois pas retarder outre	mesure	le bouclage de son dossier.
. . . bataille navale, sans émouvoir outre	mesure	lesdits marchés.

Fig 11.8 Concordances for *beyond measure* and *outre mesure*

11.3.2 Parallel corpora

The term *parallel corpus* denotes a set of corpora (two in a bilingual parallel corpus, more in a multilingual version) in which the texts in Language A correspond in some way to those in Language B (and perhaps C and D and so on). Two types of parallel corpus can be useful when you are trying to find a good translation: a *translation corpus* and a *comparable corpus*: their differences are summarized in Figure 11.9.

11.3.2.1 *The translation corpus* In a *translation corpus*, the two corpora consist of translated texts, which means of course that only 50 per cent of the texts are originals, the rest being translations. Nonetheless, a translation corpus is a rich source of equivalence material and easy to use, as software exists to align pairs of sentences, one from each language corpus. Figure 11.10, where the material is taken from the bilingual Canadian Hansard corpus,[5] shows a selection of the sentences offered in response to a search for the English verb *echo*. The aligned sentences make it easy to spot a word–word or at least a phrase–phrase equivalence.

[5] This material was extracted using TransSearch software, from the University of Montreal; see http://www.terminotix.com/eng/index.htm .

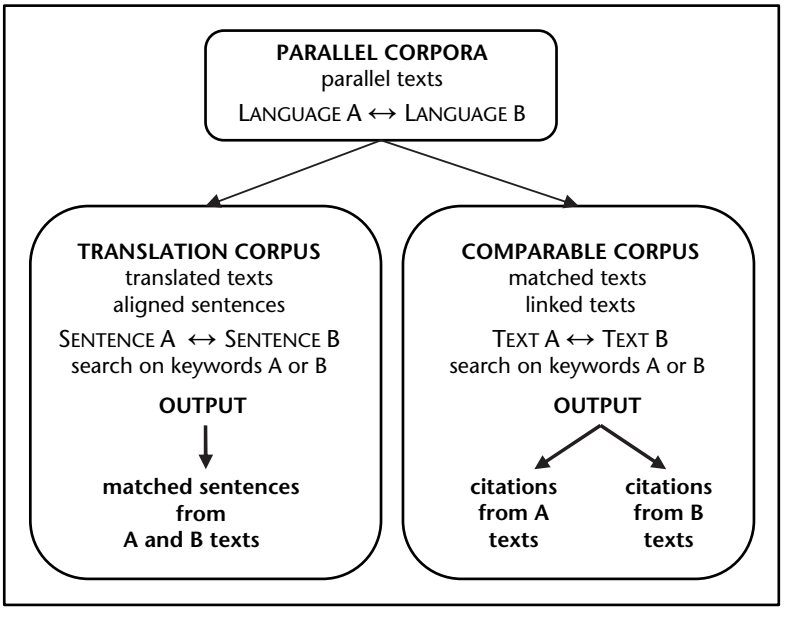


Fig 11.9 Two types of bilingual parallel corpora

These sentences contain various related uses of this English verb, and set against each one is its equivalent in French, the work of the official Hansard translators. As well as the standard equivalents *se faire l'écho de* (in 1, 2, and 4) and *faire écho à* (6, 14, and 15), we see a number of phrases which could prove useful to advanced linguists: *en écho à ce que le ministre vient de dire* . . . (7 'to echo what the minister just said . . . '), *reprendre à son compte* (3 'adopt'), *reprendre les propos du ministre* (12 'associate oneself with what the minister said'), *réitérer* (8 'reiterate'), *refléter* (11 'reflect') and indeed *abonder dans le même sens* (5 'agree wholeheartedly'). All these are grist to the lexicographer's mill, especially since electronic dictionaries will be able to relax space restrictions. Only examples 9, 10, and 13 have nothing to offer. However, the price that would have to be paid if editorial teams were to use bilingual corpora is too high.[6] The pros and cons of using such corpora for dictionary production are:

[6] An appeal in January 2007 on the EURALEX discussion list for information about any dictionary publisher using a bilingual corpus in the editing of a bilingual dictionary produced no affirmative responses, but several working lexicographers commented on how useful such corpora could be.

1.	I have to **echo** his sentiments	je me dois de **me faire l'écho** de ses sentiments.
2.	I will **echo** my Conservative predecessor.	À ce sujet, **je me ferai l'écho de** mon prédécesseur conservateur.
3.	I **echo** my colleague's comments . . .	**Je reprends également à mon compte** les propos de mon collègue . . .
4.	I want to **echo** the same message.	**je me fais l'écho de** son message.
5.	I **echo** the minister's comments in this regard.	**j'abonde dans le même sens**.
6.	Let me **echo** the declaration made by the Deputy Prime Minister	**Je ferai écho aux** paroles qu'a prononcées la vice-première ministre
7.	Mr. Speaker, I will **echo** what the minister just mentioned and say that . . .	Monsieur le Président, **en écho à** ce que le ministre vient de dire, **j'affirme que** . . .
8.	I would certainly **echo** the concerns of my colleague	je **réitérerais** certainement les préoccupations de mon collègue
9.	The member's words **echo** hollow in this chamber, literally and figuratively.	Les paroles du député **créent** peut-être **un écho** dans cette enceinte, mais elles sonnent creuses.
10.	issues . . . which I hope will **echo** in all four corners of the House.	J'espère que ses propos seront entendus aux quatre coins de la Chambre.
11.	In this respect, Grand Chief Fontaine's comments **echo** those that one would find, for example, in the . . .	À cet égard, les commentaires du grand chef Fontaine **reflètent** ceux que l'on retrouve, par exemple, dans . . .
12.	I repeat the words of the Minister which . . . **echo** what Jean Chrétien said . . .	**Je reprends les propos du** ministre qui utilise les mêmes mots qu'utilisait Jean Chrétien . . .
13.	where my words could **echo** the passion of my thoughts	où j'ai pu m'exprimer aussi haut que je pensais fort
14.	I do not think Liberals will **echo** the views and mirror the policy of the NDP.	Je ne crois pas que les libéraux **fassent écho à** la politique du NPD.
15.	I am sure premiers across the country would **echo** the statement of the premier of Manitoba.	je suis sûr que tous les premiers ministres **feraient écho à** la déclaration du premier ministre du Manitoba.

Fig 11.10 Equivalences of the verb *echo* in an English-French translation corpus

Pros

- no more hunting for equivalence candidates
- a wealth of context-sensitive translations
- contexts for all equivalence candidates

Cons

- too many equivalence candidates
- every one of them seems essential to lexicographers at that point in the editing
- the production line grinds to a halt
- the dictionaries are too big to appear in print
- the entries contain too much detail for most users.

11.3.2.2 *The comparable corpus* This corpus is made up of two individual language corpora, selected on the basis of at least one shared parameter, usually the subject matter, together with possibly other properties shared by the texts, such as the date and/or the medium (books, newspapers, conversations, etc.). An example of a bilingual comparable corpus might be one containing accounts of the same event drawn from leading quality newspapers in (say) the UK and Germany. This type of corpus provides excellent material for the translator because all the texts are original, and no translation is involved. For bilingual lexicographers it could be a rich source of inspiration, but, because the matched corpora can only be searched individually, the output is not economic to use in a serious dictionary project.[7]

11.4 Putting translations into the database

Just as the database was designed to be as rich as possible, holding all the most frequent contexts of the headword in its various senses, together with the grammar needed to use it correctly, so the translations in the database are designed to cover the whole spectrum of possibility. The reasons for doing lexicography this way are as follows:

- Everything a dictionary editor needs to know about the word before writing a bilingual entry is assembled in an orderly way that is easy to use and that allows the editor to get a fix on the word without reading through acres of concordances.
- This is a fast and effective way to compile a dictionary, because it exploits the skills of three distinct groups of people:
  - The monolingual database is compiled (in stage 1) by editors with lexicographic but not necessarily translating skills.
  - The translations are inserted (in stage 2) by skilled translators, who are not necessarily lexicographers.
  - The dictionary entries are edited (in stage 3) by skilled bilingual lexicographers.

As a result, stage 2 is not standard translating procedure. Because of the role that the translated database plays in the process of entry-writing, you don't have to translate every source-language item in the entry. You have to find a good 'direct' translation for the headword, to sit at the top of the LU

[7] This type of corpus is, however, of great help to translators of specialist terminology, especially those writing in a language not their own.

subentry. You then work systematically through the database entry reading each corpus example in turn: if the direct translation fits that context, mark it as OK and move on. Even when you come to an example where the direct translation can't be used, you don't have to translate the whole of the sentence, simply the 'core' of it, around the headword. This document is for editors' eyes only, and does not need to be polished. The aim here is to work accurately but fast, so that when the dictionary editors come to write their entry, they will have at their fingertips all the facts they need about the headword and its possible TL equivalents.

➔ When you give two TL equivalents for a particular use of the headword, make sure you explain how the TL terms differ from each other in sense, style, register, etc. (unless they are 100 per cent synonyms, which are rarer than hen's teeth).

Part of the first LU of the database entry for *bargain* (introduced to illustrate points in §9.2.6.2 ff.) is shown in Figure 11.11. For the transfer operation, this is 'opened up' so that translations may be added, together with any comments from the translator. Figure 11.12 shows what the translation process adds to this entry.[8] The examples in both figures are numbered for ease of reference, and in Figure 11.12 comments from the translator are indicated by [TR].

At the top of the entry, in the MEANING field, the database editor indicates informally which of the senses of the headword is being treated here (*transaction*, *deal*, *agreement*, etc.). After reading through the whole LU (and probably making notes along the way) the translator decides that *marché* is the best direct translation, and inserts that in the first TRANSLATION field. Since *marché* fits examples 1 and 2, the translator simply marks these as 'OK' in the TRANSLATION fields following each. However, *marché* is not a good translation for *bargain* in example 3, so the translator offers *contrat* and/or *accord* in its place. Example 4 contains the phrase *his part of the bargain*, for which the translator offers *sa part du marché*. Note that only the 'core' of the example sentence is translated each time. The phrases *one's part of the bargain* and *one's half of the bargain* appear in the context of the collocate verb *keep* in examples 5, 6, and 7. The different contexts in 5 and 6 lead the translator to find two different equivalents (*tenir sa part du marché* for 5, and *tenir sa parole* for 6). For example 7, however, the translator proposes a different translation (*j'ai fait ce qu'on attendait de moi*: literally, 'I've done what was expected of me'), adding a comment

[8] Our thanks go to Valerie Grundy for work on the English entry and the translations here.

	LEMMA	**bargain**
	LU #	**1**
	WORDCLASS	noun
	MEANING	transaction, deal, agreement between people to do something
1	EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
2	EXAMPLE	I have a bargain to offer you, sir.
3	EXAMPLE	A credit agreement could be re-opened, if the court thought just, on the grounds that the bargain was extortionate.
4	EXAMPLE	His part of the bargain is to devise methods of teaching subjects such as physics, geography or mathematics in the context of the Royal Mail's work.
	COLLOCATE	keep
5	EXAMPLE	If the sellers (in this case the manufacturer or wholesaler) do not keep their half of the bargain, the contract is broken.
6	EXAMPLE	We've got to keep our half of the bargain – we did say we'd try to persuade her.
7	EXAMPLE	I've kept my half of the bargain, though I never guessed how costly it would be for me.
	SUPPORT-VERB	make
8	EXAMPLE	Hercules arrived and made a bargain with the King
	COLLOCATE	strike
9	EXAMPLE	Buyer and seller strike a bargain with each individual purchase.
10	EXAMPLE	In return for her help the spies strike a bargain with her: she must . . .

Fig 11.11 Part of the database entry for the first LU of *bargain*

to the effect that for this particular first-person use it sounds more natural than the rather stilted *j'ai tenu ma parole* ('I have kept my word'), which otherwise might be extrapolated from the previous translation. Example 8 shows the collocate *make* as context and produces the translation *conclure un marché*, as does *strike* in example 9. However, before that the translator has proposed an extra example (8a), *let's make a bargain!*, because he or she realized that the rather formal phrase with *conclure* would not fit the context of this fairly common usage. Instead, a different, less formal French phrase with the same pragmatic meaning is proposed, *on va se mettre d'accord!* (literally, 'we're going to make an agreement').

From this brief analysis, the role of the translator is seen to be more proactive than perhaps might have been expected. Comments and

	LEMMA	**bargain**
	LU #	1
	WORDCLASS	noun
	MEANING	transaction, deal, agreement between people to do something
	TRANSLATION	**marché m**
1	EXAMPLE	Angelo offers her a bargain: if she will sleep with him her brother shall live.
	TRANSLATION	**OK**
2	EXAMPLE	I have a bargain to offer you, sir.
	TRANSLATION	**OK**
3	EXAMPLE	A credit agreement could be re-opened, if the court thought just, on the grounds that the bargain was extortionate.
	TRANSLATION	**contrat m or accord m**
	COMMENT	you can't really 'reopen' a marché, that's more the actual transaction [TR]
4	EXAMPLE	His part of the bargain is to devise methods of teaching subjects such as physics, geography or mathematics in the context of the Royal Mail's work.
	TRANSLATION	**sa part du marché**
	COLLOCATE	keep
5	EXAMPLE	If the sellers (in this case the manufacturer or wholesaler) do not keep their half of the bargain, the contract is broken.
	TRANSLATION	**tenir sa part du marché**
6	EXAMPLE	We've got to keep our half of the bargain – we did say we'd try to persuade her.
	TRANSLATION	**tenir sa parole**
	COMMENT	part du marché only OK for formal agreements, purely commercial deals etc. [TR]
7	EXAMPLE	I've kept my half of the bargain, though I never guessed how costly it would be for me.
	TRANSLATION	**j'ai fait ce qu'on attendait de moi**
	COMMENT	sounds more natural than 'j'ai tenu ma parole' [TR]
	SUPPORT-VERB	make
8	EXAMPLE	Hercules arrived and made a bargain with the King
	TRANSLATION	**conclure un marché avec**
	COMMENT	but this is fairly formal so how about the sentence below?
8a	EXAMPLE	let's make a bargain!
	TRANSLATION	**on va se mettre d'accord!**
	COMMENT	again – more natural [TR]
	COLLOCATE	strike
9	EXAMPLE	Buyer and seller strike a bargain with each individual purchase.
	TRANSLATION	**conclure un marché avec**
10	EXAMPLE	In return for her help the spies strike a bargain with her: she must . . .
	TRANSLATION	**OK**

Fig 11.12 Translated database for *bargain* sense 1

contributions, even suggestions for other examples, are offered when the translator thinks they might be useful to the editor of the dictionary entry, who is next in line to use the database.

➔ When you're putting translations into the database, remember that one way or another *everything* must have a translation: if it doesn't work with the direct translation, then translate the keyword core of the sentence.

Exercises

These exercises build on the database entries you created in Exercise 1 (1a–1d) at the end of Chapter 9. In each case the objective is to supply translations for your database material.

Exercise 1

- Choose a target language: this should be a language you know well.
- Using the database entry you wrote for your verb headword, insert translations as required (see §11.4 above).

Exercise 2

Do the same for the noun headword.

Exercise 3

Do the same for the adjective headword.

Exercise 4

Do the same for the adverb headword.

Reading

Recommended reading

Fillmore and Atkins 2000; Apresjan 1992; Atkins 1994; Corréard 1998.

Further reading on related topics

Adamska-Salaciak 2006; Bowker 2006; Citron and Widmann 2006; Cummins and Desjardins 2002; Dobrovol'skij 2000; Duval 1991; Heylen and Maxwell 1994; Leemets 1992; Lew 2002, 2004; Lewandowska-Tomaszczyk 1988; Roberts and Bossé-Andrieu 2006; Roberts and Montgomery 1996; Sharpe 1989, 1995; Sinclair et al. 1996; Tognini-Bonelli 1996; Varantola 2006; Wakely 1998.

