

SSCI RESEARCH PAPER • 双语对照版

从术语管理到术语治理：生成式人工智能辅助翻译中的人机协同术语治理框架研究

From Terminology Management to Terminology Governance: A Human-AI Collaborative Framework for Generative AI-Assisted Translation

SSCI Research Paper • 双语对照版

AUTHOR Research Team

DATE 2026-09-19

VERSION v2.0

Contents

从术语管理到术语治理：生成式人工智能辅助翻译中的人机协同术语治理框架研究

From Terminology Management to Terminology Governance: A Human-AI Collaborative Framework for Generative AI-Assisted Translation

1. Introduction

一、引言

2. Literature Review

2.1 Terminology science: from prescriptive standardization to socioterminology

2.2 Translation technology studies: from CAT to machine translation to generative AI

2.3 Human-AI collaboration and human-centred AI

2.4 AI governance and responsible AI

2.5 Bilingual lexicography and the AI turn

2.6 Synthesis and research gap

二、文献综述

2.1 术语学：从规定性标准化到社会术语学

2.2 翻译技术研究：从CAT到机器翻译再到生成式AI

2.3 人机协同与以人为本的人工智能

2.4 人工智能治理与负责任的人工智能

2.5 双语词典编纂与AI转向

2.6 综合与研究空白

3. Theoretical Framework

3.1 Terminology Management Theory

3.2 Socioterminology

3.3 Translation Technology Studies

3.4 Human-Centred AI	
3.5 AI Governance	
3.6 Integration: from management to governance	
三、理论框架	
3.1 术语管理理论	
3.2 社会术语学	
3.3 翻译技术研究	
3.4 以人为本的人工智能	
3.5 人工智能治理	
3.6 整合：从管理到治理	
4. Methodology	
4.1 Research design	
4.2 Data sources	
4.3 Case selection rationale	
4.4 Analytical procedure	
4.5 Positionality and reflexivity	
4.6 Limitations of the methodology	
四、研究方法	
4.1 研究设计	
4.2 数据来源	
4.3 案例选择依据	
4.4 分析程序	
4.5 立场性与反身性	
4.6 方法论的局限	
5. Case Study	
5.1 Case 1: yarrow(Achillea millefolium)	

5.2 Case 2:trunk	
5.3 Cross-case synthesis	
五、案例研究	
5.1 案例一：yarrow（Achillea millefolium，蓍草）	
5.2 案例二：trunk（树干/躯干/象鼻/行李箱）	
5.3 跨案例综合	
6. The Proposed Terminology Governance Framework (TGF)	
6.1 Layer 1: Source Authority Governance	
6.2 Layer 2: Semantic Alignment Governance	
6.3 Layer 3: Terminology Evaluation Mechanism	
6.4 Layer 4: Human–AI Decision Collaboration	
6.5 Layer 5: Terminology Lifecycle Management	
6.6 The Terminology Governance Layer as a system component	
六、所提出的术语治理框架（TGF）	
6.1 第一层：来源权威治理	
6.2 第二层：语义对齐治理	
6.3 第三层：术语评价机制	
6.4 第四层：人机协同决策	
6.5 第五层：术语全生命周期管理	
6.6 术语治理层作为系统组件	
7. Discussion	
7.1 Why "governance" rather than "management"?	
7.2 Prompt engineering is not governance	
7.3 The role of the human expert in the GenAI era	
7.4 The specificity of English–Chinese bilingual terminology	
七、讨论	

7.1 为何用"治理"而非"管理"?

7.2 提示词工程不是治理

7.3 人类专家在生成式AI时代的角色

7.4 英汉双语术语的特殊性

8. Theoretical Contributions

八、理论贡献

9. Practical Implications

九、实践启示

10. Conclusion, Limitations, and Future Research

10.1 Conclusion

10.2 Limitations

10.3 Future research agenda

十、结论、局限与未来研究

10.1 结论

10.2 局限

10.3 未来研究议程

References

从术语管理到术语治理：生成式人工智能辅助翻译中的人机协同术语治理框架研究

From Terminology Management to Terminology Governance: A Human-AI Collaborative Framework for Generative AI-Assisted Translation

Abstract

Generative artificial intelligence (GenAI) has rapidly transformed professional translation, lexicography, and terminology work, yet the conceptual vocabulary used to describe terminology practice has not kept pace with this shift. Traditional Terminology Management (TM), rooted in mid-twentieth-century standardization thinking, treats terminology primarily as a stored, retrievable resource to be kept consistent across texts. This paper argues that, in the era of large language models (LLMs), such an account is no longer sufficient. We introduce the construct of **Terminology Governance (TG)**, which reframes terminology as a contingent knowledge asset whose authority, semantic alignment, evaluation, human-AI division of labour, and lifecycle must be explicitly governed. Drawing on a synthesis of terminology theory (Sager, 1990; ISO 704:2022; Temmerman, 2000; Antia, 2015), translation technology studies (O'Hagan, 2013; García, 2019), human-centred AI (Floridi et al., 2018), and AI governance scholarship, the paper develops a **five-layer Terminology Governance Framework (TGF)**: (1) Source Authority Governance, (2) Semantic Alignment Governance, (3) Terminology Evaluation Mechanism, (4) Human-AI Decision Collaboration, and (5) Terminology Lifecycle Management. The framework is grounded in a qualitative case study of two representative terminology problems encountered during an AI-assisted bilingual dictionary project: *yarrow* (*Achillea millefolium*), which illustrates inter-authority conflict over standard naming, and *trunk*, which illustrates sense-level mismatch in a highly polysemous common word. We show that prompt engineering cannot substitute for governance, and that future AI translation systems require an explicit Terminology Governance Layer mediating between the LLM and translated output. The paper contributes a concept upgrade from management to governance, a testable framework, and a research agenda for responsible AI-assisted translation.

Keywords: terminology governance; terminology management; generative AI; human–AI collaboration; translation studies; AI governance; socioterminology; bilingual lexicography

摘要

生成式人工智能（GenAI）已迅速变革了专业翻译、词典编纂与术语工作，然而用于描述术语实践的概念词汇却未能跟上这一转变的步伐。传统的术语管理（Terminology Management, TM）植根于二十世纪中叶的标准化思维，主要将术语视为一种可存储、可检索的资源，旨在保持其在不同文本中的一致性。本文认为，在大语言模型（Large Language Models, LLMs）时代，这种认识已不再充分。本文提出术语治理（Terminology Governance, TG）这一构念，将术语重新界定为一种权变性知识资产，其权威性、语义对齐、评价、人机分工及生命周期均须得到明确治理。本文综合术语学理论（Sager, 1990; ISO 704:2022; Temmerman, 2000; Antia, 2015）、翻译技术研究（O'Hagan, 2013; García, 2019）、以人为本的人工智能（Floridi et al., 2018）及人工智能治理研究，构建了一个五层术语治理框架（Terminology Governance Framework, TGF）：（1）来源权威治理，（2）语义对齐治理，（3）术语评价机制，（4）人机协同决策，（5）术语全生命周期管理。该框架以一项AI辅助双语词典项目中遇到的两个代表性术语问题为定性案例基础：*yarrow*（*Achillea millefolium*，蓍草），用以说明标准命名中的权威冲突；以及 *trunk*（树干/躯干/象鼻/行李箱），用以说明高度多义常用词的义项不匹配问题。本文表明，提示词工程无法替代治理，未来的AI翻译系统需要一个明确的术语治理层（Terminology Governance Layer, TGL）作为大语言模型与翻译输出之间的中介。本文的贡献在于实现了从管理到治理的概念升级、提供了一个可检验的框架，并为负责任的AI辅助翻译提出了研究议程。

关键词： 术语治理；术语管理；生成式人工智能；人机协同；翻译研究；人工智能治理；社会术语学；双语词典编纂

1. Introduction

Generative artificial intelligence has moved, within a remarkably short period, from research prototype to a routine component of professional translation workflows. Translators, revisers, terminologists, and lexicographers now routinely interact with large language models (LLMs) that generate candidate translations, draft definitions, propose bilingual equivalents, and even assemble entry-level dictionary content. The practical appeal is obvious: speed, coverage, and a baseline level of linguistic competence

that was unavailable only a decade ago. Yet the very capacity that makes LLMs useful—their ability to produce fluent, contextually plausible language without explicit retrieval from a curated terminological database—also creates a set of problems that the existing conceptual apparatus of translation studies and terminology science was not designed to address.

Traditional **Terminology Management (TM)** is a well-established discipline. It is concerned with the systematic collection, description, storage, retrieval, and update of terminological units in order to ensure consistency across documents, products, and organizational communication (Sager, 1990; ISO 704:2022). In corporate and localization settings, TM has long been operationalized through term bases, controlled vocabularies, extraction tools, and retrieval plugs integrated into computer-assisted translation (CAT) environments. The implicit assumption of this tradition is that terminology is a relatively stable resource: once a term is approved, it is stored, retrieved, and reused, with updates made through a defined curation process.