Websites

Paraconc: bilingual/multilingual concordancer http://www.athel.com/para.html
Canadian Hansard bilingual corpus http://www.terminotix.com/eng/index.htm

12

Building the bilingual entry

In this chapter we guide you through the process of compiling entries for a bilingual dictionary. An outline of what is covered is given in Figure 12.1. Our objective in this chapter is to explain the lexicographic techniques needed for writing bilingual entries, and to do this we need illustrative material. Like the database explained in Chapter 9, our source language in this exercise is English. Our target language, for illustrative purposes, is French.

As in the case of the monolingual dictionary (cf. Chapter 10), our starting point is the database, constructed during the initial 'analysis' stage of lexicography, and populated in the way we described in Chapters 8 and 9. Each lemma in the database comes with a structured inventory of corpus-derived facts, and it is from these that the final dictionary entries will be distilled. In Chapter 8, we explained the criteria by which lemmas are divided into 'lexical units' (LUs). An LU is a bundle of information about either the headword in one of its senses or some type of multiword expression (idiom, phrasal verb, and so on: §7.2.7.1). For every LU, the database provides the following kinds of information:

- a rough characterization of its meaning
- a detailed record of its combinatorial behaviour, including:
  - syntactic patterns (§9.2.5)
  - MWEs in which it participates (§9.2.6)

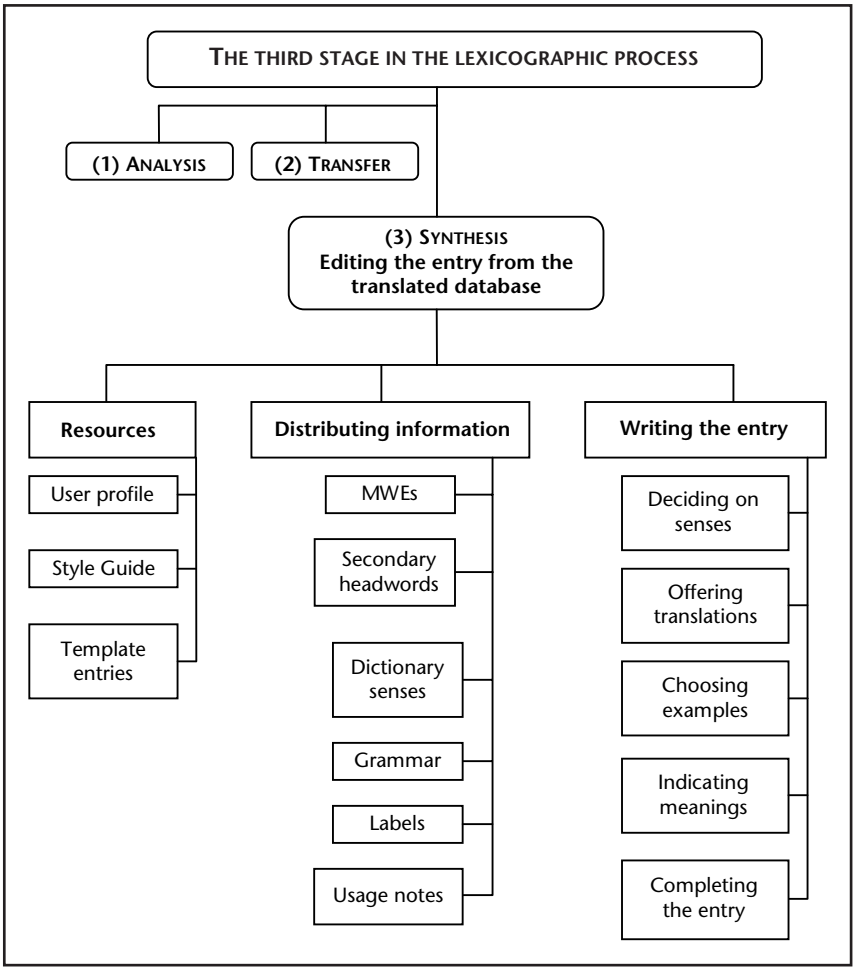


Fig 12.1 Contents of this chapter

  - lexical collocations (§9.2.7)
  - corpus patterns (§9.2.8)
- an indication of any stylistic, regional, subject-field, or other features that require a linguistic label (§9.2.9)
- one or more examples from the corpus to illustrate each individual fact which the database records.

The next stage ('transfer') entailed inserting translations into the database, and this process formed the subject of Chapter 11.

The third stage in the process ('synthesis') is the focus of this chapter. It consists of transforming the translated database records into a series of

finished entries for a specific bilingual dictionary. This involves a process of selection and presentation: selection of facts relevant to one particular dictionary and appropriate to one particular group of users. With a carefully designed and populated database, you already have all the information you need to create finished entries. The syntactic, collocational, and sociolinguistic data is already logged and supported by example sentences, so you won't – as a rule – need to go back to the corpus.

Before the actual entry-writing begins, however, there are several preparatory operations, which we will discuss in the first part of this chapter. We start at the beginning of the entry-writing process, and look at the resources you need in order to do the job successfully (§12.1), and the major decisions to be made about the distribution of information before entries can be written (§12.2). After these preliminaries, the remainder of the chapter (§12.3) deals with compiling the actual bilingual dictionary entry.

12.1 Resources for entry-building

In addition to the database itself, three other resources come into play at this stage:

- the user profile
- the Style Guide
- template entries.

We will briefly consider how each of these impacts on the entry-building process.

12.1.1 The user profile

First, catch your user. The user profile critically affects both what goes into the entry and how it is presented. A detailed user profile (see §2.3.1) will underlie the major design decisions of your dictionary, and will be reflected in the Style Guide, but even so, when you're writing an entry it's important to have some clear idea in your own mind about what you are expecting your users to be able to do and what they will be using the dictionary for. As explained in §2.4.2, a bilingual dictionary may cater for two types of users:

- speakers of the source language
- speakers of the target language.

It may thus be used for two different purposes:

- encoding (by the SL speakers), i.e. translating into, or expressing themselves in, the foreign language
- decoding (by the TL speakers), i.e. translating out of the foreign language into their own language.

Decoding is almost always easier than encoding, and so – as we saw in §2.4.2 – an entry written for the SL-speaking user has to be much fuller than one for the TL speaker. In §12.3 we'll focus on writing a dictionary entry designed to help the SL speaker to get around in a foreign language with as few mistakes as possible.

Even more than its monolingual sister, a bilingual dictionary is a tool. And because it's a tool (not an archive, or a record, or an account of the language) your main purpose in writing the entries is to make it as easy as possible for users to find the TL expression they need and to use it correctly. People who want to know how a language works need a monolingual dictionary of that language. You are writing for people who want to use the dictionary as a launchpad into another language. Important decisions must be based above all on putting the user first. A well-defined user profile helps us make the right decisions about *content*, affecting areas such as:

- Headword selection: we need to ask ourselves, for example, does our user need vocabulary items that are dated, literary, or highly technical? Are our users likely to be tourists rather than students, and so need rare words like *hydrofoil* or *verbena* or *alabaster* but not equally rare words like *denotation* or *subjunctive*?
- Sense selection: similar questions apply – we have to decide how helpful it is to our users to include rare or literary uses of a word.
- Granularity of senses: does our user need a finely split description of a word's different uses, or will a broadbrush treatment be more helpful (cf. §8.1.3)?
- Granularity of labels: the inventory of labels used in a large, unabridged volume may be quite extensive (for example, covering specific subject-fields like 'anatomy' and 'physiology'), whereas in a concise or pocket dictionary a small set of broader labels (such as 'medical') may be more appropriate.

- Grammatical and syntactic information: native speakers don't generally need to be told that *knowledge* is an uncountable noun (and can't be pluralized), or that *prevent* is typically used in the pattern *prevent sb from doing sth* (rather than **prevent sb to do sth*). But if the user is a language-learner, this is essential information. We might even go so far as to decide that most of our users won't understand terms like *transitive* or *countable*, and dispense with them.
- Examples: some types of dictionary contain very few examples, others make extensive use of them, while others again (think of Johnson or the *OED*) use only attributed citations (§10.8.2). Which of these options we choose will depend on what we know about the users' needs.

Similarly, the *presentation* of information should be guided by an understanding of users' reference skills, knowledge of the world, and linguistic competence. This can make a big difference in areas such as the following:

- The dictionary's metalanguage and conventions: will users understand abbreviations like *colloq*. or *dial*.? Can we assume they know the International Phonetic Alphabet? Will they be able to cope with lexicographic conventions like the specialized use of brackets in definitions? Will they see – far less understand – the difference between font sizes and types, which in the normal dictionary carry quite a heavy burden of information?
- Translations: how can we lead users to the appropriate translation?
- If the user needs grammatical information, what form should it take, and to what depth should it go? Can we expect the user to understand transitivity or countability, or even basic grammatical categories like subject and object?

Figure 12.2 illustrates the impact of user profiling, by comparing entries for one word in two different dictionaries derived from the same source text. The *OHFD* is a large collegiate dictionary designed for use by language students and other linguists; the *Concise OHFD* is aimed at a much less sophisticated user group, who would hardly need to be able to translate *ligament*, far less use it in complex contexts. Here, awareness of the users' needs is reflected in the type of entry accorded to this rather technical word.

ligament /. . . / **I** *n* ligament *m*; **knee/ankle ~** ligament du genou / de la cheville; **torn / strained ~** ligament déchiré / froissé. **II** *modif* [*tissue*, *fibre*] ligamenteux / -euse; [*trouble*, *injury*] ligamentaire.

OHFD-1 (1994)

ligament /. . . / *n* ligament *m*.

Concise OHFD (1995)

Fig 12.2 Entries designed for different users

12.1.2 The Style Guide

The Style Guide, as we saw earlier (§4.4), is a set of instructions for handling every aspect of the microstructure. (The fact that the word *ligament* in Figure 12.2 is handled quite differently in two English-French dictionaries is a direct result of differences in policy, embodied in the Style Guides of these dictionaries.) The Style Guide provides a detailed description of the editorial policy decisions made at the outset of the project – decisions on how to deal with each of the entry components discussed in Chapter 7. Those decisions, in turn, reflect our understanding of the needs and capabilities of the intended user. The Style Guide's principal function is to make the dictionary consistent, in both content and presentation, no matter how many editors are on the team or how long the dictionary takes to compile. This has benefits for both lexicographers and dictionary users: a well-thought-through set of editorial policies which reflect a coherent ethos will be easier for the editorial team to assimilate, while users will quickly learn the best way to find what they are looking for. The dictionary's policies on all the topics discussed in §12.2, and many more, will be set out in detail in its Style Guide. By the time the editorial team is trained and ready to start entry-writing, the senior editors will have written a hundred or more sample entries, covering all wordclasses and encompassing most of the known problems for bilingual dictionaries, and on the basis of that experience your Style Guide will be written. Some of the decisions will be built in to the dictionary writing system (§4.3.2), making it easy for you to choose (for instance, grammar codes, or linguistic labels) from a set of options, rather than hunt down what you need in the Style Guide itself. The complex guidelines in the Style Guide take some time to master, as the dictionary project gets under way. The Style Guide itself doesn't remain set in stone, but evolves as further problematic issues arise during the course of the compiling, and are reported by the lexicographers.

12.1.3 Template entries

The Style Guide incorporates the 'rules' for dealing with each individual entry component. But the lexicon includes groupings of words whose members have so much in common with one another that it makes sense to follow a standard model when compiling entries for them. These standard models are what we call 'template entries' (§4.5). A template is a kind of skeleton entry which you flesh out with information from the database, and exploits two systematic aspects of language:

- Many words belong to *lexical sets*[1] on the basis of shared semantic properties.
- The members of a lexical set often pose similar problems and require very similar treatment in a dictionary.

Before the main editing begins, the whole team may be involved in the compiling of the template entries. For bilingual dictionaries the list in §4.5.3 is a good starting point, but worth expanding – there are at least sixty or seventy categories, perhaps more, for which templates are useful. What you add to that list depends in part on your target language and how the equivalences stack up across the languages between members of similar lexical sets. The 'entry structure and contents template', discussed in §4.5.1.1, is the one to use as a model in bilingual dictionaries, since our entries do not need templates for many of the lexical sets where a consistent approach to writing definitions is important (see §10.1.3).

➔ Give careful thought to your template entries: days spent on creating a comprehensive set can save literally months of editing time in the long run.

12.2 Distributing information throughout the entry

12.2.1 Multiword expressions (MWEs)

One of the most important functions of the Style Guide is to set out a coherent policy on the handling of MWEs. These complex decisions are made by the Style Guide editors: lexicographers simply follow the rules. Five types of MWE are discussed in §6.2.2; they are:

[1] A lexical set (§4.5.1) is a group of words linked to each other by a common element of meaning, e.g. *days of the week*, *birds*, *flowers*, *metals*, *precious stones*, etc.

- idioms
- collocations
- compounds
- phrasal verbs
- support verb constructions.

The choices available when it comes to recording these in a database entry, explained in §9.2.6.2, are all valid options for the bilingual dictionary entry. There is no right and wrong about how to present the various types of MWE: every dictionary has its own approach to this, as may be seen from the extracts in Figure 12.3 from two English-French dictionaries of similar size and coverage, *OHFD-3* (2001) and *CRFD-8* (2006). Compare the way

shrug /ʃrʌg/
A *n* (*also* **~ of the shoulders**) haussement *m* d'épaules; **to give a ~** hausser les épaules.
B *vtr* (*p prés etc* **-gg-**) (*also* **~ one's shoulders**) hausser les épaules *fpl*.
Phrasal verb ▪ **shrug off ▶ ~ off [sth], ~ [sth] off** ignorer *[problem, rumour]*.

OHFD-3 (2001)

shrug /ʃrʌg/ 1 n haussement *m* d'épaules ♦ **to give a ~ of contempt** hausser les épaules (en signe) de mépris ♦ **... he said with a ~ :...** dit-il en haussant les épaules *or* avec un haussement d'épaules 2 vti **to ~ (one's shoulders)** :hausser les épaules
shrug off vt sep *[+ suggestion, warning]* dédaigner, faire fi de; *[+ remark]* ignorer, ne pas relever; *[+ infection, cold]* se débarrasser de

CRFD-8 (2006)

shower /ˈʃaʊə(r)/
A *n* 1 (for washing) douche *f*; **to have** ou **take a ~** […] 2 Meteorol averse *f*; […] 3 (of confetti, sparks, fragments) pluie *f* (**of** de); […] 4 US **bridal/baby ~** […] 5 ◦ GB péj (gang) bande *f*.
B *modif [cubicle, curtain, head, rail, spray]* de douche.
C *vtr* 1 (wash) doucher […]; 2 **to ~ sth on** ou **over sb/sth** […] 3 fig **to ~ sb with sth** […].
D *vi* 1 [person] prendre une douche; 2 **petals/sparks ~ed on me** […]
shower: **~ attachment** *n* douchette *f* de lavabo; **~ cap** *n* bonnet *m* de douche; **~proof** *adj* imperméabilisé; **~ unit** *n* douche *f*; **~ room** *n (private)* salle *f* de bains (avec douche); *(public)* douches *fpl*.

OHFD-3 (2001)

shower /ˈʃaʊəʳ/ N 1 *[of rain]* averse *f (fig)* [of blows] volée *f* [...]
2 douche *f* **to have** *or* **take a ~** prendre une douche [...]
3 *(Brit ** pej = people)* bande *f* de crétins *
4 *(before wedding etc)* **to give a ~ for sb** organiser [...]
VT *(fig)* **to ~ sb with gifts** [...]
VI 1 *(= wash)* se doucher, prendre une douche
2 *(= fall)* **hailstones ~ed (down) on to the car** [...]
COMP **shower attachment** N douchette *f* à main, douchette *f* de lavabo
shower cap N bonnet *m* de douche
shower cubicle N cabine *f* de douche
shower curtain N rideau *m* de douche
shower gel N gel *m* douche
shower stall N → **shower cubicle**
shower unit N bloc-douche *m*

CRFD-8 (2006)

Fig 12.3 Layout of MWEs in different dictionaries

they set out (in the *shrug* entries) the collocations *shrug of the shoulders* and *to shrug one's shoulders*, the support verb construction *to give a shrug*, and the phrasal verb *to shrug off*. The two *shower* entries show how the presentation of compounds also differs from dictionary to dictionary.