The arrival of generative AI complicates this picture in several interlocking ways. First, LLMs do not simply retrieve approved terms; they **generate** candidate terms, many of which have no provenance, no authority trail, and no established consensus in the relevant domain. Second, LLMs are remarkably fluent at producing plausible bilingual equivalents even when authoritative sources disagree, masking underlying conflicts behind a veneer of confidence. Third, the boundary between "source term" and "translation equivalent" becomes blurred when the model itself mediates between them. Fourth, the lifecycle of a term is no longer linear (create, approve, use, retire); instead, it is continuously renegotiated through every model interaction, every user prompt, and every revised corpus.

These shifts raise a question that this paper treats as its central problem: **what conceptual and operational vocabulary do we need in order to govern terminology practice when the system producing terms is itself a generative, probabilistic, and distributed artificial intelligence?**

We argue that the answer requires a conceptual upgrade from **Terminology Management** to **Terminology Governance (TG)**. The distinction is not merely rhetorical. Management, in its classical sense, presupposes a relatively stable object (a term base, a controlled vocabulary) and a bounded authority (a terminologist, a standards body). Governance, by contrast, presupposes a more complex and dynamic configuration of actors, sources, uncertainties, values, and accountabilities. Governance asks not only "how do we store and retrieve terms?" but also: Who has the authority to decide which term is correct? How do we adjudicate conflicts between authoritative sources? How do we evaluate the quality

of a proposed term? Where should the boundary of competence lie between the human expert and the AI system? And how do we track a term across its entire lifecycle—creation, validation, publication, revision, retirement, and archiving—in a way that is auditable and reversible?

The paper develops this argument as follows. Section 2 reviews the relevant literature across four domains: terminology science and socioterminology, translation technology studies, human-AI collaboration, and AI governance. Section 3 articulates the theoretical framework by integrating five perspectives. Section 4 describes the methodology, combining conceptual framework development with a qualitative case study. Section 5 presents two in-depth cases—*yarrow* and *trunk*—drawn from the authors' field notes during an AI-assisted bilingual dictionary project. Section 6 introduces the five-layer Terminology Governance Framework (TGF) and discusses each layer in detail. Section 7 discusses the framework in relation to the existing literature. Section 8 outlines theoretical contributions, Section 9 practical implications, and Section 10 concludes with limitations and a future research agenda.

一、引言

生成式人工智能在极短的时间内，已从研究原型转变为专业翻译 workflow 中的常规组成部分。译者、审校者、术语学家与词典编纂者如今常与大语言模型（LLMs）交互，后者能够生成候选译文、起草定义、提出双语对等词，甚至汇编词条级词典内容。其实用价值显而易见：速度、覆盖面，以及仅仅十年前尚不可得的基线语言能力。然而，正是使大语言模型具有实用性的那种能力——即无需从经过精心编纂的术语数据库中显式检索，便能产生流畅且语境上看似合理的语言——同时也造成了一系列翻译研究与术语学的现有概念工具未曾设计应对的问题。

传统的术语管理（TM）是一门成熟的学科。它关注术语单元的系统收集、描述、存储、检索与更新，以确保跨文档、产品及组织沟通的一致性（Sager, 1990; ISO 704:2022）。在企业与本地化环境中，术语管理长期以来通过术语库、受控词表、抽取工具及集成到计算机辅助翻译（CAT）环境中的检索插件来实现操作。这一传统的隐含假设是：术语是一种相对稳定的资源——一旦术语获得批准，便被存储、检索与复用，更新则通过既定的策展流程进行。

生成式人工智能的到来以若干相互关联的方式使这一图景变得复杂。第一，大语言模型并非简单地检索已批准术语，而是生成候选术语，其中许多来源不明、无权威追溯路径，也未在相关领域形成共识。第二，即使权威来源之间存在分歧，大语言模型也能极为流畅地生成看似合理的双语对等词，以自信的外表掩盖底层冲突。第三，当模型本身在源术语与翻译对等词之间充当中介时，“源术语”与“翻译对等词”之间的边界变得模糊。第四，术语的生命周期不再是线性的（创建、批准、使用、废止）；相反，它通过每一次模型交互、每一条用户提示词以及每一次修订语料而被持续重新协商。

这些转变引出了本文视为核心问题的一个问题：当生成术语的系统本身是一个生成式、概率性且分布式的人工智能时，我们需要什么样的概念与操作词汇来治理术语实践？

我们认为，答案要求从术语管理到术语治理（TG）的概念升级。这一区分并非修辞性的。经典意义上的管理以相对稳定的客体（术语库、受控词表）和有限的权威（术语学家、标准制定机构）为前提。治理则以更为复杂、动态的行动者、来源、不确定性、价值与问责配置为前提。治理不仅追问“我们如何存储和检索术语？”，还追问：谁有权决定哪个术语是正确的？我们如何裁决权威来源之间的冲突？我们如何评价候选术语的质量？人类专家与AI系统之间的能力边界应在何处？我们如何以可审计、可追溯的方式追踪一个术语的整个生命周期——创建、验证、发布、修订、废止与归档？

本文按如下结构展开论证。第二节综述四个领域的相关文献：术语学与社会术语学、翻译技术研究、人机协同以及AI治理。第三节通过整合五种视角阐述理论框架。第四节描述研究方法，结合概念框架开发与定性案例研究。第五节呈现两个深度案例——yarrow（蓍草）与 trunk——取自作者在一项AI辅助双语词典项目中的田野笔记。第六节介绍五层术语治理框架（TGF）并逐层详加讨论。第七节将框架置于既有文献中加以讨论。第八节概述理论贡献，第九节阐述实践启示，第十节以局限与未来研究议程作结。

2. Literature Review

2.1 Terminology science: from prescriptive standardization to socioterminology

Modern terminology science emerged in the mid-twentieth century in close association with standardization efforts in industry and international organization (Sager, 1990). The foundational [general theory of terminology](#), often associated with the Vienna School and with Eugen Wüster, held that concepts could be defined unambiguously, that definitions should precede designations, and that a one-to-one correspondence between concept and term was the ideal to be pursued. The International Standard ISO 704, [Terminology work—Terminology work and terminology science—Principles and methods](#), codifies much of this tradition, defining concepts, definitions, domain delimitation, and term formation in prescriptive terms (ISO 704:2022).

Sager (1990) provided the most systematic English-language account of this tradition, distinguishing between the establishment of terminology (the process of analysing concepts and designating them), the processing of terminology (storage, retrieval, presentation), and the application of terminology in translation and technical writing. Sager's model is fundamentally managerial: it assumes an identifiable expert community, a determinate domain, and a stable concept system. The terminologist's role is to

identify, define, and record terms so that translators and technical writers can retrieve and reuse them.

By the 1990s, however, a number of scholars began to challenge the realist, one-to-one presuppositions of the Wüsterian tradition. **Socioterminology**, associated particularly with Gaudin (1993) and the French terminological tradition, argued that terms are not stable entities waiting to be discovered but social constructions embedded in professional communities, discourses, and power relations. A term is not simply "correct" or "incorrect"; it is used, contested, and transformed within communities of practice. Temmerman (2000), in her seminal work *Towards New Ways of Terminology Description*, extended this critique by drawing on cognitive semantics and prototype theory. She argued that many specialized concepts—especially in the life sciences and social sciences—do not exhibit classical necessary-and-sufficient feature structure but rather prototype structure, fuzzy boundaries, and family resemblance. This implies that terminological descriptions must be flexible, context-sensitive, and sensitive to the user community rather than purely prescriptive.

Antia (2015) consolidated and extended these critiques in her work on terminology and language planning, emphasizing that terminological practice is inherently normative and political, involving choices among competing variants, and that the sociolinguistic context in which terms circulate shapes which variants become authoritative. Antia's work is particularly relevant here because it foregrounds the question of **authority**: who decides, on what grounds, and with what accountability. This question, latent in socioterminology, becomes unavoidable in the GenAI context, where the "decision" is distributed across a model, a prompt, a retrieval context, and a human reviewer.

2.2 Translation technology studies: from CAT to machine translation to generative AI

Translation technology has evolved through several distinct phases (O'Hagan, 2013). The first phase—computer-assisted translation (CAT)—introduced translation memory (TM) and terminology databases as productivity tools that augmented but did not replace the human translator. The second phase—statistical machine translation (SMT)—introduced probabilistic translation that could produce novel output but was often unreliable at the level of terminology. The third phase—neural machine translation (NMT)—sharpened fluency but introduced a new set of consistency and reliability problems, including "hallucinated" terms and unpredictable register shifts.

García (2019) and O'Hagan (2013) have been prominent voices in documenting these shifts. O'Hagan (2013) argued that translation technology is not a neutral tool but a

sociotechnical arrangement that reshapes the translator's skillset, identity, and epistemic position. García (2019), examining the impact of machine translation on the translation profession, noted that the division of labour between human and machine is itself an object of political and economic negotiation. Both scholars emphasize that "translator" is not a stable role but one that is continuously reconstructed as technology changes.

The arrival of **generative AI**—and specifically LLMs that can produce fluent, context-aware bilingual output—marks a further shift. Unlike NMT systems, which are typically trained on parallel corpora and constrained by segment-level inputs, LLMs can engage in multi-turn dialogue, justify their choices, cite sources, and revise output on request. This makes them, in principle, much better partners for terminological work. But it also means that the translator's epistemic task has changed: the question is no longer "is this segment correct?" but "is this **claim about terminology** justified, and by whom?" Existing translation technology studies have not yet fully theorized this shift. Most scholarship on NMT-era terminology (e.g., reviews of termbase integration in CAT tools) assumes that terminology is a retrievable resource. GenAI forces us to recognize that terminology is increasingly a **generated** one, and that governance, not retrieval, is the central challenge.

2.3 Human–AI collaboration and human-centred AI

A third relevant literature concerns the design of collaborative systems in which humans and artificial intelligences share a task. The notion of **human-centred AI**, articulated by Floridi et al. (2018) and by the European Commission's High-Level Expert Group on AI, holds that AI systems should augment human agency, support human autonomy, and remain under meaningful human control. In the translation context, this means that the AI should be designed to **complement** rather than replace the translator's terminological judgment.

However, the empirical record on human–AI collaboration in professional settings is mixed. On the one hand, well-designed AI systems can reduce cognitive load, surface alternatives, and catch errors. On the other hand, over-reliance on AI—sometimes called "automation bias"—can lead humans to endorse incorrect AI output simply because it is fluent and confident. In the specific domain of terminology, automation bias is particularly dangerous: an LLM that produces a plausible-sounding equivalent for an obscure botanical term is, to a non-specialist, indistinguishable from one that has retrieved a correct authoritative name. This makes the design of the **division of labour** between human and AI not merely an efficiency question but an epistemic and ethical one.

2.4 AI governance and responsible AI

Finally, the AI governance literature provides a set of concepts that are directly applicable to the terminology problem. Floridi and Cowls (2019), in their work on the European approach to AI, identify four principles—respect for human autonomy, prevention of harm, fairness, and explicability—as the foundations of a responsible AI ecosystem. Translated into the terminology domain, these principles imply that terminological decisions made by AI systems should be explicable (i.e., traceable to sources), fair (i.e., not systematically biased toward one linguistic or cultural tradition), and subject to meaningful human oversight.

This literature has, however, largely focused on high-stakes domains such as medical diagnosis, criminal justice, and autonomous vehicles. It has not yet been systematically applied to language work—translation, lexicography, and terminology—despite the fact that language work is itself high-stakes when it shapes public understanding of science, law, or policy. This paper aims to extend the AI governance literature into the specifically linguistic domain, arguing that terminology is a paradigmatic case of a "knowledge asset" whose governance becomes urgent once generative AI enters the workflow.

2.5 Bilingual lexicography and the AI turn

A fourth, smaller literature addresses the specific implications of AI for bilingual lexicography. Traditional bilingual dictionary making is a painstaking, human-curated enterprise in which lexicographers select headwords, define senses, identify equivalents, and craft illustrative examples. The arrival of corpus-driven and, more recently, LLM-driven approaches has raised questions about the role of the lexicographer, the status of "equivalence" between languages, and the reliability of generated entries. Relevant here is the long-standing lexicographical distinction between **translation equivalent** (a word in the target language that, in a particular sense, replaces the source word) and **explanatory translation** (a gloss that helps the user understand the meaning of a source word, even if it is not a conventional target-language equivalent). This distinction is directly relevant to the terminology governance problem, and we return to it in the case studies.

2.6 Synthesis and research gap

The literature review reveals a clear gap. Terminology science has developed rich theories of concepts, definitions, and communities, but has not systematically addressed the governance challenges introduced by generative AI. Translation technology studies have documented the sociotechnical reshaping of the translator's role but have not produced a terminologically specific governance model. Human-centred AI and AI governance offer general principles but have not been applied to the language-technical domain. Bilingual

lexicography has a sophisticated understanding of equivalence but has not, until recently, confronted a system that generates equivalents at scale. This paper addresses that gap by integrating these four traditions into a single, actionable framework.

二、文献综述

2.1 术语学：从规定性标准化到社会术语学

现代术语学于二十世纪中叶兴起，与工业及国际组织的标准化工作密切相关（Sager, 1990）。奠基性的普通术语学理论与维也纳学派及 Eugen Wüster 相关联，主张概念可以被无歧义地界定，定义应先于称名，概念与术语之间的一一对应应是应追求的理想目标。国际标准 ISO 704《术语工作——术语工作与术语学——原则与方法》将这一传统的大部分内容法典化，以规定性方式界定概念、定义、领域划定与术语构词（ISO 704:2022）。

Sager（1990）对这一传统做出了最系统的英文论述，区分了术语的确立（分析概念并为其命名的过程）、术语的处理（存储、检索、呈现）以及术语在翻译与技术写作中的应用。Sager 的模型在根本上是管理性的：它假设存在一个可识别的专家共同体、一个确定的领域以及一个稳定的概念系统。术语学家的角色是识别、定义和记录术语，以便译者与技术写作者检索并复用。

然而，到了二十世纪九十年代，若干学者开始挑战 Wüster 传统的实在论与一一对应预设。社会术语学尤其与 Gaudin（1993）及法国术语学传统相关联，主张术语并非等待被发现的稳定实体，而是嵌入专业共同体、话语与权力关系之中的社会建构。一个术语并非简单地“正确”或“错误”；它在实践共同体中被使用、被争议、被转化。Temmerman（2000）在其开创性著作《迈向术语描写的新路径》中，借助认知语义学与原型理论扩展了这一批判。她认为，许多专门概念——尤其是生命科学与社会科学中的概念——并不呈现经典的必要且充分特征结构，而是呈现原型结构、模糊边界与家族相似性。这意味着术语描写必须灵活、语境敏感，并对用户共同体保持敏感，而非纯粹规定性的。

Antia（2015）在其关于术语与语言规划的著作中整合并扩展了这些批判，强调术语实践本质上是规范性与政治性的，涉及在竞争变体之间做出选择，而术语流通的社会语言学语境塑造了哪些变体成为权威。Antia 的工作在此尤为相关，因为它将权威问题前置：谁来决定、依据什么理由、承担什么问责。这一潜在于社会术语学中的问题，在生成式AI语境中变得不可避免，因为“决策”分布在模型、提示词、检索语境与人类审校者之间。

2.2 翻译技术研究：从CAT到机器翻译再到生成式AI

翻译技术经历了若干不同阶段的演进（O'Hagan, 2013）。第一阶段——计算机辅助翻译（CAT）——引入了翻译记忆库（TM）与术语数据库作为增强人类译者但不替代其工作的生产力工具。第二阶段——统计机器翻译（SMT）——引入了概率性翻译，能够产生新颖输出，但在术语层面往往不

可靠。第三阶段——神经机器翻译（NMT）——提升了流畅度，但也带来了一系列新的一致性与可靠性问题，包括“幻觉”术语与不可预测的语域偏移。

García (2019) 与 O'Hagan (2013) 一直是记录这些转变的重要声音。O'Hagan (2013) 认为，翻译技术并非中性工具，而是一种社会技术安排，它重塑译者的技能组合、身份认同与认知地位。García (2019) 考察了机器翻译对翻译职业的影响，指出人机之间的劳动分工本身就是政治与经济协商的对象。两位学者均强调，“译者”并非一个稳定角色，而是随技术变迁而持续重构的角色。

生成式AI的到来——特别是能够产生流畅、语境感知双语输出的大语言模型——标志着又一次转变。与通常基于平行语料训练且受句段级输入约束的NMT系统不同，大语言模型能够进行多轮对话、论证其选择、引用来源并按需修订输出。这使它们在原则上成为术语工作更好的伙伴。但这也意味着译者的认知任务已经改变：问题不再是“这个句段正确吗？”，而是“这个关于术语的主张是否成立、由谁佐证？”现有的翻译技术研究尚未充分理论化这一转变。大多数关于NMT时代术语的研究（例如对CAT工具中术语库集成的综述）都假设术语是可检索的资源。生成式AI迫使我们认识到，术语日益成为一种生成的资源，治理而非检索才是核心挑战。

2.3 人机协同与以人为本的人工智能

第三类相关文献涉及人机共享任务的协同系统设计。以人为本的人工智能理念由 Floridi 等人 (2018) 及欧盟人工智能高级专家组阐述，主张AI系统应增强人类能动性、支持人类自主，并保持在有意义的人类控制之下。在翻译语境中，这意味着AI应被设计为补充而非替代译者的术语判断。