The way the two dictionaries show MWEs is briefly as follows:

- Collocations (*to shrug one's shoulders*)
  - *OHFD* package together *to shrug* and *to shrug one's shoulders*
  - *CRFD* the same, slightly different layout
- Support verb constructions (*to give a shrug*)
  - *OHFD* include as an example
  - *CRFD* the same, but adding the useful construction with *of*
- Phrasal verbs (*shrug off*)
  - *OHFD* within the entry, but at the end, set out as secondary headword, signalled by 'Phrasal verb'; the fact that *off* is an adverbial particle is shown by the variations in wording 'to ~ off [sth], to ~ [sth] off'
  - *CRFD* similarly, within the entry, but as a secondary headword at the end
- Compounds (*shower attachment*, *shower cap*, *shower cubicle*, etc.)
  - *OHFD* The compounds are treated here in two different ways: some (perhaps those seen as posing few translation problems) are included in the *modif* (modifier) section of the entry, while the others are presented in their alphabetical order in the headword list, headed by the place-marker 'shower:'. Figure 12.4 shows how the compound list (e.g. of the *bargain* compounds) may be interrupted by main entries (e.g. *bargaining* and *bargaining chip*).
  - *CRFD* All the compounds lie within the entry, at the very end, in a special COMP (compound) section.

The discussion in §10.2.1 on variations in the wording of MWEs, exemplified by *take something with a pinch of salt*, is also relevant to the editors of bilingual dictionaries.

12.2.2 Secondary headwords

The question of whether or not the dictionary should have 'secondary headwords' (§7.2.10.1) is a decision for the policy-makers. (Run-ons, cf. §7.2.10.2, are not an option in bilinguals as every SL word needs its TL translation.) The tendency nowadays is to avoid secondary headwords if possible, as

embedding one entry (however reduced) within another simply makes it more difficult for the user to find anything. An exception to this rule is the way some dictionaries treat MWEs like phrasal verbs and compounds. This is illustrated by the entries in Figure 12.4, where full headwords in this dictionary (like *bargain* and *bargaining*) are given a phonetic transcription (indicated by the symbol /. . . / in the illustration).

bargain /…/
A *n* I (*deal*) marché *m* (**between** entre); **to make** ou **strike a ~** conclure un marché; **to keep one's side of the ~** tenir sa part du marché; **to drive a hard ~** négocier ferme *or* serré; **into the ~** par-dessus le marché; 2 (*good buy*) affaire *f*; **what a ~!** quelle bonne affaire! **to get a ~** faire une affaire; **a ~ at £10** une affaire à 10 livres sterling.
B *modif* [*buy*, *book*, *house*] à prix réduit.
C *vi* I (*for deal*) négocier (**with** avec); **to ~ for** négocier [*freedom*, *release*, *increase*]; 2 (*over price*) marchander (**with** avec); **to ~ for a lower price** marchander un prix plus bas.
Phrasal verb ▪ **bargain for, bargain on**: ▶ **~ for, ~ on something** s'attendre à quelque chose; **we got more than we ~ed for** nous ne nous attendions pas à ça.
bargain: **~ basement** *n* coin *m* des affaires; **~ hunter** *n* personne *f* à l'affût d'une bonne affaire.
bargaining /…/ A *n* (*over pay*) négociations *fpl*. B *modif* [*framework*, *machinery*, *position*, *power*, *procedure*, *rights*] de négociation.
bargaining chip *n* atout *m* dans les négociations.
bargain: **~ offer** *n* promotion *f*; **~ price** *n* prix *m* avantageux.

Fig 12.4 Secondary headwords in *OHFD-3* (2001)

This run of consecutive entries from the English-French dictionary contains several types of secondary headword:

- The phrasal verbs *bargain for*, *bargain on* are secondary headwords in a special section at the end of the main entry for *bargain.* They are 'declared' in full, in bold, but have no pronunciation, and no wordclass markers other than the introductory boxed 'Phrasal verb'. Within the secondary headword subentry, the headword *bargain* is replaced by the tilde (~).
- The entries for *bargain basement* and *bargain hunter* are grouped under a 'place marker' headword *bargain* followed by a colon. These compounds have no pronunciation but do have their wordclass shown, and the headword *bargain* here is replaced by the tilde.
- The same treatment is accorded to *bargain offer* and *bargain price*, which follow *bargaining* and *bargaining chip* in the headword list.

- The compound *bargaining chip* has an almost complete entry, but is without pronunciation. It slots into its alphabetical order in the headword list.

The material in Figure 12.4 gives an idea of how detailed the Style Guide needs to be on a comparatively simple subject such as secondary headwords, what kind of words should be treated in that way, and how they should be set out in the entry.

12.2.3 Dictionary senses

With MWEs and secondary headwords accounted for, we now turn our attention to dictionary senses. In Chapter 8, we described an approach for identifying distinct 'lexical units' (LUs) in words that exhibit polysemy (cf. in particular §8.5, §8.6.3). These are the building blocks of the database, and it is from these LUs that we will derive the inventory of senses for each of the headwords in a particular dictionary. But these LUs belong to a *monolingual* description of the language. We are now writing a *bilingual* dictionary entry, and the difference between the sense divisions in the database entry and the final dictionary entry is often dramatic: a word in the source language may have a large number of LUs with very different meanings – but they might all be translated by a single target-language equivalent. There is more about bilingual dictionary senses in §12.3.1.

12.2.4 Grammar

In the database, the grammatical properties of each LU are described in detail, individually (cf. §9.2.5). How much of this information finds its way into the final entry, and in what form, will depend on how large and complex your dictionary is, and what you think your users will need (and will be able to understand). The user profile will give you some idea about this, and it will certainly form the basis of all policy decisions on grammar. These are embodied in the Style Guide, which outlines the dictionary's general approach to grammar, lists the categories, codes, or other systems used for describing grammatical behaviour in both source and target languages, and explains the circumstances in which each of these elements may be used. Many of these categories and codes will be stored

in the dictionary writing system, thus ensuring consistency throughout the project.

Before the real entry-writing begins, it's important to get a fix on the grammar of the dictionary – the kind of grammatical information you are expected to include in the entry, and how this should be expressed. At the entry-editing stage, our task is to evaluate all of the database grammatical material in the light of its TL equivalence, and select from it the facts needed for the dictionary entry. The three major entry components used to hold grammar information in the dictionary – WORDCLASS, VALENCY, and GRAMMAR – form the focus of §7.2.6. Figure 12.5 re-visits the *question* entry from *CRFD* in order to show how these components are used to give grammar information about both the SL and the TL.

question ['kwestʃən] 1 **N** a question *f* (*also Parl*); **to ask sb a ~. to put a ~ to sb, to put down a ~ for sb** (*Parl*) poser une question à qn [...]
b (NonC = *doubt*) doute *m*; [...] **to accept/obey without ~** accepter/obéir sans poser de questions; [...]
2 **VT** a interroger, questionner **(on** sur, **about** au sujet de, a propos de) […]
b [+ *motive, account, sb's honesty]* mettre en doute *or* en question; [+ *claim*] contester; **to ~ whether…** douter que … (+ *subj*) […]

CRFD-1998

WORDCLASS of SL items only
VALENCY of SL and TL items
GRAMMAR of SL and TL items

Fig 12.5 Grammar information in SL and TL

The GRAMMAR component is very flexible (cf. for its description in §7.2.6.3, and for ways in which it is used §9.2.5.3 and §9.2.7.2) and is meant to hold a ragbag of grammatical facts: everything, indeed, that isn't wordclass or valency information. This can range from facts about the countability of SL nouns to the gender of TL nouns or the need for the subjunctive in certain TL constructions.

The way the dictionary expresses grammatical information – e.g. *n* for noun, *v* for verb, etc. – is of course a matter for the Style Guide, which also helps you to identify items to be so labelled. One additional point however should be noted, since it impacts much more on the length of a bilingual entry than it does on a monolingual one. Many dictionaries

eat /i:t/
A *vtr* (*prét* **ate;** *pp* **eaten)** I (consume) [*person, animal*] manger [*cake, food, snack*]; prendre [*meal*]; **I don't ~ meat** je ne mange pas de viande; […]
B *vi* (*prét* **ate;** *pp* **eaten)** I (take food) manger**; to ~ from** *ou* **out of** manger dans [*plate, bowl*]; 2 (have a meal) manger**; I never ~ in the canteen** je ne mange jamais à la cantine; […]

OHFD-3 (2001)

Fig 12.6 Part of a bilingual entry for *eat*, duplications highlighted

set out the intransitive uses of a verb headword quite separately from the transitive uses. This is particularly true of bilingual dictionaries and leads to a lot of duplication, as may be seen from the highlighted parts of the entry in Figure 12.6. The transitive sense of *eat* in A I and its TL equivalent (*manger*) are simply repeated in the two intransitive senses in B. The uses shown in B I and B 2 are instances of *indefinite null instantiation* (INI: see §9.2.5.5), and this phenomenon is found in so many classes of verbs that it is worthwhile considering instances of INI as part of the transitive spectrum. Cases of *definite null instantiation* are local to individual lemmas, not lexical sets, and require more explanation in a bilingual entry.

12.2.5 Labels

The vocabulary types represented by the linguistic labels are introduced in §6.4.1.4; the way the labels function is explained in §7.2.8; and their use in the database is set out in §9.2.9. However, conventional labels are at best a blunt instrument: categories like 'formal' and 'literary' are umbrella terms that conceal a good deal of variation. For instance, the word *purchase* has a more formal ring than *buy*, and would sound pompous if used in ordinary conversation. But the data suggests that in certain situations (for example when talking about buying 'major' items like land, companies, or military hardware) it is a perfectly natural word to choose. A 'formal' or 'commerce' label may not be much help here. The corpus can help us to a degree, but in general, labelling is an area of lexicography where there is more work to be done.

But for the moment we must do the best we can with what we've got. As was the case for grammar, the sociolinguistic properties of each LU are

described in detail in the database, and as necessary in the dictionary. Before entry-writing proper begins, it's important to have a good understanding of where the boundaries are drawn between 'database-only' labels and those that will appear in the dictionary itself. In the dictionary, both SL and TL items must be scrutinized to ensure that labels are used systematically and correctly (i.e. according to the Style Guide). The actual labels for use in any particular dictionary will be selected at the start of the project, listed in the Style Guide and available in the menus of the dictionary writing software. If the dictionary is to be sold in both SL and TL communities, decisions will be made with an eye to labelling practice in dictionaries of the target language. It's particularly important to get the scope of the labels right, on both SL and TL items (§7.2.8.10), and the Style Guide must carry detailed instructions on how to do this.

bird [bɜ:d] 1 N a oiseau *m*; (*Culin*) volaille *f;* **they shot six ~ s** ils ont abattu six oiseaux *or* six pièces de gibier (à plumes); **~ of ill omen** (*liter*) oiseau *m* de mauvais augure *or* de malheur; **a ~ in the hand is worth two in the bush** (*PROV*) :un tiens vaut mieux que deux tu l'auras (*PROV*); **~ s of a feather flock together** (*PROV*) qui se ressemble s'assemble (*PROV*); **they're ~ s of a feather** (*gen*) ils se ressemblent beaucoup; (*pej*) ils sont à mettre dans le même sac; **a little ~ told me*** mon petit doigt me l'a dit; **the ~ has flown** (*fig*) l'oiseau s'est envolé; **to give sb the ~** †* (*Theatre, Sport*) huer *or* siffler qn; (= *send sb packing*) envoyer bouler* *or* paître* qn; **to get the ~** †* (Theatre) se faire siffler *or* huer; **for the birds**** (= *worthless*) nul*; (= *silly*) débile*; **he'll have to be told about the ~ s and the bees** (*hum*) il va falloir lui expliquer que les bébés ne naissent pas dans les choux; → **early, jailbird, kill.**

CRFD-8 (2006)

Fig 12.7 Labels in the bilingual entry

The policy in database building, and in translating the database, is that everything in the source language that can be labelled should be labelled, and labels should be used whenever appropriate in dealing with database translations. This is not however the case for the dictionary entry: neither SL nor TL speakers need to be told, for instance, that the noun *violin* when translated into French by *violon* belongs to the domain of music. Standard practice for print dictionaries[2] is to include a label only when the item – whether SL or TL – needs clarifying. There are several situations in

[2] Labels can be more widely used in electronic dictionaries, where they could be switched off if not wanted.

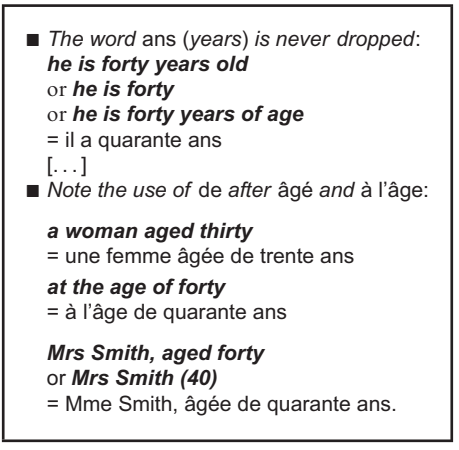
which this arises in a bilingual entry and most of these are illustrated in Figure 12.7. They are:

(1) The label functions as a sense indicator for the benefit of the SL speaker where the headword or phrase is polysemous, e.g. (*Culin*) and (*Theatre*, *Sport*).
(2) It functions as a sense indicator for the benefit of the TL speaker when the direct translation is polysemous, e.g. the asterisk (meaning 'informal') on the translation *débile**, which is supposed to tell the French speaker that this word has the sense of 'stupid', not that of 'sickly'. (However, this is not the most user-friendly way to indicate meanings, see below at §12.3.4.)
(3) It warns both types of user when an item does not belong to the default 'unmarked' general language, e.g. (*liter*), all the (*Prov*)s, the asterisks (which are this dictionary's way of indicating the expression's position on the register scale of informality), and the † symbol (meaning 'old-fashioned').
(4) It indicates a non-literal interpretation of the item it attaches to, thus reassuring users that the translation can be used in the same way, e.g. (*fig*) and (*hum*).

➔ Don't rely too much on labels in your entry: they usually mean more to you than they do to the user.

12.2.6 Usage notes

The type of note discussed here (the 'subject-oriented' usage note) is described in §7.2.9.1, where it is illustrated with an excerpt from the *OHFD-2001* note on *Countries and Continents.* As for template entries, the subjects to be covered in any bilingual dictionary will depend on how the source and target languages match up, or diverge. The *OHFD*, for instance, lists over forty topics for which quite long and complex usage notes are supplied for the English user, in English, about translations into French. These include such diverse topics as Age, Capacity Measurement, The Clock, Currencies and Money, and Date. As well as offering translations for the commonest phrases such as (for 'Age') *how old are you?*, *what age is she?*, *he's about fifty*, and so on, the notes also contain other information on usage and grammar for which there is no obvious location in a dictionary. This type of note is


■ *The word* ans (*years*) *is never dropped*:
he is forty years old
or ***he is forty***
or ***he is forty years of age***
= il a quarante ans
[. . .]
■ *Note the use of* de *after* âgé *and* à l'âge:

a woman aged thirty
= une femme âgée de trente ans
at the age of forty
= à l'âge de quarante ans

Mrs Smith, aged forty
or ***Mrs Smith (40)***
= Mme Smith, âgée de quarante ans.

Fig 12.8 From the usage note on 'Age' in *OHFD-3* (2001)

illustrated in the excerpt from the *OHFD-3* (2001) 'Age' usage note shown in Figure 12.8.

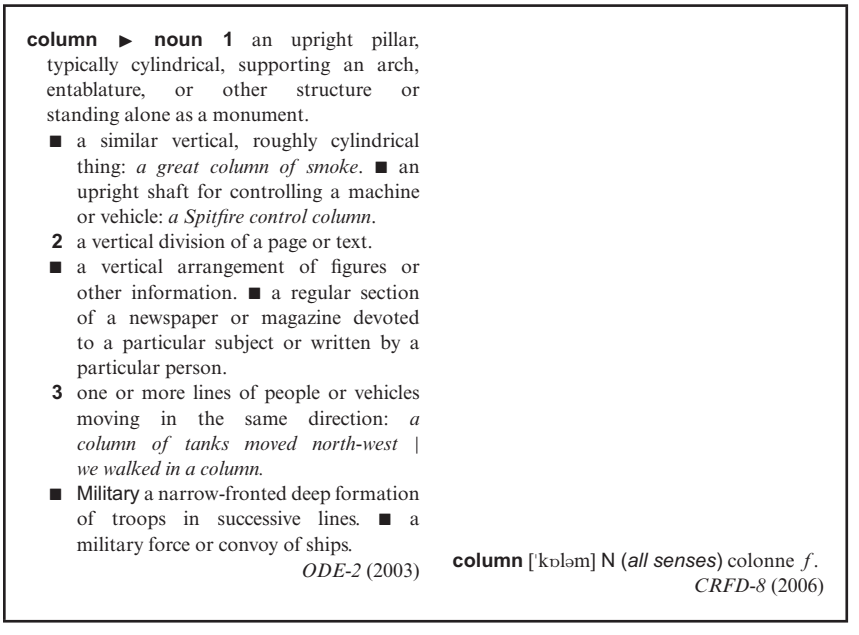
➔ Avoid duplication by preparing at least a working draft of such usage notes early in the project, so that the team don't fill entries with material that proves redundant in the light of the notes.

12.3 Writing the entry

In this section we look at the major types of activity involved in putting together a bilingual entry on the basis of a thorough analysis of corpus data, with translations inserted where appropriate.

12.3.1 Deciding on senses

When writing a dictionary entry, you start by setting up provisionally the skeleton of the entry, the dictionary senses. The first thing to note is that these 'dictionary senses' are not the same as lexical units (LUs), the basic building blocks of the monolingual database entry. For bilingual entry writers, LUs are the 'deep structure' of the entry. They plot out the essential senses of a polysemous headword, and let you see the full potential of your headword in the language. They have no place in the 'surface structure' of the bilingual entry, where their role in ordering the material is taken by what we'll call 'dictionary senses' (or 'senses' for short). The reason for this should be clear from a comparison of the two entries shown in Figure 12.9.



column ▶ **noun 1** an upright pillar, typically cylindrical, supporting an arch, entablature, or other structure or standing alone as a monument.
- a similar vertical, roughly cylindrical thing: *a great column of smoke*. ■ an upright shaft for controlling a machine or vehicle: *a Spitfire control column*.

2 a vertical division of a page or text.
- a vertical arrangement of figures or other information. ■ a regular section of a newspaper or magazine devoted to a particular subject or written by a particular person.

3 one or more lines of people or vehicles moving in the same direction: *a column of tanks moved north-west* | *we walked in a column.*
- Military a narrow-fronted deep formation of troops in successive lines. ■ a military force or convoy of ships.

ODE-2 (2003)

column ['kɒləm] N (*all senses*) colonne *f*.

CRFD-8 (2006)

Fig 12.9 Monolingual and bilingual 'senses' of the same word

The two dictionaries from which the entries in Figure 12.9 are taken are both large one-volume standard works, destined for the adult market. But that's where the resemblance ends. The first (*ODE*) is a monolingual dictionary, the second (*CRFD*) a bilingual. The dictionary senses of the first (three main senses, each with two subsenses, or nine in all) correspond to the LUs of the lemma *column*. The French equivalent of this word in each of its LUs is *colonne*. Space is too precious to repeat this fact nine times, and users would find searching such an entry time-consuming and irritating. As a result, 'dictionary senses' in a bilingual are not really senses of the headword at all, but simply the most user-friendly way to structure the material. Bilingual dictionary senses[3] are predicated more on the TL than on the actual meanings of the SL headword. While it's not always possible to compress into one dictionary sense all the uses with the same TL equivalent, it's quite acceptable to do so if you can, since the semantic content of the LUs involved is often fairly similar.

The ordering of dictionary senses (discussed in §7.3.3) is again normally subordinated to the needs of the typical users. A dictionary made for use only by SL speakers can use their knowledge of the various meanings of

[3] This use of the word *sense* is common among lexicographers.

the headword. However, in a dictionary made for use by both SL and TL speakers, only the SL speakers can rely on the sense of the headword to help them navigate the entry. By definition, that's probably what the TL speakers don't know (otherwise, why look up the word?). On the other hand, what they may be able to identify is the wordclass of the unknown word, i.e. the headword in their own context. Consequently, the usual way of structuring the material in a bilingual entry is to treat it as follows:

(1) Separate out the uses on the basis of wordclass, with nouns, verbs, adjectives, and adverbs etc. in separate sections, and keep them apart (but see the note on INI constructions and transitivity in §12.2.4 above).
(2) Look at the TL equivalents of each of these uses.
(3) If it is reasonable (i.e. if you can read through the resultant entry without an unpleasant shock of surprise) see if you can collapse some of the uses together into dictionary senses.

Once you have got this far, the Style Guide should tell you whether to present the material in a hierarchy of senses and subsenses, or in a simple flat structure (see §7.3.2). The hierarchical approach is more satisfying for you, as an SL speaker, because it chimes with your perception of (for instance) sense 1c being 'closer' to sense 1a and 1b than it is to sense 2. Whether this helps TL-speaking users is not clear, but it probably encourages SL speakers in their search through the entry. At this point, too, you should consult the Style Guide for instructions on how to deal with the MWEs in your entry, and plot them in to your draft entry.

12.3.2 Offering translations

Once you have a clear idea of the senses your entry will include, and where the various types of MWE should be located, then you work on each sense individually until the entry is complete. (Of course, it's never as clear as that – sometimes you have to tweak another sense as you go along.) However, from now on, when we say 'headword' we mean 'headword in the particular sense you're dealing with'. In bilingual entries, your objective is to give users a clear idea of the safest direct translation of the headword, of where the boundaries of that translation lie, and of other TL expressions that could come in handy in translating the headword or expressing the

design [dɪ'zaɪn] 1 N a (= *ornamental pattern*) motif *m*, dessin *m* (**on** sur); **the ~ on the material/ the cups** le dessin *or* le motif du tissu/des tasses; **a leaf ~** un motif de feuille(s).

b (= *plan drawn in detail*) (*of building, machine, car etc*) plan *m*, dessin *m* (**of, for** de); (*of dress, hat*) croquis *m*, dessin *m* (**of, for** de) (= *preliminary sketch*) ébauche *f*, étude *f* (**for** de); **have you seen the ~s for the new cathedral?** avez-vous vu les plans de la nouvelle cathédrale?

c (= *way in which sth is planned and made*) (*of building, book*) plan *m*, conception *f* (**of** de); (*of clothes*) style *m*, ligne *f* (**of** de); (*of car, machine etc*) conception *f* ; (= *look*) esthétique *f*, design *m*; **the ~ was faulty** la conception était défectueuse, c'était mal conçu; **the ~ of the apartment facilitates ...** le plan de l'appartement facilite ...; **the general ~ of "Paradise Lost"** le plan général *or* l'architecture *f* du "Paradis perdu"; **a dress in this summer's latest ~** une robe dans le style de cet été; **the ~ of the car allows ...** la conception de la voiture *or* la façon dont la voiture est conçue permet ...; **the grand** *or* **overall ~** le plan d'ensemble; **this is a very practical ~** c'est conçu de façon très pratique; **these shoes are not of (a) very practical ~** ces chaussures ne sont pas très pratiques.

d (= *completed model*) modèle *m*; **a new ~ of car** un nouveau modèle de voiture; **the dress is an exclusive ~ by ...** cette robe est un modèle exclusif de ...

e (*subject of study*) (*for furniture, housing*) design *m*; (*for clothing*) stylisme *m*; **industrial ~** l'esthétique *f* *or* la création industrielle; **he has a flair for ~** il est doué pour le design.

f (= *intention*) intention *f*, dessein *m*; **his ~s became obvious when ...** ses intentions *or* ses desseins sont devenu(e)s manifestes quand ...; **to conceive a ~ to do sth** former le projet *or* le dessein de faire qch; **imperialist ~s against their country** les visées impérialistes sur leur pays; **by ~** (=*deliberately*) délibérément, à dessein; **whether by ~ or accident he arrived just at the right moment** que ce soit à dessein *or* délibérément ou par hasard, il est arrivé juste au bon moment; **truly important events often occur not by ~ but by accident** les événements vraiment importants sont souvent le fruit du hasard plutôt que d'une volonté précise; **to have ~s on sb/sth** avoir des visées sur qn/qch; **to have evil ~s on sb/sth** nourrir de noirs desseins à l'encontre de qn/qch; **we believe they have aggressive ~s on our country** nous pensons qu'ils ont l'intention d'attaquer notre pays.

2 VT a (= *think out*); [+ *object, car, model, building*] concevoir; [+ *scheme*] élaborer; **well-~ed** bien conçu

b (= *draw on paper*) [+ *object, building*] concevoir, dessiner; [+ *dress, hat*] créer, dessiner.

c (= *destined for a particular purpose*) **room ~ed as a study** pièce conçue comme cabinet de travail; **car seats ~ed for maximum safety** des sièges *mpl* de voiture conçus pour une sécurité maximale; **software ~ed for use with a PC** un logiciel conçu pour être utilisé sur un PC; **to be ~ed for sb** (= *aimed at particular person*) s'adresser à qn; **a course ~ed for foreign students** un cours s'adressant aux étudiants étrangers ; **to be ~ed to do** (= *be made for sth*) être fait or conçu pour faire; (= *be aimed at sth*) être destiné à faire, viser à faire; **~ed to hold wine** fait *or* conçu pour contenir du vin; **a peace plan ~ed to end the civil war** un plan de paix visant or destiné à mettre fin à la guerre civile; **the legislation is ~ed as a consumer protection measure** cette loi vise à protéger les consommateurs; **clothes that are ~ed to appeal to young people** des vêtements qui sont conçus pour plaire aux jeunes.

3 COMP ▶ **design award** N prix *m* de la meilleure conception *or* du meilleur dessin ▶ **design engineer** N ingénieur *m* concepteur ▶ **design fault** N défaut *m* de conception ▶ **design office** N (*Ind*) bureau *m* d'études

CRFD-5 (1998)

Fig 12.10 The *CRFD* entry for *design*

concept underlying it. Most of the components that you can use to transmit this information are discussed in §7.2.4. We shall now look at these components in turn.

The complete *CRFD* entry for *design*[4] is shown in Figure 12.10 and will serve as a comprehensive illustration of most of the points to be made in the rest of this chapter about putting translations into a bilingual entry.

12.3.2.1 *Direct translation* The major translation-carrying component is of course the DIRECT TRANSLATION (§7.2.4.1). At this point you have in front of you a lot of different sentences and half-sentences using the headword in this one sense, and your first objective is to find a TL word that fits as many of these contexts as possible – a translation that is as near *context-free* as you can make it (see §11.1 for an explanation of this term).

design [dɪ'zaɪn] 1 N a (= *ornamental pattern*) motif *m*, dessin *m* (**on** sur); **the ~ on the material/ the cups** le dessin *or* le motif du tissu/des tasses; **a leaf ~** un motif de feuille(s).

Fig 12.11 Direct translations

The first sense of *design* (Figure 12.11) offers two apparently equally good direct translations (although not synonyms), *motif* and *dessin*, together with the genders of these nouns and the translation of the construction ('**on** sur') needed to slot them into the most common sentence pattern in which the word is found; a sample of corpus contexts is given in Figure 12.12.

the diamond	design	on the traditional Aran sweater
the pattern echoed by the	design	on the vase
a vessel with floral	design	on the exterior
a carrier bag that's got the	design	on the front
the rest of the	design	on the front of the coin
the	design	on the hand of one of the men
a spoon with a flowerlike	design	on the handle

Fig 12.12 Sample concordances for *design on*

Normally, when you offer two direct translations that are not totally synonymous (and what two words ever are?) it's good practice to include

[4] Experience showed this to be the most translation-resistant word of all 50,000-odd headwords in an English-French dictionary.

design [dɪ'zaɪn] 1 N a […]
b (= *plan drawn in detail*) (*of building, machine, car etc*) plan *m*, dessin *m* (**of, for** de); (*of dress, hat*) croquis *m*, dessin *m* (**of, for** de) (= *preliminary sketch*) ébauche *f*, étude *f* (**for** de); […]
c (= *way in which sth is planned and made*) (*of building, book*) plan *m*, conception *f* (**of** de); (*of clothes*) style *m*, ligne *f* (**of** de); (*of car, machine etc*) conception *f*; (= *look*) esthétique *f*, design *m*; […]
d […]
e (*subject of study*) (*for furniture, housing*) design *m*; (*for clothing*) stylisme *m*; […]
2 VT a (= *think out*); [+ *object, car, model, building*] concevoir; [+ *scheme*] élaborer; […]
b (= *draw on paper*) [+ *object, building*] concevoir, dessiner; [+ *dress, hat*] créer, dessiner.[…]

Fig 12.13 Sense indicators

sense indicators (cf. §7.2.5) to clarify the differences in their use. This is well done in the rest of the *design* entry: senses **1b**, **1c**, **1e**, **2a**, and **2b** (Figure 12.13) all offer a selection of direct equivalents and a wealth of sense indicators to help the SL speakers choose the most appropriate for their needs. They will also find the constructions they'll need if they are to use the word correctly, in the form of '(**of**, **for** de)' and so on throughout the noun senses.

Many SL–TL pairs will be the source of recurring problems for direct translations: difficulties that depend on a systematic mismatch of semantic distribution between the two languages. Policy decisions on how to handle these have to be taken ahead of the entry-compiling and embodied in the Style Guide. Since actual instances have no interest except for specific language pairs, one example will suffice. It is a problem that arises in dictionaries from English into many Romance languages, and is illustrated in Figure 12.14.

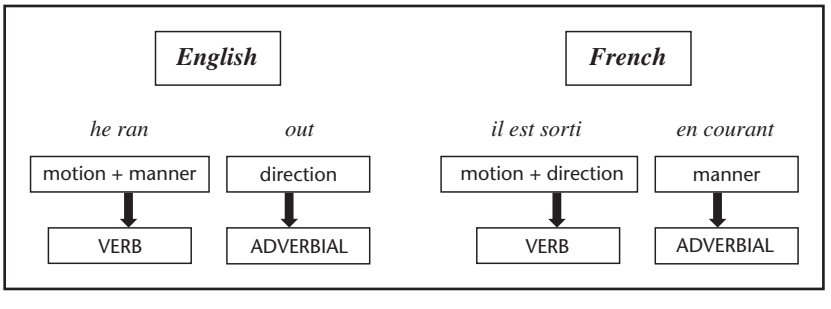


Fig 12.14 Regular equivalence pattern in English and French motion verbs

English manner of motion verbs with directional particles are normally translated by French directional motion verbs with adverbials of manner. The slash (/) should be used to show the productivity of this patterning, as follows:

to run in / out *etc*. entrer / sortir *etc*. en courant.

Note that the slash is used elsewhere in the dictionary when alternatives in the SL item are paralleled in its translation, and when it is used it *must* appear in both the SL and TL items.