然而，关于专业环境中人机协同的实证记录好坏参半。一方面，设计良好的AI系统可以降低认知负荷、呈现备选方案并捕捉错误。另一方面，过度依赖AI——有时被称为“自动化偏差”——可能导致人类仅仅因为AI输出流畅自信就予以认可。在术语这一特定领域，自动化偏差尤其危险：一个大语言模型为一个生僻植物学术语生成一个听起来合理的对等词，对非专业人士而言，与它检索到正确的权威名称无法区分。这使得人机之间劳动分工的设计不仅是效率问题，更是一个认识论与伦理问题。

2.4 人工智能治理与负责任的人工智能

最后，AI治理文献提供了一套可直接应用于术语问题的概念。Floridi 与 Cowls (2019) 在其关于欧洲AI路径的著作中，确立了四项原则——尊重人类自主、预防伤害、公平性与可解释性——作为负责任AI生态系统的基石。将其转译至术语领域，这些原则意味着AI系统做出的术语决策应当是可解释的（即可追溯至来源）、公平的（即不系统性偏向某一语言或文化传统），并接受有意义的人类监督。

然而，该文献主要聚焦于高风险领域，如医疗诊断、刑事司法与自动驾驶。它尚未系统地应用于语言工作——翻译、词典编纂与术语——尽管语言工作在塑造公众对科学、法律或政策的理解时本身就是高风险的。本文旨在将AI治理文献延伸至特定的语言领域，论证术语是一种“知识资产”的典型示例，一旦生成式AI进入工作流，其治理便变得紧迫。

2.5 双语词典编纂与AI转向

第四类规模较小的文献探讨了AI对双语词典编纂的具体影响。传统双语词典编纂是一项艰苦的、由人工精心策展的事业，词典编纂者选择词头、界定义项、识别对等词并精心撰写例证。语料库驱动以及更近的大语言模型驱动方法的出现，引发了关于词典编纂者角色、语言间"对等"地位以及生成词条可靠性的问题。此处相关的是词典学中长期存在的区分：**翻译对等词**（目标语中在特定定义项下替代源语词的词）与**解释性翻译**（帮助用户理解源语词意义的释义，即使它不是常规的目标语对等词）。这一区分与术语治理问题直接相关，我们将在案例研究中回到这一点。

2.6 综合与研究空白

文献综述揭示了一个清晰的空白。术语学发展了关于概念、定义与共同体的丰富理论，但尚未系统应对生成式AI引入的治理挑战。翻译技术研究记录了译者角色的社会技术重塑，但尚未产出一个术语学特定的治理模型。以人为本的AI与AI治理提供了一般原则，但尚未应用于语言技术领域。双语词典编纂对对等性有精细理解，但直到最近才面对一个大规模生成对等词的系统。本文通过将这四种传统整合为一个单一、可操作的框架来填补这一空白。

3. Theoretical Framework

The framework proposed in this paper draws on five theoretical perspectives. Each contributes a distinct lens through which the terminology governance problem can be understood.

3.1 Terminology Management Theory

The first lens is the classical theory of terminology management (Sager, 1990; ISO 704:2022). This perspective provides the foundational vocabulary: **concept**, **definition**, **designation**, **domain**, **term system**, **term base**. It also provides the baseline model of good practice: concepts are delimited, definitions are formulated, designations are recorded, and terms are managed as retrievable resources. The TGF builds on this tradition but extends it: where classical TM assumes a stable object, the TGF assumes a dynamic, AI-mediated object.

3.2 Socioterminology

The second lens is socioterminology (Gaudin, 1993; Temmerman, 2000; Antia, 2015). This perspective reminds us that terms are social, contested, and community-bound. It forces us to ask not only "what is the correct term?" but "correct for whom, in which community, and with what authority?" In the GenAI context, this question becomes particularly sharp: the "community" now includes the model's training data, which may reflect a global, heterogeneous, and unaccountable mix of usage.

3.3 Translation Technology Studies

The third lens is translation technology studies (O'Hagan, 2013; García, 2019). This perspective situates terminology work within the broader sociotechnical ecosystem of professional translation. It emphasizes that tools reshape practice, and that the introduction of GenAI is not merely a productivity change but a change in the epistemic structure of translation work. The TGF is designed to be embedded within this ecosystem rather than to abstract away from it.

3.4 Human-Centred AI

The fourth lens is human-centred AI (Floridi et al., 2018). This perspective provides normative principles: AI should augment human agency, remain under meaningful control, and be explicable. Translated into the terminology domain, this means that the AI's role should be that of a candidate-generating, evidence-summarizing assistant, while the human expert retains final authority over terminological decisions. The TGF explicitly codifies this division of labour.

3.5 AI Governance

The fifth lens is AI governance (Floridi & Cowls, 2019). This perspective provides the macro-level principles of accountability, transparency, fairness, and robustness. Translated into the terminology domain, it means that terminological decisions must be traceable, that conflicts between sources must be visible rather than hidden, and that errors must be correctable across the lifecycle of the term.

3.6 Integration: from management to governance

These five lenses are integrated into a single argument. Classical TM provides the starting point: terminology is a real practice with real conventions. Socioterminology problematizes the idea that "correctness" is self-evident. Translation technology studies locate the practice within a changing sociotechnical system. Human-centred AI provides the normative principle that humans must remain in control. AI governance provides the structural principles of accountability and traceability. Together, they justify the conceptual upgrade from **management** to **governance**: where management asks how to **administer** a stable resource, governance asks how to **steer** a dynamic, distributed, and AI-mediated knowledge asset.

三、理论框架

本文提出的框架借鉴了五种理论视角。每一种视角都提供了一个独特的透镜，透过它可以理解术语治理问题。

3.1 术语管理理论

第一个透镜是经典的术语管理理论（Sager, 1990; ISO 704:2022）。这一视角提供了基础词汇：概念、定义、称名、领域、术语系统、术语库。它还提供了良好实践的基线模型：概念被划定、定义被表述、称名被记录，术语作为可检索资源被管理。TGF建立在这一传统之上但加以扩展：经典TM假设一个稳定的客体，而TGF假设一个动态的、AI中介的客体。

3.2 社会术语学

第二个透镜是社会术语学（Gaudin, 1993; Temmerman, 2000; Antia, 2015）。这一视角提醒我们，术语是社会的、有争议的、受共同体约束的。它迫使我们不仅追问"正确的术语是什么？"，还要追问"对谁正确、在哪个共同体中、依据什么权威？"在生成式AI语境中，这一问题变得尤为尖锐："共同体"现在包括模型的训练数据，后者可能反映一种全球的、异质的、不可问责的用法混合。

3.3 翻译技术研究

第三个透镜是翻译技术研究（O'Hagan, 2013; García, 2019）。这一视角将术语工作置于专业翻译更广泛的社会技术生态系统中。它强调工具重塑实践，而生成式AI的引入不仅是生产力变化，更是翻译工作认识论结构的变化。TGF旨在嵌入这一生态系统之中，而非从中抽象出来。

3.4 以人为本的人工智能

第四个透镜是为本的AI（Floridi et al., 2018）。这一视角提供了规范性原则：AI应增强人类能动性、保持在有意义的控制之下，并具有可解释性。转译至术语领域，这意味着AI的角色应是候选生成、证据汇总的助手，而人类专家保留术语决策的最终权威。TGF明确将这一劳动分工法典化。

3.5 人工智能治理

第五个透镜是AI治理（Floridi & Cows, 2019）。这一视角提供了问责、透明、公平与稳健性等宏观层面原则。转译至术语领域，它意味着术语决策必须可追溯，来源之间的冲突必须可见而非隐藏，错误必须在术语生命周期中可纠正。

3.6 整合：从管理到治理

这五个透镜被整合为一个单一论证。经典TM提供了起点：术语是一种有真实惯例的真实实践。社会术语学将"正确性"自明这一问题化。翻译技术研究将实践置于一个变化的社会技术系统中。以人为本的AI提供了人类必须保持控制的规范原则。AI治理提供了问责与可追溯的结构原则。它们共同论

证了从管理到治理的概念升级：管理追问如何管理一个稳定资源，治理追问如何引导一个动态的、分布式的、AI中介的知识资产。

4. Methodology

4.1 Research design

This paper adopts a **design-oriented conceptual research** approach. It combines two methodological strands: (a) **Conceptual Framework Development** (CFD), in which a novel conceptual distinction—terminology **management** vs. terminology **governance**—is articulated, refined, and operationalized through engagement with prior theory; and (b) a **Qualitative Case Study**, in which two purposively selected terminology problems are analysed in depth in order to illustrate, stress-test, and refine the emerging framework.

This design is appropriate for a problem that is at once theoretical (it requires a new conceptual vocabulary) and practical (it must be grounded in real terminological decisions). It is also consistent with methodological guidance in translation studies and information science, which advocates for design-science approaches when a novel sociotechnical arrangement requires both theoretical articulation and empirical grounding.

4.2 Data sources

The empirical material for the case study is drawn from two complementary sources, both of which were generated during the authors' participation in an ongoing AI-assisted bilingual dictionary project (a Webster's-style English-Chinese bilingual dictionary augmented by generative AI).

First source: multi-round human–AI dialogue archives. Over a period of several days, the authors engaged in repeated, structured dialogue with two general-purpose LLMs (ChatGPT and DeepSeek) about specific terminology problems encountered during the dictionary project. These dialogues were recorded, exported, and anonymized. They capture the full lifecycle of a terminology question: initial proposal, source retrieval, conflict identification, expert adjudication, and final resolution. The dialogues are particularly valuable because they reveal the AI's epistemic behaviour under uncertainty: how it handles conflicting sources, how it hedges, how it (sometimes) hallucinates authorities, and how it responds to expert correction.

Second source: public forum discussion records. The project also maintained a discussion forum in which human contributors—lexicographers, subject-matter specialists, and advanced users—debated specific entries. These records were coded to

identify recurring patterns of disagreement, the kinds of evidence participants invoked, and the criteria they implicitly used to evaluate proposed terms. The forum data complements the dialogue data by showing how **human** participants themselves reason about terminology authority, and where their disagreements lie.

4.3 Case selection rationale

Two cases were selected for in-depth analysis. The selection was purposive, designed to cover the two most structurally distinct kinds of terminology problem that the project encountered.

Case 1: yarrow (*Achillea millefolium*). This case represents the **proper noun / species-name problem**: a lexical item whose correct target-language equivalent is a **standardized** name that should in principle be determinable by consulting authoritative sources. The problem is that different authoritative sources (bilingual dictionaries, floras, standard nomenclature references) give different Chinese renderings, and the AI has no principled way to choose among them. This case tests the **source authority** and **evaluation** dimensions of the framework.

Case 2: trunk. This case represents the **polysemy / sense-mismatch problem**: a highly polysemous common word whose senses include "woody stem of a tree," "human or animal torso," "elongated muscular nose of an elephant," and "large travelling case." A naïve word-level binding of **trunk** to a single Chinese equivalent produces systematic errors. The problem is structural: the unit of terminological matching must be the **sense**, not the word. This case tests the **semantic alignment** dimension of the framework.

Together, the two cases cover the two poles of the terminology governance problem: (a) **which source should win when sources conflict**, and (b) **at what granularity should the match be made**. They are complementary rather than redundant.

4.4 Analytical procedure

Analysis proceeded in four steps.

Step 1: Dialogue coding. The dialogue archives were coded for segments in which (a) the AI proposed a term, (b) the AI cited or failed to cite a source, (c) the human identified a conflict or an error, (d) the AI revised its proposal, and (e) a final decision was recorded. Coding was conducted iteratively, with emergent categories refined against the source materials.

Step 2: Case extraction. For each of the two selected cases, the full dialogue thread was extracted and reconstructed as a chronological narrative, noting the sequence of proposals, evidence, revisions, and resolutions.

Step 3: Framework induction. Emerging patterns were compared against the five theoretical perspectives (Section 3) to identify which aspects of the problem the existing theories did and did not capture. The five-layer TGF was inductively derived from these comparisons, with each layer corresponding to a recurring problem identified in the cases.

Step 4: Verification. The emerging framework was reviewed against the forum discussion records to check whether the layers and their sub-components resonated with the kinds of disagreements that human participants actually raised. Discrepancies were used to refine the framework.

4.5 Positionality and reflexivity

The authors were not neutral observers of the terminology process; they were active participants in the dictionary project. This positionality is both a limitation (it risks confirmation bias) and a strength (it provides access to the fine-grained epistemic work of terminology adjudication that external researchers rarely see). The case narratives below are written in the third person and anonymized, but the analytical categories are grounded in the authors' own practical reasoning. We report this reflexively in the spirit of action research in translation studies.

4.6 Limitations of the methodology

The methodology has three important limitations, which we state explicitly. First, the case material is drawn from a single bilingual project (English–Chinese). The framework's transferability to other language pairs and domains is an empirical question that requires future research. Second, the dialogue data were generated by two specific LLMs; different models may exhibit different epistemic behaviours. Third, this paper does not report a controlled experimental evaluation of the framework. The Delphi studies, expert surveys, and AI-output quality experiments that would be needed to validate the framework quantitatively are identified as future work in the Conclusion.

四、研究方法

4.1 研究设计

本文采用设计导向的概念研究路径。它结合了两条方法论脉络：（a）概念框架开发（Conceptual Framework Development, CFD），即通过与既有理论对话，阐述、精炼并操作化一个新颖的概念区分——术语管理与术语治理；以及（b）定性案例研究，即深度分析两个有目的选取的术语问题，以说明、压力测试并精炼正在形成的框架。

这一设计适用于一个同时具有理论性（需要新的概念词汇）与实践性（必须扎根于真实术语决策）的问题。它也与翻译研究与信息科学中的方法论指导一致，后者主张当一种新颖的社会技术安排既需要理论阐述又需要实证基础时，应采用设计科学路径。

4.2 数据来源

案例研究的实证材料取自两个互补来源，二者均产生于作者参与一项正在进行的AI辅助双语词典项目（一部由生成式AI增强的韦氏风格英汉双语词典）期间。

第一来源：多轮人机对话档案。 在数天时间里，作者就词典项目中遇到的特定术语问题，与两个通用大语言模型（ChatGPT与DeepSeek）进行了反复的、结构化的对话。这些对话被记录、导出并匿名化。它们捕捉了一个术语问题的完整生命周期：初始提议、来源检索、冲突识别、专家裁决与最终解决。这些对话尤为宝贵，因为它们揭示了AI在不确定性下的认知行为：它如何处理冲突来源、如何模糊措辞、如何（有时）幻觉权威来源，以及如何回应专家纠正。

第二来源：公开论坛讨论记录。 该项目还维护了一个讨论论坛，人类贡献者——词典编纂者、学科专家与高级用户——在其中辩论特定词条。这些记录被编码，以识别反复出现的分歧模式、参与者援引的证据类型以及他们隐含用于评价候选术语的标准。论坛数据补充了对话数据，展示了人类参与者自身如何推理术语权威，以及他们的分歧何在。

4.3 案例选择依据

选取了两个案例进行深度分析。选择是有目的的，旨在覆盖项目所遇到的两种结构上最不同的术语问题。

案例一：yarrow (*Achillea millefolium*, 蓍草)。 本案例代表专名/物种名问题：一个词汇条目，其正确的目标语对等词是一个标准化名称，原则上应通过查阅权威来源来确定。问题在于不同权威来源（双语词典、植物志、标准命名参考文献）给出了不同的中文译法，而AI没有原则性的方法在它们之间做出选择。本案例测试了框架的来源权威与评价维度。

案例二：trunk (树干/躯干/象鼻/行李箱)。 本案例代表多义性/义项不匹配问题：一个高度多义的常用词，其义项包括"树的木质茎干"、"人或动物的躯干"、"大象伸长的肌肉质鼻子"以及"大型旅行箱"。将 trunk 朴素地词级绑定为单一中文对等词会产生系统性错误。问题是结构性的：术语匹配的单位必须是义项，而非词。本案例测试了框架的语义对齐维度。

两个案例共同覆盖了术语治理问题的两极：（a）当来源冲突时哪个来源应当胜出，以及（b）匹配应在什么粒度上进行。它们是互补的，而非冗余的。

4.4 分析程序

分析分四步进行。

第一步：对话编码。 对对话档案进行编码，标记以下片段：（a）AI提出术语；（b）AI引用或未能引用来源；（c）人类识别出冲突或错误；（d）AI修订其提议；以及（e）记录最终决策。编码以迭代方式进行，涌现类别参照原始材料加以精炼。

第二步：案例提取。 对于两个选定案例中的每一个，提取完整对话线索并按时间顺序重构为叙事，记录提议、证据、修订与解决的序列。

第三步：框架归纳。 将涌现的模式与五种理论视角（第三节）进行比较，以识别既有理论捕捉到与未捕捉到问题的哪些方面。五层TGF从这些比较中归纳得出，每一层对应案例中识别出的一个反复出现的问题。

第四步：验证。 参照论坛讨论记录审查正在形成的框架，检验各层及其子组件是否与人类参与者实际提出的分歧类型相呼应。差异被用于精炼框架。

4.5 立场性与反身性

作者并非术语过程的中立观察者；他们是词典项目的积极参与者。这种立场性既是局限（有确认偏误的风险），也是优势（能够接触到外部研究者很少看到的术语裁决的精细认知工作）。下文的案例叙事以第三人称撰写并匿名化，但分析类别扎根于作者自身的实践推理。我们本着翻译研究中行动研究的精神，反身性地报告这一点。

4.6 方法论的局限

本研究有三个重要局限，我们在此明确指出。第一，案例材料取自单一双语项目（英汉）。框架向其他语言对与领域的可迁移性是一个需要未来研究的实证问题。第二，对话数据由两个特定大语言模型生成；不同模型可能表现出不同的认知行为。第三，本文未报告框架的受控实验评估。量化验证框架所需的德尔菲研究、专家调查与AI输出质量实验，已在结论中列为未来工作。

5. Case Study

5.1 Case 1: yarrow (*Achillea millefolium*)

5.1.1 Background

Yarrow is the common English name for a Eurasian flowering plant of the genus *Achillea*, species *millefolium*, family Asteraceae. In scientific nomenclature, the Latin binomial

Achillea millefolium L. is unambiguous and is governed by the International Code of Nomenclature for algae, fungi, and plants. In English, "yarrow" is the standard common name. In Chinese, however, the situation is more complex. The plant has been known in Chinese botanical literature under several renderings, and different authoritative sources do not agree on which rendering should be treated as the [standard](#) Chinese name.