Fig 12.15 Style Guide on a recurrent SL-TL pattern

The French equivalent of *he ran out* is *il est sorti en courant* (literally, 'he went out running'). The Style Guide dealing with this would have to contain a paragraph like the one in Figure 12.15. This would result in entries like those in Figure 12.16, where the productive nature of this equivalence is shown by the SL and TL material to the left and right of the slashes. This implies (and hopefully users infer) that the formula may be used for all similar instances, so that they will translate *to amble back* (or *hobble back* etc.) as *revenir d'un pas tranquille* (or *revenir en clopinant* etc.).

amble ['æmbl] aller d'un pas tranquille; **to amble in/out** *etc* entrer/sortir *etc* d'un pas tranquille.	**hobble** ['hɒbl] clopiner; **to hobble in/out** *etc* entrer/sortir *etc* en clopinant
crawl [krɔ:l] *vi* ramper; **to crawl in/out** *etc* entrer/sortir *etc* en rampant.	**stagger** ['stægəʳ] chanceler, tituber; **to stagger in/out** *etc* entrer/sortir *etc* en chancelant *or* titubant.

Fig 12.16 Treatment of some English manner of motion verbs

12.3.2.2 *Other translation components* There are of course cases where no direct translation exists, and you have to resort to one of two alternatives: the NEAR-EQUIVALENT (§7.2.4.2) and the GLOSS (§7.2.4.3), shown here again in Figure 12.17. Here again, the Style Guide must give clear guidance on what is acceptable in these components.

It is also possible to supplement the direct translation by using USAGE NOTES. All members of the editing team are expected to be aware of the SUBJECT-ORIENTED USAGE NOTES (see §7.2.9.1 and §12.2.6) planned for the dictionary, and to include cross-references to these as appropriate. LOCAL USAGE NOTES (see §7.2.9.2) are normally written by the lexicographer compiling the associated entry. Less central to our purpose here is

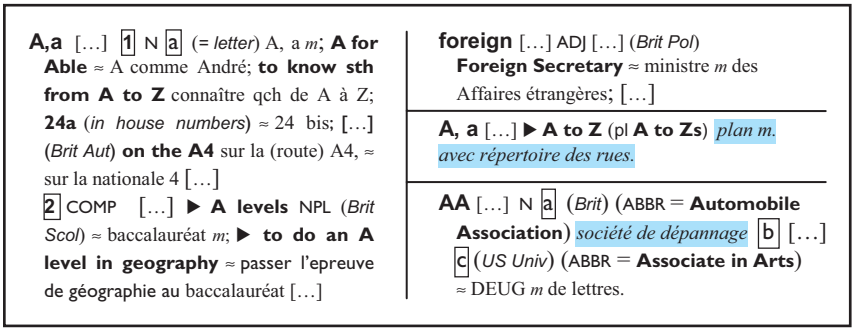


Fig 12.17 NEAR-EQUIVALENTS (≈) and GLOSSES in *CRFD-5* (1998)

the non-lexical material that may also be used to convey TL equivalents. This includes in-text graphic illustrations and photographs, and – in the dictionary back matter – extra-textual materials like tables, organigrams, and charts.

However, by far the most useful way to handle cases where there is simply no direct translation of the headword is to move straight into examples with their translations, set out according to a typographical protocol which indicates the absence of a direct translation. This tactic is explained in §7.2.4.4 and illustrated again here in Figure 12.18.

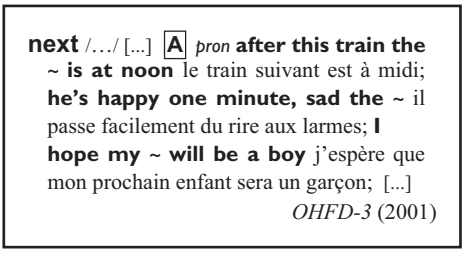


Fig 12.18 Translated examples take the place of a direct translation

12.3.3 Choosing examples[5]

Examples in an 'active' bilingual dictionary (designed for encoding SL speakers) supplement the information given in the direct translation(s). Their purpose is to help SL speakers choose the appropriate TL equivalent and use it correctly. This involves:

[5] Everything we say in this section refers to dictionary examples *plus their translations*. Although some 'passive' dictionaries designed specifically for decoding TL speakers contain SL examples without translations, examples in an 'active' entry for use by encoding SL speakers must be translated.

- showing them which sense of the headword is being translated
- reassuring them about the use of the direct translation
- complementing (or replacing) the direct translation: offering translations for when it can't be used, or simply for when the user is looking for a different wording
- pinpointing the meaning of polysemous TL words.

(The second of these is the least important.) This means that the great bulk of examples in a bilingual entry are chosen entirely on the basis of their translations. There is no room – literally as well as figuratively – for discussions on whether we should use 'real' examples direct from corpus. We don't have the luxury of such a choice. If the dictionary is not to be too long, or too confusing, we have to offer examples that shed light on the uses of the unfamiliar target language. Here collocates (§9.2.7) come into play – the entry should help users translate the headword in the context of its principal corpus collocates. This means composing simple examples using these collocates (based on facts drawn from the corpus), in order to show how the headword is translated in these very frequent contexts. It is naïve to think that you can lift stretches of corpus text, translate them, and produce really useful examples. Crafting a bilingual entry is more complex than that, and demands more skills: as well as a native-speaker command of the source language, editors must have an excellent knowledge of the target language.

The 'context-free' versus 'context-sensitive' distinction (§11.1) holds good for translations of examples as well as for direct translations. Take the case of the English construction *could always*. The BNC has 412 '*could always*' sentences, of which a few are shown in Figure 12.19.

I	could always	ask Canon Wheeler to hear my confession
I'd rather miss the news . . . I	could always	watch it at nine.
I	could always	work from home
If she took the bandage off she	could always	wrap it up again
you	could always	try using some fresh extra thick cream
If you are feeling brave you	could always	try phoning.
it	could always	hope to attract supporters from allied groups
you	could always	join her later
If they're the worse for wear you	could always	just paint and perhaps stencil them
he	could always	claim he knew nothing of the layout

Fig 12.19 Some concordances for *could always*

Recently, a bilingual entry, one of a set of sample entries, included this construction in the most neutral context possible, in the form of an example:

you could always do something else. This was translated as *on pourrait toujours faire autre chose.* A reviewer criticized this as 'much too literal', deploring the absence of the picturesque idiom *avoir plus d'une corde à son arc* (literally 'to have more than one string to one's bow'), ignoring the fact that the 'vanilla' example gives the language-learner a construction that can be adapted to fit many contexts. The almost context-free *on pourrait toujours faire autre chose* would be good for most of the 412 contexts found in the BNC. The excessively context-sensitive *avoir plus d'une corde à son arc* would simply be a distraction, and lead all but the most skilful user into error.

➔ Go for context-free if you can: always for the direct translation and for the examples too as far as possible.

When you reach the point of choosing examples, you've already chosen one or two direct translations, and are working from either your own notes or a translated database entry showing how the headword is used and how it is translated in a variety of different contexts. Now the task is to choose examples that complement the direct translations. It's important to bear in mind what SL speakers and TL speakers will infer when they read these examples. When you've finished an entry, it's worthwhile putting it aside for a day or two, and returning to it when some of the details are not so fresh in your mind. Then you read it through as though you were a typical user – first as an SL speaker then (if it's a dictionary meant for both language communities) as a TL speaker. How will each interpret the various examples? Will they assume anything about the foreign language on the basis of these multiple contexts? It's easy to lead users into error through ignoring that aspect.

The *design* entry in Figure 12.10 illustrates very well the fourfold role that examples play in a bilingual entry. For ease of reference, the relevant parts of this entry are repeated in Figure 12.20.

(1) Indicating the sense of the headword
- **1d**: the sense indicator 'completed model' is not easy to understand, but the two examples differentiate this sense from **1b** and **1c**, pointing the user to *modèle* rather than *plan* or *dessin* as the correct translation.
- **1b** and **1c** are difficult to distinguish, but the first example in each of these clarifies the distinction: *have you seen the designs for the new cathedral?* in **1b**, and *the design was faulty* in **1c**.

design [dɪ'zaɪn] **1** N a (= *ornamental pattern*) motif *m*, dessin *m* (**on** sur); **the ~ on the material/ the cups** le dessin *or* le motif du tissu/des tasses; **a leaf ~** un motif de feuille(s) b (= *plan drawn in detail*) (*of building, machine, car etc*) plan *m*, dessin *m* (**of, for** de); (*of dress, hat*) croquis *m*, dessin *m* (**of, for** de) (= *preliminary sketch*) ébauche *f*, étude *f* (**for** de); **have you seen the ~s for the new cathedral?** avez-vous vu les plans de la nouvelle cathédrale?
c (= *way in which sth is planned and made*) (*of building, book*) plan *m*, conception *f* (**of** de); (*of clothes*) style *m*, ligne *f* (**of** de); (*of car, machine etc*) conception *f* ; (= *look*) esthétique *f*, design *m*; **the ~ was faulty** la conception était défectueuse, c'était mal conçu; [...] **the ~ of the car allows ...** la conception de la voiture *or* la façon dont la voiture est conçue permet ...; [...] **this is a very practical ~** c'est conçu de façon très pratique;
d (= *completed model*) modèle *m*; **a new ~ of car** un nouveau modèle de voiture; **the dress is an exclusive ~ by ...** cette robe est un modèle exclusif de ...
e (*subject of study*) (*for furniture, housing*) design *m*; (*for clothing*) stylisme *m*; **industrial ~** l'esthétique *f or* la création industrielle; **he has a flair for ~** il est doué pour le design.
f (= *intention*) intention f, dessein m; **his ~s became obvious when ...** ses intentions *or* ses desseins sont devenu(e)s manifestes quand ...; […] **by ~** (=*deliberately*) délibérément, à dessein; **[...] to have ~s on sb/sth** avoir des visées sur qn/qch;
2 VT a [...]
c (= *destined for a particular purpose*) **room ~ed as a study** pièce conçue comme cabinet de travail; [...] **to be ~ed to do** (= *be made for sth*) être fait *or* conçu pour faire; (= *be aimed at sth*) être destiné à faire, viser à faire; **[...] a peace plan ~ed to end the civil war** un plan de paix visant or destiné à mettre fin à la guerre civile; **the legislation is ~ed as a consumer protection measure** cette loi vise à protéger les consommateurs; [...]

Fig 12.20 Examples (highlighted) pulling their weight in a bilingual entry

- **1e**: the two examples where the headword is used without the article indicate the sense at once.
- **1f**: here again, the first example tells you the sense of the headword, as do others like *to have designs on sb/sth*.

(2) Confirming the direct translation

- **1a**: after giving *motif* and *dessin* as direct translations, the example *the design on the material / the cups* uses both these TL words.
- **1d**: *modèle* is the direct translation, confirmed by both examples.
- **1f**: the direct translations *intention* and *dessein* are offered in the translation of the first example *his designs became obvious when* . . .

(3) Complementing (or replacing) the direct translation

This is the principal *raison d'être* of the example in the bilingual dictionary, and the **design** entry is full of examples included for that purpose. A selection follows.

- **1c**: the entry makes it clear that the verb *concevoir* is a must for translating this sense of the noun, but it won't fit as a direct translation, so we find the past participle *conçu* in the translations of three of the eight examples: *the design was faulty*, *the design of the car allows* . . . and *this is a very practical design*. In particular, *concevoir*

is useful to translate *design* when modified by *faulty*, a significant collocate.

- **1e**: the direct translations *design* and *stylisme* are not enough, since *design* when modified by a pertainym like *industrial* is normally translated by *esthétique* or *création*, as the example shows.
- **1f**: the entry contains two idiomatic collocations where the direct translation cannot simply be plugged in to a word-by-word translation – *by design*, *to have designs on* – and this is clearly set out in the examples.
- **2c**: here the examples replace a direct translation. The passive of this sense (*be designed to do*...) is a very significant pattern in the corpus behaviour of this verb, and this is reflected in the fact that all the examples are passive forms; not all of them however accept *concevoir* as a translation, and so we see examples like *to be designed to do* using not only *conçu pour faire*, but also *fait pour faire*, *destiné à faire*, and *visé à faire*. Further down two examples show the active use of *viser* translating the English passive: *a peace plan designed to end the civil war* and *the legislation is designed as a consumer protection measure*.

(4) Pinpointing meanings of polysemous words

Occasionally, a headword is translated perfectly correctly by one sense of a polysemous TL word. Such a case is the English noun *story*, whose French equivalent, *histoire*, means both 'story, tale' and 'history'. In the entry for *histoire* the two senses will be distinguished, but in the entries for both *story* and *history*, the appropriate sense of the TL language item must be indicated for the sake of TL users (and such clarifications must also be clearly indicated when translations are put into the database). Figure 12.21 illustrates this point with extracts from *OHFD-3* (2001).

story /…/ *n* I (*account*) histoire *f* (**of** de); **to tell a ~** raconter une histoire; […]

history /…/ A *n* I (*past*) histoire *f*; **ancient/modern ~** histoire *f* ancienne/moderne; […]

histoire /…/ *nf* I (*discipline*) history; **aimer/enseigner/étudier l'~** to like/teach/study history; […] 2 (*récit*) story; **raconter une ~ de fantômes à quelqu'un** to tell somebody a ghost story; […]

Fig 12.21 Examples pinpointing meanings of polysemous words

The two meanings of *histoire* present no problem to English speakers.

- If they are using the English-French entry to translate either *story* or *history* into French, the fact that *histoire* has a second meaning is irrelevant.
- If they are using the French-English entry to translate *histoire* into English, they won't have any problem in choosing between *story* and *history*.

It is the French speakers who need help in this situation.

- If they are using the English-French entry to translate either *story* or *history* into French, they need to know which sense of *histoire* the English word means, and this is clarified for them by the French of the two distinctive examples, 'raconter une histoire' and 'histoire ancienne/moderne'.
- If they are using the French-English entry to translate *histoire* into English, they need to choose correctly for their context between *story* and *history*, and this is clarified for them by the examples again, 'aimer/enseigner/étudier l'histoire' and 'raconter une histoire de fantômes'.

12.3.4 Indicating meanings

An entry as complicated and difficult as the *design* entry in Figure 12.10 would be virtually unusable without careful indication of the meanings of the word being handled in the various senses. The function of sense indicators (see §7.2.5 for an introduction to these) is self-evident, but it's worth picking out a few from the *design* entry, where they certainly pull their weight. The three main types – specifiers, collocators, and domain labels – are shaded in Figure 12.22.

Specifiers (§7.2.5.2)

- All senses: such is the difficulty of this entry that a specifier introduces every one of the senses, always in the form of a paraphrase or synonym, for instance for **1a** this is *ornamental pattern*, for **1b** it is *plan drawn in detail*, for **1c** *way in which sth is planned and made*, and so on.

Collocators (§7.2.5.3)

- Most senses also include collocators:

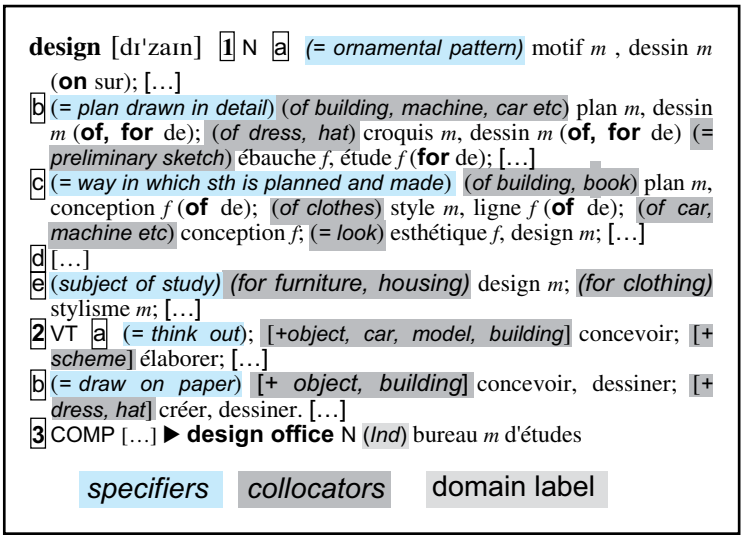


Fig 12.22 Various ways to indicate meanings

- noun collocators for the noun senses, for instance **1b**: *of building, machine, car* and *of dress*, *hat*; **1c** *of building, book* and so on; and **1e** *for furniture, housing* and *for clothing*
- noun collocators as typical objects for the verb senses, for instance **2a**: *object, car, model, building* and *scheme;* and **2b**: *object, building* and *dress, hat*.