5.1.2 The conflict

During the dictionary project, the question "How should [yarrow](#) be translated into Chinese?" triggered a multi-round dialogue in which the AI, when prompted, produced several candidate renderings:

- 蓍草 (shīcǎo): a name with deep historical roots in Chinese, associated since antiquity with the plant used for divination stalks in the *I Ching*;
- 西洋蓍草 (xīyáng shīcǎo, literally "Western yarrow"): a name used in some modern Chinese floras and pharmacological references to distinguish the introduced Western species from the native Chinese *Achillea* species;
- 欧蓍 (ōushī, literally "European yarrow"): a shorter form used in some pharmaceutical and horticultural contexts;
- 千叶蓍 (qiānyè shī, literally "thousand-leaf yarrow"): a name derived from the species epithet *millefolium* ("thousand-leaved").

The AI, when first asked, produced a single confident answer: "蓍草." When pressed, it acknowledged that other renderings existed, but did not reliably identify which source supported which rendering, and it at one point attributed a particular rendering to a source that did not, in fact, use it—a classic LLM hallucination pattern.

5.1.3 Why this is a governance problem, not just a translation problem

At first glance, this looks like a routine dictionary question: pick the best equivalent. But closer analysis reveals that the question cannot be answered by retrieval alone, for three reasons.

First, the sources themselves disagree. The Chinese [Flora of China](#) (the English-language edition) and various Chinese provincial floras use different renderings. Pharmacopoeia references use still others. Bilingual dictionaries differ among themselves. There is no single "correct" rendering that can be looked up.

Second, the renderings are not mere synonyms; they encode different semantic and historical claims. "蓍草" carries the full weight of the classical Chinese association with divination and medicine. "西洋蓍草" explicitly marks the plant as an introduced species.

"千叶蓍" is a calque on the Latin epithet. Choosing among these is not a neutral lexical choice; it is a choice about which **semantic frame** the Chinese user will encounter.

Third, the AI's behaviour in this case is revealing. It produced a fluent answer, cited a plausible-sounding authority when pressed, and only revealed the existence of alternatives when directly challenged. This is precisely the pattern that makes prompt engineering insufficient: a well-crafted prompt ("Please consult authoritative bilingual dictionaries") will produce a better **first draft**, but it will not create a mechanism for adjudicating the conflict that remains.

5.1.4 Resolution through governance

The resolution reached in the project was not to choose one rendering at the expense of the others, but to record all of them with their sources and domains, and to designate one as the **standard name** on the basis of an explicit evaluation. Specifically:

- **Standard Chinese name (标准名):** 蓍草, on the grounds that it is the most widely recognized Chinese common name in general dictionaries and has the strongest historical warrant;
- **Aliases (别名):** 西洋蓍草, 欧蓍, 千叶蓍, each tagged with the domain in which it is typically used (pharmacology, horticulture, academic botany respectively);
- **Scientific name (学名):** *Achillea millefolium* L., recorded for unambiguous cross-language reference;
- **Source attribution (来源依据):** each rendering is tagged with the authoritative source(s) that support it, so that future reviewers can trace the choice.

This resolution is not "the correct translation of **yarrow**." It is a **governed** decision: explicit, attributable, revisable, and transparent about the alternatives. It is precisely the kind of decision that a management-only approach (store the approved term, retrieve it) cannot produce, because the existence of multiple legitimate alternatives is the point, not a failure to be resolved by better retrieval.

5.2 Case 2: trunk

5.2.1 Background

Trunk is a common English word with at least four major senses: (a) the thick wooden stem of a tree; (b) the human or animal torso; (c) the elongated muscular proboscis of an elephant; and (d) a large, sturdy rectangular travelling case. Each sense has a different Chinese equivalent, and these equivalents are not interchangeable.

5.2.2 The problem

In the early stages of the dictionary project, an initial AI-assisted draft proposed a single Chinese equivalent for **trunk** at the headword level. When this was reviewed, it became clear that a headword-level binding would produce systematic errors: a user looking up "elephant's trunk" would be offered the equivalent for "tree trunk," and a user looking up "pack a trunk" would be offered the equivalent for "torso."

This is not a problem of **authority** (the sources largely agree on the equivalent for each sense). It is a problem of **granularity**. The unit of terminological matching in classical TM is often the term as stored in a term base. But for polysemous common words, the stored unit must be the **sense**, not the word.

5.2.3 The structural insight

The **trunk** case forces a structural conclusion: matching must move from **word-level** (English word → Chinese word) to **sense-level** (English sense → Chinese equivalent). This is a well-known principle in lexicography, but it has not been systematically incorporated into the design of AI translation systems, which typically operate at the segment level and rely on contextual disambiguation rather than explicit sense inventory.

The relevant Chinese equivalents, by sense, are roughly:

- **Tree trunk** (树干, shùgàn): the woody main stem of a tree;
- **Torso** (躯干, qūgàn): the human or animal body excluding the head and limbs;
- **Elephant's trunk** (象鼻, xiàngbí): the proboscis of an elephant;
- **Travelling trunk** (行李箱 / 皮箱, xínglǐxiāng / píxiāng): a large rigid case for travel.

The project's resolution was to record **trunk** as a polysemous headword with four separately defined senses, each with its own Chinese equivalent, its own illustrative examples, and its own domain tags. Critically, this structure also clarifies the distinction—central to bilingual lexicography—between a true **translation equivalent** and a mere **explanatory gloss**. For example, a dictionary entry might explain that "torso" and "trunk" overlap in some but not all contexts; the Chinese equivalent for the torso sense is a genuine equivalent, while a longer explanatory gloss might be needed to convey the range of the word.

5.2.4 Why this is also a governance problem

At first glance, the **trunk** case looks like a routine lexicographical matter: disambiguate senses. But in the GenAI context, it becomes a governance problem for two reasons.

First, LLMs are, in practice, remarkably good at **contextual disambiguation** at the moment of translation: given a sentence, they can usually pick the right sense. But this implicit disambiguation is not **explicit** in the dictionary. The user of a bilingual dictionary needs to know that the four senses exist and are separately lexicalized in Chinese. A governance system must make this structure explicit, not rely on the model's implicit choices.

Second, the boundary between senses is itself contestable. Should "trunk" in the sense of "the luggage compartment of a car" (a fifth sense, used in American English) be recorded separately? Should "trunk" in the sense of "a long box in a studio" be conflated with the travelling-case sense? These are not facts to be retrieved; they are decisions to be governed.

5.3 Cross-case synthesis

The two cases together reveal the two fundamental challenges of terminology governance in the GenAI era:

- 1. **Authority challenge (illustrated by yarrow):** When authoritative sources disagree, how do we make a transparent, traceable, revisable decision?
- 2. **Granularity challenge (illustrated by trunk):** When a word is polysemous, at what level of analysis do we match, and how do we make the sense structure explicit rather than implicit?

The five-layer framework developed in Section 6 is designed to address both challenges simultaneously.

 **Cross-reference** 【← Governance Plan §6: Dispute Resolution and Escalation Mechanism】

五、案例研究

5.1 案例一：yarrow (Achillea millefolium, 蓍草)

5.1.1 背景

Yarrow 是菊科蓍草属 (Achillea) 千叶蓍种 (millefolium) 一种欧亚开花植物的通用英文名称。在科学命名中，拉丁双名法名称 Achillea millefolium L. 是无歧义的，受《国际藻类、真菌和植物命名法规》管辖。在英语中，"yarrow"是标准通用名。然而在汉语中，情况更为复杂。该植物在

中文植物学文献中以若干译法为人所知，而不同权威来源对哪种译法应被视为标准中文名并无一致意见。

5.1.2 冲突

在词典项目期间，"yarrow 应如何译为中文？"这一问题触发了多轮对话，AI在被提示时产生了若干候选译法：

- 蓍草 (shǐcǎo)：一个在汉语中具有深厚历史根基的名称，自古与《易经》中用于占卜的蓍草茎相关联；
- 西洋蓍草 (xīyáng shǐcǎo，字面义"西方蓍草")：一些现代中文植物志与药理学参考文献中使用的名称，用以区分引入的西方物种与中国本土的蓍草属物种；
- 欧蓍 (ōushī，字面义"欧洲蓍草")：在某些药学与园艺语境中使用的较短形式；
- 千叶蓍 (qiānyè shī，字面义"千叶蓍草")：一个源自种加词 *millefolium* ("千叶的") 的名称。

AI在首次被问及时给出了一个自信的单一答案："蓍草。"当被追问时，它承认存在其他译法，但未能可靠识别哪个来源支持哪个译法，并且一度将某一译法归因于一个实际上并未使用该译法的来源——这是一种典型的大语言模型幻觉模式。

5.1.3 为何这是一个治理问题，而非仅仅是翻译问题

乍看之下，这像是一个常规的词典问题：选择最佳对等词。但更细致的分析表明，这一问题不能仅靠检索来回答，原因有三。

第一，来源本身存在分歧。中文《中国植物志》（英文版）与各省级植物志使用不同译法。药典参考文献使用另外的译法。双语词典之间也互不相同。不存在一个可以查阅的单一"正确"译法。

第二，这些译法并非单纯的同义词；它们编码了不同的语义与历史主张。"蓍草"承载着古典汉语中与占卜和医学相关的全部历史重量。"西洋蓍草"明确标示该植物为引入物种。"千叶蓍"是对拉丁种加词的仿译。在这些译法之间做出选择并非中性的词汇选择；它是关于中国用户将遭遇何种语义框架的选择。

第三，AI在本案中的行为具有揭示性。它产生了一个流畅的答案，在被追问时引用了一个听起来合理的权威来源，并且仅在被直接质疑时才揭示其他备选译法的存在。这正是使提示词工程不足的模式：一个精心设计的提示词（"请查阅权威双语词典"）会产生更好的初稿，但它不会为仍然存在的冲突创建裁决机制。

5.1.4 通过治理达成的解决

项目中达成的解决方案不是以牺牲其他译法为代价选择某一译法，而是将所有译法连同其来源与领域一并记录，并在明确评价的基础上指定一个为标准名。具体而言：

- **标准中文名（标准名）**： 蓍草，理由是它是通用词典中最广为人知的中文通用名，且具有最强的历史依据；
- **别名（别名）**： 西洋蓍草、欧蓍、千叶蓍，每一个都标注其通常使用的领域（分别为药理学、园艺学、植物学学术）；
- **学名（学名）**： *Achillea millefolium* L.，记录以备无歧义的跨语言参照；
- **来源依据（来源依据）**： 每个译法都标注支持它的权威来源，以便未来审校者追溯选择。

这一解决并非"yarrow 的正确翻译"。它是一个受治理的决策：明确的、可归因的、可修订的，并且对备选方案保持透明。这恰恰是仅靠管理方法（存储已批准术语、检索它）无法产生的决策类型，因为多个合法备选方案的存在本身就是要点，而非通过更好的检索来解决的失败。

5.2 案例二：trunk（树干/躯干/象鼻/行李箱）

5.2.1 背景

Trunk 是一个常见英文词，至少有四个主要义项：（a）树的粗木质茎干；（b）人或动物的躯干；（c）大象伸长的肌肉质长鼻；以及（d）一个大型、坚固的长方形旅行箱。每个义项有不同的中文对等词，且这些对等词不可互换。

5.2.2 问题

在词典项目早期阶段，一份AI辅助初稿在词头层面为 trunk 提出了单一中文对等词。当对此进行审校时，词头级绑定会产生系统性错误这一点变得清晰：查阅"elephant's trunk"（象鼻）的用户将获得"tree trunk"（树干）的对等词，而查阅"pack a trunk"（装箱）的用户将获得"torso"（躯干）的对等词。

这不是一个权威问题（各来源对每个义项的对等词基本一致）。这是一个粒度问题。经典TM中术语匹配的单位通常是术语库中存储的术语。但对于多义常用词，存储的单位必须是义项，而非词。

5.2.3 结构性洞见

trunk 案例迫使我们得出一个结构性结论：匹配必须从词级（英文词→中文词）转向义项级（英文义项→中文对等词）。这是词典学中一个众所周知的原则，但它尚未被系统地纳入AI翻译系统的设计，后者通常在句段层面操作，依赖语境消歧而非显式义项清单。

按义项划分的相关中文对等词大致如下：

- **树干（树干, shùgàn）**： 树的木质主干；
- **躯干（躯干, qūgàn）**： 人或动物除去头部与四肢的身体部分；

- 象鼻 (象鼻, xiàngbí) : 大象的长鼻;
- 行李箱/皮箱 (行李箱/皮箱, xínglixiāng / píxiāng) : 用于旅行的大型硬质箱。

项目的解决方案是将 **trunk** 记录为一个多义词头, 具有四个独立定义的义项, 每个义项有其自己的中文对等词、例证与领域标签。关键在于, 这一结构还厘清了双语词典编纂中处于核心地位的区分——真正的翻译对等词与单纯的解释性释义之间的区分。例如, 一个词典条目可能解释"torso"与"trunk"在某些语境中重叠但并非全部; 躯干义项的中文对等词是真正的对等词, 而要传达该词的语义范围则可能需要更长的解释性释义。

5.2.4 为何这也是一个治理问题

乍看之下, **trunk** 案例像是一个常规的词典学事务: 义项消歧。但在生成式AI语境中, 它之所以成为治理问题有两个原因。

第一, 大语言模型在实践中非常擅长在翻译时刻进行语境消歧: 给定一个句子, 它们通常能选择正确的义项。但这种隐含的消歧在词典中并非显式的。双语词典的用户需要知道四个义项存在且在汉语中分别词汇化。治理系统必须使这一结构显式化, 而非依赖模型的隐含选择。

第二, 义项之间的边界本身就是可争议的。"trunk"在"汽车行李箱舱"义项(美式英语中的第五个义项)中是否应单独记录? "trunk"在"录音棚中的长箱"义项中是否应与旅行箱义项合并? 这些不是待检索的事实; 它们是待治理的决策。

5.3 跨案例综合

两个案例共同揭示了生成式AI时代术语治理的两大根本挑战:

1. 权威挑战 (以 **yarrow** 为例): 当权威来源存在分歧时, 我们如何做出透明、可追溯、可修订的决策?
2. 粒度挑战 (以 **trunk** 为例): 当一个词是多义的, 我们应在什么分析层面进行匹配, 以及如何使义项结构显式而非隐含?

第六节提出的五层框架旨在同时应对这两大挑战。

 关联方案 【←方案 § 6 争议处理与升级机制】

6. The Proposed Terminology Governance Framework (TGF)

We now introduce the Terminology Governance Framework (TGF), a five-layer model that operationalizes the concept of terminology governance in the context of generative

AI-assisted translation. Each layer addresses a specific challenge identified in the case study and grounded in the theoretical perspectives of Section 3.

6.1 Layer 1: Source Authority Governance

The first layer addresses the question of *who/what is authoritative*. The central insight, illustrated by *yarrow*, is that not all sources are equal, and that the AI should not be free to privilege one source over another without an explicit ranking. The framework proposes a **source authority hierarchy** with five tiers:

Tier	Source type	Examples	Authority weight
1	Professional standards and databases	ISO standards, ICN nomenclature, national terminology standards	Highest
2	Professional institutional resources	Subject-society glossaries, research-institution corpora	High
3	Authoritative bilingual dictionaries	Major published bilingual dictionaries with editorial boards	Medium-high
4	General (monolingual or learner) dictionaries	General-purpose dictionaries	Medium
5	AI-generated candidates	LLM-proposed renderings with no independent source	Low

This hierarchy is not a rigid rule; it is a *default weighting* that can be overridden when the specific context demands it. But it solves the *yarrow* problem by ensuring that, when the AI proposes a candidate, it is required to identify the source tier from which the candidate is drawn, and that conflicts between tiers are flagged for human review rather than silently resolved.

Importantly, the hierarchy also encodes a critical principle: **AI-generated candidates are the lowest tier, not the highest**. This is a direct response to the observation that LLMs can produce fluent but ungrounded renderings, and that prompt engineering alone cannot prevent this. The AI is a candidate generator, not an authority.

 **Cross-reference** 【← Governance Plan § 5.2: Source Authority Hierarchy Table】

6.2 Layer 2: Semantic Alignment Governance

The second layer addresses the question of at what granularity the match should be made. The central insight, illustrated by [trunk](#), is that matching must move from the word level to the sense level. The framework proposes the following principles:

1. **Sense-level matching, not word-level binding.** For any polysemous headword, the dictionary record must explicitly enumerate senses, each with its own target-language equivalent.
2. **Explicit sense delimitation.** Each sense must be accompanied by a brief definition or delimitation that makes the sense boundary visible to both the human reviewer and the downstream translation system.
3. **Domain tagging.** Each sense is tagged with the domain in which it is typically used (botany, medicine, everyday language, etc.), enabling domain-aware retrieval.
4. **Distinction between equivalent and gloss.** The dictionary record must distinguish between a [true translation equivalent](#) (a target-language word that, in the relevant sense, replaces the source word) and an [explanatory gloss](#) (a phrase that helps the user understand the source word but is not a conventional target equivalent).