Domain labels (§7.2.5.1)

- Sense 3 ('Ind' for 'industry' at the compound *design office*), presumably to help the SL speaker understand what kind of a 'design office' this is. Although domain labels are commonly used in bilingual dictionaries as sense indicators, they are hardly user-friendly, and this is the only domain label in the entry. The editor of this very complex entry thought it better to use more explicit sense indicators.

➔ When you're including sense indicators, remember they're mainly for the SL speaker: the TL speaker rarely needs them.

12.3.5 Completing the entry

Finally, you've reached the end of the entry. All the senses of the headword have been teased out and translations inserted. You've checked that all these

translations carry with them information about when they would be chosen and how they are used. Perhaps you've added a few cross-references to other entries where your headword may be found. All that it remains for you to do now is to re-read the whole entry: the best time to do that is after a week or two. You want to be sure you can still make sense of it when you come afresh to it. And you want to double-check that you're not inadvertently leading any of your users into error. Finally, you check it for length. If it's too long (and it always is) the painful process of cutting it down is always easier to do at a distance from the actual compiling.

Exercises

Exercise 1

This exercise builds on the translated database entries you created in the exercises attached to Chapter 11. In each case the objective is to create a dictionary entry from your translated database material.

- Write a brief user profile for your dictionary (see §2.3).
- Using §12.3 as a guide, create a dictionary entry for the verb.
- Do the same for the noun.
- Do the same for the adjective.
- Do the same for the adverb.

Exercise 2 A new electronic dictionary

- Choose a large bilingual dictionary that you know well.
- You are planning to convert this dictionary into a ground-breaking new electronic dictionary. For once, there are no budget constraints. Propose two new features:
  - an improvement: something that the dictionary does now that will be done better in your model
  - an innovation: an entirely new function the e-dictionary will have, one that will be of real benefit to its users.

(Do not include the speed of searches as an improvement! That is a given.)

- Any new feature you propose should benefit the SL-speaking users in one or more of the following respects:
  - finding their way about the entry
  - information about the headword

  - help with grammar
  - help with finding the right translations
  - help with using the target language correctly
  - help with the messages carried in the metalanguage (labels, sense indicators, collocators, etc).
- For each proposal you should
  - specify any additional material that must be compiled in order to make the new feature possible, and give one or two examples of this material
  - choose an entry which you will use to illustrate the proposal
  - describe the electronic function you propose
  - use your word processing program to show what the new entry might look like in the proposed electronic version.
- Can you suggest some way of improving the dictionary for TL speakers?

Reading

Recommended reading

Atkins 1994, 1996, 2002; Atkins, Rundell, and Sato 2003.

Further reading on related topics

Adamska-Salaciak 2006; Apresjan 1992; Atkins and Varantola 1997, 1998; Béjoint and Thoiron 1996; Bogaards 1990; Bogaards and Hannay 2004; Bogaards and van der Kloot 2001; Corréard 1998; Cowie 1987a; Cummins and Desjardins 2002; Dobrovol'skij 2000; Duval 1991, 2002; Fontenelle 1992, 1996, 1997, 2000; Heylen and Maxwell 1994; Jarosova 2000; Katzaros 2004; Kilgarriff 1997a; Lew 2002, 2004; Macklovitch 1996; Marello 1989; Martin 1992; Neubert 1992; O'Neill and Palmer 1992; Piotrowski 1994; Roberts 1992; Roberts and Montgomery 1996; Rodger 2002; Rogers and Ahmad 1998; Salerno 1999; Snell-Hornby 1984; Zgusta 1984.

Bibliography

Adamska-Salaciak, A. (2006). 'Translation of Dictionary Examples – Notoriously Unreliable?', in Corino et al. (2006). 493–501.

Aitchison, J. (2003). *Words in the Mind.* 3rd edition. Oxford: Blackwell Publishing Ltd.

Algeo, J. (1993). 'Desuetude among New English Words', in *International Journal of Lexicography* 6.4. 281–293.

Alvar Ezquerra, M. (1992) (ed.). *EURALEX '90 Proceedings*. Barcelona: Bibliograf.

Apresjan, J. D. (1973). 'Regular Polysemy', in *Linguistics* 142. 5–39.

——(1992). 'Systemic Lexicography', in Tommola et al. (1992). 3–16.

——(2002). 'Principles of Systematic Lexicography', in Corréard (2002). 91–104. Reprinted in Fontenelle (2008).

Atkins, B. T. S. (1985). 'Monolingual & Bilingual Learners' Dictionaries: a Comparison', in R. F. Ilson (ed.), *Dictionaries, Lexicography & Language Learning*. Oxford: Pergamon. 15–24.

——(1993). 'Theoretical Lexicography and its Relation to Dictionary-making', in W. Frawley (ed.), *Dictionaries: the Journal of the Dictionary Society of North America*. Cleveland, OH: DSNA. 4–43. Reprinted in Fontenelle (2008).

——(1994). 'A Corpus-Based Dictionary', in *Preface to Oxford Hachette English & French Dictionary*. Oxford: Oxford University Press. xix–xxvi.

——(1995). 'Analysing the Verbs of Seeing: A Frame Semantics Approach to Corpus Lexicography', in S. Gahl, C. Johnson and A. Dolbey (eds.), *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society, 1994*. Berkeley, CA: BLS. 42–56.

——(1996). 'Bilingual Dictionaries: Past, Present and Future', in Gellerstam et al. (1996). 515–590. Reprinted in Corréard (2002).

——(1998) (ed.). *Using Dictionaries*: *Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Niemeyer.

——(2002). 'Then and Now: Competence and Performance in 35 Years of Lexicography', in Braasch and Povlsen (2002). 1–28. Reprinted in Fontenelle (2008).

——Clear, J. and Ostler, N. (1992). 'Corpus Design Criteria', in *Journal of Literary and Linguistic Computing* 7.1. 1–16.

——Fillmore, C. J. and Johnson, C. R. (2003). 'Lexicographic Relevance: Selecting Information from Corpus Evidence', in *International Journal of Lexicography* 16.3. 251–280.

——and Grundy, V. (2006). 'Lexicographic Profiling: an Aid to Consistency in Dictionary Entry Design', in Corino et al. (2006). 1097–1107.

Atkins, B. T. S., Kegl, J. and Levin, B. (1988). 'Anatomy of a Verb Entry: From Linguistic Theory to Lexicographic Practice', in *International Journal of Lexicography* 1.2. 84–126.

——and Levin, B. (1988). 'Admitting impediments', in *Information in Text*, Proceedings of the Fourth Annual Conference of the New OED Centre, University of Waterloo, Canada. 97–114. Also in U. Zernik (ed.), *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Assoc. Inc. (1991). 233–262.

————(1995). 'Building on a Corpus: A Linguistic and Lexicographical Look at some Near-Synonyms', in *International Journal of Lexicography* 8.2. 85–114.

————and Song, G. (1997). 'Making Sense of Corpus Data: A Case Study of Verbs of Sound', in *International Journal of Corpus Linguistics* 2.1. 23–64.

——Rundell, M. and Sato, H. (2003). 'The Contribution of FrameNet to Practical Lexicography', in *International Journal of Lexicography* 16.3. 333–358.

——and Varantola, K. (1997). 'Monitoring Dictionary Use', in *International Journal of Lexicography* 10.1. 1–45. Reprinted in Atkins (1998) and Fontenelle (2008).

————(1998). 'Language Learners Using Dictionaries: The Final Report of the EURALEX- and AILA-Sponsored Research Project into Dictionary Use', in Atkins (1998). 21–82.

Ayto, J. (1988). 'Fig. Leaves. Metaphor in Dictionaries', in Snell-Hornby (1988). 49–54.

Baroni, M., Kilgarriff, A., Pomikalek, J. and Rychly, P. (2006). 'WebBootCaT: a Web Tool for Instant Corpora', in Corino et al. (2006). 123–131.

Béjoint, H. (1990). 'Monosemy and the Dictionary', in Magay and Zigány (1990). 13–26.

——and Thoiron, P. (1996) (eds.). *Les dictionnaires bilingues*. Louvain-la-Neuve: Aupelf-Uref-Duculot.

Benson, M. (1990). 'Collocations in a General-Purpose Dictionary' *International Journal of Lexicography* 3.1. 23–34.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

——(1993). 'Representativeness in Corpus Design', in *Literary and Linguistic Computing* 8. 243–257.

——Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bogaards, P. (1990). 'Où cherche-t-on dans le dictionnaire?', in *International Journal of Lexicography* 3.2. 79–102.

——(1992). 'French Dictionary Users and Word Frequency', in Tommola et al. (1992). 51–62.

——(1996). 'Dictionaries for Learners of English', in *International Journal of Lexicography* 9.4. 277–320.

——(1998a). 'What Type of Words do Language Learners Look up?', in Atkins (1998). 152–157.
——(1998b). 'Scanning Long Entries in Learners' Dictionaries', in Fontenelle et al. (1998). 555–563.
——and Hannay, M. (2004). 'Towards a New Type of Bilingual Dictionary', in Williams and Vessier (2004). 463–474.
——and van der Kloot, W. A. (2001). 'The Use of Grammatical Information in Learners' Dictionaries', in *International Journal of Lexicography* 14.2. 97–121.
Bolinger, D. (1965). 'The Atomization of Meaning', in *Language* 41. 555–573.
——(1975). *Aspects of Language.* Second edition. New York: Harcourt Brace Jovanovich, Inc.
——(1980). *Language – The Loaded Weapon.* London: Longman.
——(1985) 'Defining the Indefinable', in R. F. Ilson (ed.), *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press. 69–75. Reprinted in Fontenelle (2008).
Bowker, L. (2006) (ed.). *Lexicography, Terminology, and Translation: Text-based Studies in Honour of Ingrid Meyer.* Ottawa: University of Ottawa Press.
Braasch, A. (2004). 'A Health Corpus Selected and Downloaded from the Web: is it Healthy enough?', in Williams and Vessier (2004). 71–78.
——(2006). Exploitation of Syntactic Patterns for Sense-Group Identification', in Corino et al. (2006). 133–139.
——and Povlsen, C. (2002) (eds.). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen: Center for Sprogteknologi.
Bullon, S. (1990). 'The Treatment of Connotation in Learners' Dictionaries', in Magay and Zigány (1990). 27–34.
Carter, R. (1989). Review article of LDOCE2 and COBUILD1, *International Journal of Lexicography* 1989, 2.1. 30–43.
Čermak, F. (2006). 'Collocations, Collocability and Dictionary', in Corino et al. (2006). 929–937.
Church, K. W. and Hanks, P. W. (1990). 'Word Association Norms, Mutual Information, and Lexicography', in *Computational Linguistics* 16. 22–29. Reprinted in Fontenelle (2008).
Citron, S. and Widmann, T. (2006). 'A Bilingual Corpus for Lexicographers', in Corino et al. (2006). 251–255.
Clark, E. and Clark, H. (1979). 'When Nouns Surface as Verbs', in *Language* 55. 767–781.
Clear, J. (1987). 'Computing', in Sinclair (1987). 41–61.
Coffey, S. (2006). 'Delexical Verb+Noun Phrases in Monolingual English Learners' Dictionaries', in Corino et al. (2006). 939–949.
Copestake, A. and Briscoe, T. (1995). 'Semi-Productive Polysemy and Sense Extension', in *Journal of Semantics* 12. 15–67.
Corino, E., Marello, C. and Onesti, C. (2006) (eds.). *Proceedings of 12th EURALEX International Congress, EURALEX 2006*. Alessandria: Edizioni Dell'Orso.

Corréard, M.-H. (1998). 'Traduire avec un dictionnaire, traduire pour un dictionnaire', in Fontenelle et al. (1998). 17–24.

——(2002) (ed.). *Lexicography and Natural Language Processing*: *A Festschrift in Honour of B. T. S. Atkins.* UK: Euralex.

Cowie, A. P. (1981). 'The Treatment of Collocations and Idioms in Learners' Dictionaries', in *Applied Linguistics* 2.3. 223–235.

——(1987a) (ed.). *The Dictionary and the Language Learner*. Tübingen: Niemeyer.

——(1987b). 'Syntax, the Dictionary and the Learner's Communicative Needs', in Cowie (1987a). 183–192.

——(1994). 'Phraseology', in R. E. Asher and J. M. Y. Simpson (eds.), *The Encyclopedia of Language and Linguistics Vol. 6*. Oxford: Pergamon Press. 3168–3171. Reprinted in Fontenelle (2008).

——(1998) (ed.). *Phraseology: Theory, Analysis and Applications*. Oxford: Clarendon Press. (Paperback edition, 2001.)

——(1999a). 'Phraseology and Corpora: Some Implications for Dictionary-Making', in *International Journal of Lexicography* 12.4. 307–323.

——(1999b). *English Dictionaries for Foreign Learners – a History*. Oxford: Oxford University Press.

——(2001). 'Homonymy, Polysemy and the Monolingual English Dictionary', in F. M. Dolezal et al. (eds.), *Lexicographica* 17. Tübingen: Niemeyer.

——and Howarth, P. (1996). 'Phraseology – a Select Bibliography', in *International Journal of Lexicography* 9.1. 38–51.

Crowdy, S. (1993). 'Spoken corpus design', in *Literary and Linguistic Computing* 8. 259–265.

——(1994). 'Spoken corpus transcription', in *Literary and Linguistic Computing* 9. 25–28.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

——(1990). 'Prototype Theory and Lexical Semantics', in Tsohatzidis (1990). 382–402.

——(2002). 'Aspects of the Micro-Structure of Word Meanings', in Ravin and Leacock (2000). 30–51.

——(2004). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Second edition. Oxford: Oxford University Press.

Csábi, S. (2002). 'Polysemous Words, Idioms, and Conceptual Metaphors: Cognitive Linguistics and Lexicography', in Braasch and Povlsen (2002). 249–254.

Cummins, S. and Desjardins, I. (2002). 'A Case Study in Lexical Research for Translation', in *International Journal of Lexicography* 15.2. 139–156.

Dalen-Oskam, K. van, Geirnaert, D. and Kruyt, J. (2002). 'Text Typology and Selection Criteria for a Balanced Corpus: the Integrated Language Database of 8th–21st Century Dutch', in Braasch and Povlsen (2002). 401–406.

de Schryver, G.-M. (2003). 'Lexicographers' Dreams in the Electronic-Dictionary Age', in *International Journal of Lexicography* 16.2. 143–199.

——and Joffe, D. (2004). 'On how electronic dictionaries are really used', in Williams and Vessier (2004). 187–196.

——and Prinsloo, D. (2000). 'Dictionary-making Process with "Simultaneous Feedback" from the Target Users to the Compilers', in Heid et al. (2000). 197–209.

Dobrovol'skij, D. (2000). 'Contrastive Idiom Analysis: Russian and German Idioms in Theory and in the Bilingual Dictionary', in *International Journal of Lexicography* 13.3. 169–186.

Drysdale, P. D. (1987). 'The Role of Examples in a Learners' Dictionary', in Cowie (1987a). 213–223.

Duval, A. (1991). 'L'équivalence dans le dictionnaire bilingue', in F. J. Hausmann, O. Reichmann, H. E. Wiegand and L. Zgusta (eds.), *Wörterbücher / Dictionaries / Dictionnaires.* Vol. 3. Berlin/New York: De Gruyter. 2817–2824. Reprinted in English in Fontenelle (2008).

——(1992). 'From the printed dictionary to the CD-ROM', in Alvar Ezquerra (1992). 79–88.

——(2002). 'La métalangue, un mal nécessaire du dictionnaire actif', in Corréard (2002). 45–59.

Dziemianko, A. and Lew, R. (2006). 'When you are Explaining the Meaning of a Word: The Effect of Abstract Noun Definition Format on Syntactic Class Identification', in Corino et al. (2006). 857–863.

Eijk, P. van der, Alejandro, O. and Florenza, M. (1995). 'Lexical Semantics and Lexicographic Sense Distinction', in *International Journal of Lexicography* 8.1. 1–27.

Fedorova, I. V. (2004). 'Style and Usage Labels in Learners' Dictionaries: Ways of Optimisation', in Williams and Vessier (2004). 265–272.

Fillmore, C. J. (1975). 'An Alternative to Checklist Theories of Meaning', in C. Cogen et al. (eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA: BLS. 123–131.

——(1989). Review article of LDOCE2 and COBUILD1, *International Journal of Lexicography* 1989, 2.1. 57–83.

——(1992). 'Corpus Linguistics or Computer-Aided Armchair Linguistics', in J. Svartvik (ed.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82, Stockholm, August 1991*. Berlin: Mouton de Gruyter. 35–66. Reprinted in Fontenelle (2008).

——(1995). 'The Hard Road From Verbs To Nouns', in M. Chen and O. Tzeng (eds.), *In Honor of William S-Y. Wang*. Taipei, Taiwan: Pyramid Press. 105–129.

——(1997). *On Deixis*. Stanford: CSLI Publications.

——(2002). 'Lexical Isolates', in Corréard (2002). 105–124.

——(2003). 'Double-decker Definitions: the Role of Frames in Meaning Explanations', in *Sign Language Studies* (Gallaudet University Press). 3.3. 263–295.

——(2005). 'Frame Semantics' in K. Brown (ed.), *Encyclopedia of Language and Linguistics*. Oxford: Elsevier.

——and Atkins, B. T. S. (1994). 'Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography', in B. T. S. Atkins and A. Zampolli (eds.), *Computational Approaches to the Lexicon*. Oxford: Oxford University Press. 350–393.

Fillmore, C. J. and Atkins, B. T. S. (2000). 'Describing Polysemy: The Case of *crawl*', in Ravin and Leacock (2000). 91–110.

Fletcher, W. H. (2004). 'Making the Web More Useful as a Source for Linguistic Corpora', in U. Connor and T. Upton (eds), *Applied Corpus Linguistics. A Multidimensional Perspective*. Amsterdam: Rodopi. 191–205.

Fontenelle, Th. (1992). 'Collocation Acquisition from a Corpus or from a Dictionary: a Comparison', in Tommola et al. (1992). 221–228.

——(1996). 'Ergativity, Collocations and Lexical Functions', in Gellerstam et al. (1996). 209–222.

——(1997). 'Using a Bilingual Dictionary to Create Semantic Networks', in *International Journal of Lexicography* 10.3. 275–303. Reprinted in Fontenelle (2008).

——(2000). 'A Bilingual Lexical Database for Frame Semantics', in *International Journal of Lexicography* 13.4. 232–248.

——(2002). 'Lexical Knowledge and Natural Language Processing', in Corréard (2002). 216–229.

——(2008) (ed.). *Practical Lexicography: A Reader.* Oxford: Oxford University Press.

——Hiligsmann, P., Michiels, A., Moulin, A. and Theissen, S. (1998) (eds.). *Proceedings of the Eighth EURALEX Congress.* Liège: University of Liège.

Fox, G. (1987). 'The Case for Examples', in Sinclair (1987). 137–149.

Geeraerts, D. (1990) 'The Lexicographical Treatment of Prototypical Polysemy', in Tsohatzidis (1990). 195–210.

——(1994). 'Vagueness's Puzzles, Polysemy's Vagaries', in *Cognitive Linguistics* 4.3. 223–272.

——(2000). 'Adding Electronic Value: the Electronic Version of the *Grote Van Dale*', in Heid et al. (2000). 75–84.

Gellerstam, M., Järborg, J., Malmgren, S.-G., Norén, K., Rogström, L. and Papmehl, C. R. (1996) (eds.). *EURALEX'96 Proceedings*. Gothenburg: Gothenburg University.

Grefenstette, G. (1998). 'The Future of Linguistics and Lexicographers: Will there be Lexicographers in the Year 3000?', in Fontenelle et al. (1998). 25–41. Reprinted in Fontenelle (2008).

——(2002). 'The WWW as a Resource for Lexicography' in Corréard (2002). 199–215.

Grossmann, F. and Tutin, A. (2003) (eds.). *Les Collocations: Analyse et traitement.* Travaux et Recherches en Linguistique Appliquée Série E. Amsterdam: De Werelt.

Hanks, P. W. (1979). 'To what Extent does a Dictionary Definition Define?', in R. R. K. Hartmann (ed.), *Dictionaries and their Users*. Exeter: University of Exeter. 32–38.

——(1987). 'Definitions and Explanations', in Sinclair (1987). 116–136.

——(1988). 'Typicality and Meaning Potentials', in Snell-Hornby (1988). 37–48.

——(1990). 'Evidence and Intuition in Lexicography', in J. Tomaszczyk and B. Lewandowska-Tomaszczyk (eds.), *Meaning and Lexicography*. Amsterdam/Philadelphia: John Benjamins. 31–41.
——(1993). 'Lexicography: theory and practice', in W. Frawley (ed.), *Dictionaries: the Journal of the Dictionary Society of North America*. Cleveland, OH: DSNA. 97–112.
——(1994). 'Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers Need Prototype Theory, and Vice Versa', in F. Kiefer, G. Kiss and J. Pajzs (eds.), *Papers in Computational Lexicography: Complex '94*. Budapest: Hungarian Academy of Sciences.
——(1998). 'Enthusiasm and Condescension', in Fontenelle et al. (1998). 151–166.
——(2000a). 'Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance', in Heid et al. (2000). 3–14.
——(2000b). 'Do Word Meanings Exist?', in *Computers and the Humanities* 34: 205–215. Reprinted in Fontenelle (2008).
——(2001). 'The Probable and the Possible: Lexicography in the Age of the Internet', in Sangsup Lee (ed.), *AsiaLex 2001 Proceedings*. Seoul: Yonsei University. 1–15.
——(2002). 'Mapping Meaning onto Use', in Corréard (2002). 156–198.
——(2004a). 'Corpus Pattern Analysis', in Williams and Vessier (2004). 87–97.
——(2004b). 'The Syntagmatics of Metaphor and Idiom', in *International Journal of Lexicography* 17.3. 245–274.
——(2005). 'Johnson and Modern Lexicography', in *International Journal of Lexicography* 18.2. 243–266.
——Urbschat, A. and Gehweiler, E. (2006). 'German Light Verb Constructions in Corpora and Dictionaries', in *International Journal of Lexicography* 19.4. 439–457.
Hartmann, R. R. K. (1984). (ed.) *LEXeter'83 Proceedings*. Tübingen: Niemeyer.
Hausmann, F. J. (1989). 'Le dictionnaire de collocations', in Hausmann et al. (1989). 1010–1019.
——(1991). 'Collocations in Monolingual and Bilingual English Dictionaries', in V. Ivir and D. Kalogjera (eds.), *Languages in Contact and Contrast*. Berlin: Mouton de Gruyter. 225–236.
——and Gorbahn, A. (1989). Review article of *LDOCE2* and *COBUILD1*, in *International Journal of Lexicography* 2.1. 44–56.
——Reichman, O., Wiegand, H. E. and Zgusta, L. (1989) (eds.). *Wörterbücher /Dictionaries / Dictionnaires*. Vol. 1. Berlin/New York: Walter de Gruyter.
——and Wiegand, H. E. (1989). 'Component Parts and Structures of General Monolingual Dictionaries: A Survey', in Hausmann et al. (1989). 328–360.
Heid, U. (1994). 'On Ways Words Work Together – Topics in Lexical Combinatorics', in Martin et al. (1994). 226–257.
——(1998). 'Towards a Corpus-Based Dictionary of German Noun–Verb Collocations', in Fontenelle et al. (1998). 301–312.

Heid, U., Evert, S., Lehmann, E. and Rohrer, C. (2000) (eds.). *Proceedings of the Ninth EURALEX Congress*. Stuttgart: University of Stuttgart.

Herbst, T. (1996). 'On the Way to the Perfect Learners' Dictionary: A First Comparison of OALD5, LDOCE3, COBUILD2 and CIDE', in *International Journal of Lexicography* 9.4. 321–357.

Heylen, D. and Maxwell, K. (1994). 'Lexical Functions and the Translation of Collocations', in Martin et al. (1994). 298–303.

Heyvaert, F. J. (1994). 'The Edges of Definition', in Martin et al. (1994). 84–92.

Hoey, M. (2005). *Lexical Priming: a New Theory of Words and Language*. London: Routledge.

Hulstijn, J. H. and Atkins, B. T. S. (1998). 'Empirical Research on Dictionary Use in Foreign-Language Learning: Survey and Discussion', in Atkins (1998). 7–19.

Humble, P. (1998). 'The Use of Authentic, Made-up, and "Controlled" Examples in Foreign Language Dictionaries', in Fontenelle et al. (1998). 593–600.

Hunston, S. (2007). 'Semantic Prosody Revisited', in *International Journal of Corpus Linguistics* 12.2. 249–268.

Jansen, J., Mergeai, J. and Vanandroye, J. (1987). 'Controlling LDOCE's Controlled Vocabulary', in Cowie (1987a). 78–94.

Jarosova, A. (2000). 'Problems of Semantic Subdivisions in Bilingual Dictionary Entries', in *International Journal of Lexicography* 13.1. 12–28.

Joffe, D. and de Schryver, G-M. (2004). 'TshwaneLex, a State-of the-Art Dictionary Compilation Program', in Williams and Vessier (2004). 99–104.

Johnson, Samuel (1747). *The Plan of a Dictionary of the English Language*; *Addressed to the Right Honourable Philip Dormer, Earl of Chesterfield*. London: J. and P. Knapton [etc.]. (Edited by Jack Lynch). Reprinted in Fontenelle (2008).

——(1755). *Preface to a Dictionary of the English Language*. (Edited by Jack Lynch). <http://andromeda.rutgers.edu/~jlynch/Texts/preface.html>

Katzaros, V. (2004). 'The Different Functions of Illustrative Examples in Learners' Bilingual Dictionaries', in Williams and Vessier (2004). 487–494.

Kay, M. (1983). 'The Dictionary of the Future and the Future of the Dictionary', in A. Zampolli and A. Cappelli (eds.), *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries: Proceedings of the European Science Foundation Workshop, Pisa '81*. (Linguistica Computazionale III.) Pisa: Giardini Editori. 161–74.

Keller, F., Lapata, M. and Ourioupina, O. (2002). 'Using the Web to Overcome Data Sparseness', in J. Hajič and Y. Matsumoto (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Philadelphia. 230–237.

Kilgarriff, A. (1994). 'The Myth of Completeness and some Problems with Consistency', in Martin et al. (1994). 101–116.

——(1997a). 'I don't Believe in Word Senses', in *Computers and the Humanities* 31.2. 91–113. Reprinted in Fontenelle (2008).

——(1997b). 'Putting Frequencies in the Dictionary', in *International Journal of Lexicography* 10.2. 135–155.
——(1998). 'SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation', in Fontenelle et al. (1998). 167–174.
——(2006a). 'Word Senses', in E. Agirre and P. Edmonds (eds.), *Word Sense Disambiguation*: *Algorithms and Applications*. New York: Springer. 29–45.
——(2006b). 'Collocationality (and how to Measure it)', in Corino et al. (2006). 997–1004.
——and Grefenstette, G. (2003). 'Introduction to the Special Issue on the Web as Corpus', in *Computational Linguistics*, 29.3. 333–348. Reprinted in Fontenelle (2008).
——and Rundell, M. (2002). 'Lexical Profiling Software and its Lexicographic Applications: Case Study', in Braasch and Povlsen (2002). 807–819.
————and Uí Dhonnchadha, E. (2007). 'Efficient Corpus Development for Lexicography: Building the New Corpus for Ireland', in *Language Resources and Evaluation* 40.2: 127–152.
——and Tugwell, D. (2002). 'Sketching Words', in Corréard (2002). 125–137.
——Rychly, P., Smrz, P. and Tugwell, D. (2004). 'The Sketch Engine', in Williams and Vessier (2004). 105–116. Reprinted in Fontenelle (2008).
Knowles, F. (1996). 'Lexicographical Aspects of Health Metaphors in Financial Text', in Gellerstam et al. (1996). 789–796.
Knowles, M. and Moon, R. E. (2006). *Introducing Metaphor*. London: Routledge.
Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.
Lakoff, G. (1987). *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.
——and Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.
——(1993). 'Wierzbicka's Theory and the Practice of Lexicography', in W. Frawley (ed.), *Dictionaries: the Journal of the Dictionary Society of North America.* Cleveland, OH: DSNA. 113–119.
Laufer, B. (1992). 'Corpus-based Examples versus Lexicographer Examples in Comprehension and Production of New Words', in Tommola et al. (1992). 71–76.
Leemets, H. (1992). 'Translating the "Untranslatable" Words', in Tommola et al. (1992). 473–478.
Lehrer, A. (1990). 'Prototype Theory and its Implications for Lexical Analysis', in Tsohatzidis (1990). 368–381.
Lemmens, M. and Wekker, H. (1986). *Grammar in English Learners' Dictionaries* (Lexicographica Series Maior 16). Tübingen: M. Niemeyer.
Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Lew, R. (2002). 'A Study in the Use of Bilingual and Monolingual Dictionaries by Polish Learners of English: A Preliminary Report', in Braasch and Povlsen (2002). 759–771.

——(2004). *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English.* Poznań: Motivex.

——and Dziemianko, A. (2006). 'A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and its Usefulness for Conveying Syntactic Information', in *International Journal of Lexicography* 19.3. 225–242.

Lewandowska-Tomaszczyk, B. (1988). 'Universal Concepts and Language-Specific Meaning', in Snell-Hornby (1988). 17–26.

——(1990). 'Conversational Data and Lexicographic Practice', in Magay and Zigány (1990). 207–214.

Littré, E. (1880). *Comment j'ai fait mon dictionnaire.* Reprinted (1995) with postface by J. Cellard. Arles: Editions Philippe Picquier.

Lorentzen, H. (1996). 'Lemmatization of MultiWord Lexical Units: In which Entry?', in Gellerstam et al. (1996). 415–422.

Louw, Bill (1993). 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in M. Baker et al. (eds.), *Text and Technology.* Amsterdam: Benjamins 157–176.

Luna, P. (2004). 'Not just a Pretty Face: the Contribution of Typography to Lexicography', in Williams and Vessier (2004). 847–858.

Lyons, J. (1969). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

——(1977). *Semantics I & II.* Cambridge: Cambridge University Press.

——(1981). *Language and Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Mackintosh, K. (1998). 'An Empirical Study of Dictionary Use in L2-L1 Translation', in Atkins (1998). 123–149.

——(2006). 'How Dictionaries are Influenced by Social Values', in Bowker (2006). 45–63.

Macklovitch, E. (1996). 'Les dictionnaires bilingues en-ligne et le poste de travail du traducteur', in Béjoint and Thoiron (1996). 169–180.

Magay, T. and Zigány, J. (1990) (eds.). *BudaLEX'88 Proceedings*. Budapest: Akadémiai Kiadó.

Marello, C. (1989). *Dizionari bilingui.* Bologna: Zanichelli Editore S.p.A.

——(1998). 'Hornby's Bilingualized Dictionaries', in *International Journal of Lexicography* 11.4. 292–314.

Martin, W. (1992). 'On the organization of semantic data in passive bilingual dictionaries', in Alvar Ezquerra (1992). 193–201.

——and Al, B. P. F. (1990). 'User-Orientation in Dictionaries: Nine Propositions', in Magay and Zigány (1990).

——Meijs, W., Moerland, M., ten Pas, E., van Sterkenburg, P. and Vossen, P. (1994) (eds.). *Proceedings of the Sixth EURALEX Congress*. Amsterdam.

McArthur, T. (1986). *Worlds of Reference.* Cambridge: Cambridge University Press.
McCreary, D. R. (2002). 'American Freshmen and English Dictionaries: "I had *aspersions* of becoming an English Teacher" ', in *International Journal of Lexicography* 15.3. 181–205.
—— and Amacker, E. (2006). 'Experimental Research on College Students' Usage of Two Dictionaries', in Corino et al. (2006). 871–885.
—— and Dolezal, F. T. (1999). 'A Study of Dictionary Use by ESL Students in an American University', in *International Journal of Lexicography* 12.1. 107–146.
McEnery, T. and Wilson, A. (2001). *Corpus Linguistics – An Introduction*. 2nd edition. Edinburgh: Edinburgh University Press.
Meer, G. van der (1996). 'The Treatment of Figurative Meanings in the English Learner's Dictionaries (OALD, LDOCE, CC and CIDE)', in Gellerstam et al. (1996). 423–430.
—— (1998). 'Collocations as One Particular Type of Conventional Word Combination: their Definition and Character', in Fontenelle et al. (1998). 313–322.
—— (1999). 'Metaphors and Dictionaries: the Morass of Meaning, or How to Get Two Ideas for One', in *International Journal of Lexicography* 12.3. 195–208.
—— (2000). 'Core, Subsense, and the *New Oxford Dictionary of English* (*NODE*)', in Heid et al. (2000). 419–432.
—— (2004). 'On Defining: Polysemy, Core Meanings, and "Great Simplicity" ', in Williams and Vessier (2004). 807–815.
—— (2006). 'It's about *Time*: On Coherence and Simplicity in Dictionary Entries', in *English Studies* (Routledge) 87.5. 602–616.
Mel'čuk, I. A. (1988). 'Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria', in *International Journal of Lexicography* 1.3. 165–188.
—— (1996). 'Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon', in L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins. 37–102.
—— (1998). 'Collocations and Lexical Functions', in Cowie (1998). 23–54.
—— Clas, A. and Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve: Duculot.
Meyer, I., Zaluski, V., Mackintosh, K. and Foz, C. (1998). 'Metaphorical Internet Terms in English and French', in Fontenelle et al. (1998). 523–531.
Miller, G. and Gildea, P. (1985). 'How to Misread a Dictionary', in *AILA Bulletin* 1985. 13–26.
—— et al. (1990). Special Issue of *IJL* devoted to WordNet. *International Journal of Lexicography* 3.4. 235–244.
Moon, R. E. (1987a). 'The Analysis of Meaning', in Sinclair (1987). 86–103.
—— (1987b). 'Monosemous Words and the Dictionary' in Cowie (1987a). 173–182.
—— (1988). ' "Time" and Idioms', in Snell-Hornby (1988). 107–116.
—— (1992). ' "There is Reason in the Roasting of Eggs": a Consideration of Fixed Expressions in Native-Speaker Dictionaries', in Tommola et al. (1992). 493–502.

Moon, R. E. (1996). 'Data, Description, and Idioms in Corpus Lexicography', in Gellerstam et al. (1996). 245–256.

——(1998). *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

——(2004). 'On Specifying Metaphor: an Idea and its Implementation', in *International Journal of Lexicography* 17.2. 195–222.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms*. Cambridge: Cambridge University Press.

Murray, K. E .M. (1979). *Caught in the Web of Words: James A. H. Murray and the Oxford English Dictionary.* Oxford: Oxford University Press.

Nakamoto, K. (1998). 'From Which Perspective Does the Definer Define the Definiendum', in *International Journal of Lexicography* 11.3. 205–218.

Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries.* Tübingen: Niemeyer. Lexicographica Series Maior 98.

——and Haill, R. (2002). 'A Study of Dictionary Use by International Students at a British University', in *International Journal of Lexicography* 15.4. 277–305.

Neubert, A. (1992). 'Fact and Fiction of the Bilingual Dictionary', in Alvar Ezquerra (1992). 29–44.

Norri, J. (1996). 'Regional Labels in Some British and American Dictionaries', in *International Journal of Lexicography* 9.1. 1–29.

——(2000). 'Labelling of Derogatory Words in Some British and American Dictionaries', in *International Journal of Lexicography* 13.2. 71–106.

Nuccorini, S. (1994). 'On Dictionary Misuse', in Martin et al. (1994). 586–597.

Nunberg, G. and Zaenen, A. (1992). 'Systematic Polysemy in Lexicology and Lexicography', in Tommola et al. (1992). 387–396.

O'Neill, M. and Palmer, C. (1992). 'Editing a bilingual dictionary entry within the framework of a bidirectional dictionary', in Alvar Ezquerra (1992). 211–218.

Ostler, N. and Atkins, B. T. S. (1992). 'Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules', in J. Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Commonsense Reasoning*. New York: Springer-Verlag. 87–98.

Palmer, M. (2000). 'Consistent Criteria for Sense Distinctions', in *Computers and the Humanities* 34. 217–222.

Pinker, S. (1997). *How the Mind Works.* London: Penguin Books.

Piotrowski, T. (1988). 'Defining Natural-Kind Words', in Snell-Hornby (1988). 55–62.

——(1994). *Problems in Bilingual Lexicography*. Wrocław: Wrocław University Press.

Potter, L. (1998). 'Setting a Good Example: What Kind of Examples Best Serve the Users of Learners' Dictionaries?', in Fontenelle et al. (1998). 357–362.

Prinsloo, D. (2008). 'Criteria for Corpus Design for the Creation of Bilingual Dictionaries', in R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.),

Dictionaries: An International Encyclopedia of Lexicography. Supplementary volume: Recent developments, with special focus on computational lexicography. Berlin: Mouton de Gruyter.

Ravin, Y. and Leacock, C. (2000) (eds.). *Polysemy: Linguistic and Computational Approaches*. Oxford: Oxford University Press. (Paperback 2002).

Renouf, A. (1987). 'Corpus Development', in Sinclair (1987). 10–40.

Roberts, R. P. (1992). 'Organization of Information in a Bilingual Dictionary Entry', in Alvar Ezquerra (1992). 291–314.

—— and Bossé-Andrieu, J. (2006). 'Corpora and Translation', in Bowker (2006). 201–214.

—— and Montgomery, C. (1996). 'The Use of Corpora in Bilingual Lexicography', in Gellerstam et al. (1996). 457–464.

Robins, R. H. (1987). 'Polysemy and the Lexicographer', in R. Burchfield (ed.), *Studies in Lexicography*. Oxford: Clarendon Press. 52–75.

Rodger, L. (2002). 'Is a Bilingual Dictionary Possible?', in Braasch and Povlsen (2002). 435–440.

Rogers, M. A. and Ahmad, K. (1998). 'The Translator and The Dictionary: Beyond Words?', in Atkins (1998). 193–204.

Rosch, E. H. (1973). 'Natural Categories', in *Cognitive Psychology* 4. 328–350.

—— (1975). 'Cognitive Representations of Semantic Categories', in *Journal of Experimental Psychology: General* 104: 192–233.

Rundell, M. (1988). 'Changing the Rules: Why the Monolingual Learner's Dictionary Should Move Away from the Native-Speaker Tradition', in Snell-Hornby (1988). 127–138.

—— (1998). 'Recent Trends in English Pedagogical Lexicography', in *International Journal of Lexicography* 11.4. 315–342.

—— (1999). 'Dictionary Use in Production', in *International Journal of Lexicography* 12.1. 35–53.

—— (2002). 'Good Old-Fashioned Lexicography: Human Judgment and the Limits of Automation', in Corréard (2002). 138–155.

—— (2006). 'More than One Way to Skin a Cat: Why Full-Sentence Definitions have not been Universally Adopted', in Corino et al. (2006). 323–337. Reprinted in Fontenelle (2008).

—— and Atkins, B. T. S. (forthcoming). 'Criteria for Corpus Design for the Creation of Monolingual Dictionaries', in R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.), *Dictionaries: An International Encyclopedia of Lexicography.* Supplementary volume: Recent developements, with special focus on computational lexicography. Berlin: Mouton de Gruyter.

—— and Stock, P. (1992). 'The Corpus Revolution', in *English Today* 30, 31, and 32.

Ruppenhofer, J., Baker, C. and Fillmore, C. J. (2002). 'Collocational Information in the FrameNet Database', in Braasch and Povlsen (2002). 359–369.

Ruus, H. (2002). 'A Corpus-based Electronic Dictionary for (Re)search', in Braasch and Povlsen (2002). 175–185.

Salerno, L. (1999). 'Grammatical Information in the Bilingual Dictionary: A Study of Five Italian-French Dictionaries', *International Journal of Lexicography* 12.3. 209–222.

Scholfield, P. (1999). 'Dictionary Use in Reception', in *International Journal of Lexicography* 12.1. 13–34.

Schutz, R. (2002). 'Indirect Offensive Language in Dictionaries', in Braasch and Povlsen (2002). 637–642.

Selva, T., Verlinde, S. and Binon, J. (2002). 'Le DAFLES, un nouveau dictionnaire électronique pour apprenants de français', in Braasch and Povlsen (2002). 199–208.

Sharpe, P. (1989). 'Pragmatic Considerations for an English-Japanese Dictionary', in *International Journal of Lexicography* 2.4. 315–323.

——(1995). 'Electronic Dictionaries with Particular Reference to the Design of an Electronic Bilingual Dictionary for English-speaking Learners of Japanese', in *International Journal of Lexicography* 8.1. 39–54.

Shcherba, L. V. (1995). 'Towards a General Theory of Lexicography', translated and annotated by Donna M. T. Farina, in *International Journal of Lexicography* 8.4. 314–350.

Siepmann, D. (2005). 'Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects', in *International Journal of Lexicography* 18.4. 409–443.

——(2006). 'Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects', in *International Journal of Lexicography* 19.1. 1–39.

Silva, P. (2000). 'Time and Meaning: Sense and Definition in the OED', in L. Mugglestone (ed.), *Lexicography and the OED*. Oxford: Oxford University Press. 77–95.

Simpson, J. (2003). 'The Production and Use of Occurrence Examples', in van Sterkenburg (2003). 260–272.

Sinclair, J. M. (1987) (ed.). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

——(1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

——(1996). 'The Search for Units of Meaning', in *TEXTUS IX, 1*, Special Volume, L. Merlini and J. M. Sinclair (eds.), *Lessico e Morfología*, 75–106. Reprinted in J. M. Sinclair and R. Carter (eds.), *Trust the Text*. London: Routledge (2004).

——(2003). 'Corpora for Lexicography', in van Sterkenburg (2003). 167–178.

——(2004). 'In Praise of the Dictionary', in Williams and Vessier (2004). 1–12.

——(2005). 'Corpus and Text – Basic Principles' in M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. 1–16. Also available online at <http://ahds.ac.uk/linguistic-corpora/>

——et al. (1996). 'Corpus to Corpus: A Study of Translation Equivalence', in *International Journal of Lexicography* 9.3. 171–178.

Snell-Hornby, M. (1984). 'The Bilingual Dictionary – Help or Hindrance?', in Hartmann (1984). 274–282.

——(1988) (ed.). *ZüriLEX '86 Proceedings*. Tübingen: Francke Verlag.
Sterkenburg, P. van (2003) (ed.). *A Practical Guide to Lexicography*. Amsterdam: J. Benjamins.
Stock, P. (1984). 'Polysemy'in Hartmann (1984). 131–140. Reprinted in Fontenelle (2008).
——(1988). 'The Structure and Function of Definitions', in Snell-Hornby (1988). 81–90.
——(1992). 'The Cultural Dimension in Defining', in Tommola et al. (1992). 113–120.
Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.
——(2001). 'Texts, Corpora, and Problems of Interpretation', in *Applied Linguistics* 22:2. 149–172.
Summers, D. (1993). 'Longman/Lancaster English Language Corpus: Criteria and Design', in *International Journal of Lexicography* 6.3. 181–208.
Swanepoel, P. (2006). 'Capturing Semantic Relativity in Dictionary Definitions – the Case of Defining "Imaginary" Beings and "Imaginary" Attributes', in Corino et al. (2006). 1271–1276.
Taylor, J. R. (1990). 'Schemas, Prototypes, and Models: in Search of the Unity of the Sign', in Tsohatzidis (1990). 521–534.
——(1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Second edition. Oxford: Clarendon Press.
Tognini-Bonelli, E. (1996). 'Towards Translation Equivalence from a Corpus Linguistic Perspective', in *International Journal of Lexicography* 9.3. 197–217.
Tommola, H., Varantola, K., Salmi-Tolonen, T. and Schopp, J. (1992). (eds.) *EURALEX'92 Proceedings*. Tampere: University of Tampere.
Tsohatzidis, S. L. (1990) (ed.) *Meanings and Prototypes*. London: Routledge.
Vandeloise, C. (1990). 'Representation, prototypes, and centrality', in Tsohatzidis (1990). 403–437.
Varantola, K. (1998). 'Translators and their Use of Dictionaries: User Needs and User Habits', in Atkins (1998). 179–192.
——(2002). 'Use and Usability of Dictionaries: Common Sense and Context Sensibility', in Corréard (2002). 30–44.
——(2006). 'The Contextual Turn in Learning to Translate', in Bowker (2006). 215–226.
Veisbergs, A. (2002). 'Defining Political Terms in Lexicography: Recent Past and Present', in Braasch and Povlsen (2002). 657–668.
Verlinde, S., Dancette, J. and Binon, J. (1998). 'Redéfinir la Définition', in Fontenelle et al. (1998). 375–384.
Vilpula, M. (1995). 'The Sun and the Definition of Day', in *International Journal of Lexicography* 8.1. 29–38.
Vossen, P. (2004). 'EuroWordNet: A Multilingual Database of Autonomous and Language-Specific WordNets Connected via an Inter-lingual Index', in *International Journal of Lexicography* 17.2. 161–173.

Wakely, R. (1998). 'The Treatment of French Reflexive Verbs in Bilingual Dictionaries', in Fontenelle et al. (1998). 421–430.

Walter, E. (1992). 'Semantic set-defining: Benefits to the Lexicographer and the User', in Tommola et al. (1992). 129–136.

—— and Harley, A. (2002). 'The Role of Corpus and Collocation Tools in Practical Lexicography', in Braasch and Povlsen (2002). 851–864.

Whitcut, J. (1988). 'Lexicography in Simple Language', in *International Journal of Lexicography* 1.1. 49–55.

Whitsitt, S. (2005). 'A Critique of the Concept of Semantic Prosody', in *International Journal of Corpus Linguistics* 10.3. 283–305.

Wiegand, H. E. (1984). 'On the Structure and Contents of a General Theory of Lexicography', in R. R. K. Hartmann (1984). 13–30.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor, MI: Karoma.

——(1987). *English Speech Act Verbs*. Sydney: Academic Press.

——(1990). ' "Protypes Save": on the Uses and Abuses of the Notions of "Prototype" in Linguistics and Related Fields', in Tsohatzidis (1990). 347–365.

——(1993). 'What are the Uses of Theoretical Lexicography?', in W. Frawley (ed.), *Dictionaries: the Journal of the Dictionary Society of North America*. Cleveland, OH: DSNA. 44–78.

Williams, G. and Vessier, S. (2004) (eds.). *EURALEX 2004 Proceedings*. Lorient: Université de Bretagne-Sud.

Wilson, D. (forthcoming). 'Relevance and lexical pragmatics' in *Italian Journal of Linguistics/Rivista di Linguistica*, Special Issue on Pragmatics and the Lexicon. 15.2. 273–291.

Zaenen, A. (2002). 'Musings about the Impossible Electronic Dictionary', in Corréard (2002). 230–244.

Zgusta, L. (1971). *Manual of Lexicography*. The Hague: Mouton.

——(1984). 'Translation Equivalence in the Bilingual Dictionary', in Hartmann (1984). 155–165.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Cambridge, MA: Houghton Mifflin.

Index