This layer operationalizes the socioterminological insight that words are not stable units but context-dependent, community-bound lexical items. It also operationalizes the human-centred AI principle of explicability: the sense structure is made explicit rather than left implicit in the model's contextual choices.

 **Cross-reference** 【← Governance Plan § 4, Rule 2: Polysemous Words Must Be Handled by Sense】

6.3 Layer 3: Terminology Evaluation Mechanism

The third layer addresses the question of [how a candidate term is evaluated](#). The central insight is that source authority alone is insufficient: even within a single authoritative source, multiple candidates may exist, and the choice among them requires a multi-dimensional evaluation. The framework proposes five evaluation dimensions:

1. **Authority (来源权威性):** The source tier from which the candidate is drawn (see Layer 1).
2. **Semantic Match (语义匹配):** The degree to which the candidate captures the relevant sense of the source term, without over- or under-translation.

3. **Domain Acceptance** (领域接受度): The degree to which the candidate is actually used in the relevant target-language community of practice (e.g., by pharmacists, botanists, or lawyers).
4. **Usage Stability** (使用稳定性): The degree to which the candidate has been used consistently over time, as opposed to being a recent coinage or a passing fashion.
5. **Accessibility** (用户可理解性): The degree to which the candidate is comprehensible to the intended end-user (e.g., a general dictionary user vs. a subject specialist).

These five dimensions are combined into a composite evaluation that produces a recommended primary term and a set of secondary aliases, each tagged with the dimensions on which it scores well. This is precisely the procedure used to resolve the yarrow case: 蓍草 was selected as the primary term on the basis of its high authority, high usage stability, and high accessibility, while the other renderings were retained as domain-tagged aliases.

The evaluation mechanism is explicitly designed to be human-augmented: the AI can compute preliminary scores on each dimension (e.g., by counting frequency in a reference corpus), but the final decision is made by the human expert, who can weigh the dimensions contextually.

 **Cross-reference** 【← Governance Plan § 3: Standard Workflow SOP】

6.4 Layer 4: Human–AI Decision Collaboration

The fourth layer addresses the question of who decides. The central insight, grounded in human-centred AI theory (Floridi et al., 2018), is that the AI and the human expert should have complementary roles, not overlapping or competing ones. The framework proposes the following division of labour:

AI responsibilities (candidate generation and evidence assembly):

- Proposing multiple candidate renderings for a given source term;
- Retrieving and organizing evidence from authoritative sources;
- Identifying conflicts between sources and flagging them;
- Preparing preliminary evaluation scores on the five dimensions;
- Suggesting sense delimitations for polysemous headwords.

Human responsibilities (authoritative decision):

- Confirming or rejecting the concept behind the term;
- Choosing the primary term from among the AI's candidates;
- Resolving conflicts that the AI has flagged;
- Approving the sense structure for polysemous headwords;
- Taking final accountability for the published term.

This division of labour operationalizes the principle of **meaningful human control**. The AI is not a black box that outputs "the answer"; it is an assistant that assembles evidence and proposes candidates. The human is not a mere approver who rubber-stamps AI output; they are the decision-maker who exercises expert judgment. This is critical for addressing the automation bias problem: by explicitly assigning conflict identification to the AI and conflict resolution to the human, the framework builds in a structural incentive for the human to scrutinize the AI's output.

 **Cross-reference** 【← Governance Plan § 2: Organizational Roles and RACI Matrix】

6.5 Layer 5: Terminology Lifecycle Management

The fifth layer addresses the question of how a term evolves over time. The central insight is that a term is not a static object but a living record that changes as new sources appear, as usage shifts, and as errors are discovered. The framework proposes a six-stage lifecycle:

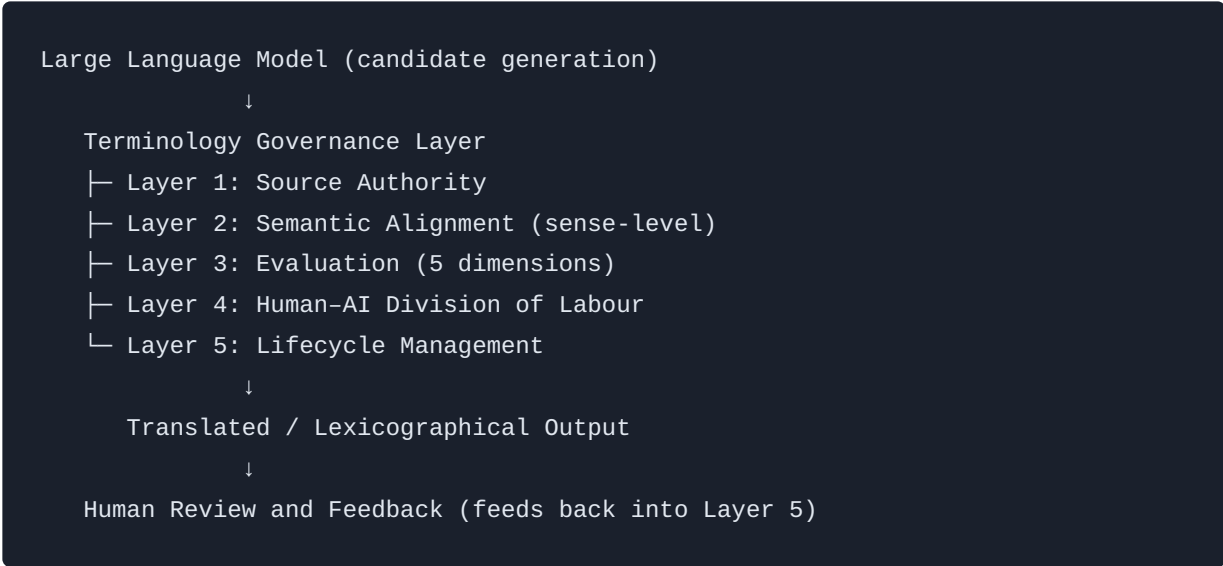
1. **Creation (创建)**: A new candidate term is entered, with its sources and preliminary evaluation.
2. **Review (审核)**: The candidate is reviewed by the human expert against the five dimensions.
3. **Publication (发布)**: The term is approved and entered into the term base, with full source attribution.
4. **Revision (修订)**: The term is updated when new evidence emerges (e.g., a new authoritative source, a shift in usage).
5. **Retirement (废弃)**: The term is retired when it is no longer appropriate (e.g., when a standard nomenclature changes).
6. **Archiving (归档)**: The retired term is retained in the archive for audit and traceability, with a record of when and why it was retired.

This layer operationalizes the AI governance principles of accountability and traceability. Every terminological decision is a record: it has a date, a decision-maker, a set of sources, and a rationale. It can be revisited, revised, or retired, and the history is preserved. This is a direct response to the GenAI context, in which the "decision" is distributed across a model version, a prompt, a retrieval context, and a human reviewer: only an explicit lifecycle record can make such a distributed decision auditable.

 **Cross-reference** 【← Governance Plan § 9: Version Management and Archiving】

6.6 The Terminology Governance Layer as a system component

Taken together, the five layers constitute what we call a **Terminology Governance Layer (TGL)**: an explicit mediating component that sits between the LLM and the translated output, as illustrated schematically below.



The central claim of this paper is that **prompt engineering cannot substitute for this layer**. A prompt that says "please consult authoritative dictionaries" is a substitute for the governance layer only in the most superficial sense: it produces a better first draft, but it does not create the mechanisms for adjudicating conflicts, for making sense structure explicit, for evaluating candidates, for dividing labour, or for tracking the lifecycle. Those mechanisms require an explicit, structured layer, not a prompt.

六、所提出的术语治理框架（TGF）

我们现在介绍术语治理框架（Terminology Governance Framework, TGF），这是一个五层模型，在生成式AI辅助翻译语境中操作化术语治理概念。每一层应对案例研究中识别出的特定挑战，并以第三节的理论视角为基础。

6.1 第一层：来源权威治理

第一层解决谁/什么具有权威性的问题。以 *yarrow* 为例说明的核心洞见是：并非所有来源都同等重要，AI不应在没有明确排序的情况下自由地偏袒某一来源。框架提出了一个五层来源权威层级：

层级	来源类型	示例	权威权重
1	专业标准与数据库	ISO标准、ICN命名法规、国家术语标准	最高
2	专业机构资源	学科协会术语表、研究机构语料库	高
3	权威双语词典	有编辑委员会的主要出版双语词典	中高
4	通用（单语或学习者）词典	通用词典	中
5	AI生成候选	无独立来源的LLM提议译法	低

这一层级不是一条刚性规则；它是一个默认权重，在特定语境需要时可以被覆盖。但它通过以下方式解决了 *yarrow* 问题：当AI提出候选时，要求它标识该候选来自哪个来源层级，并且层级之间的冲突被标记以供人类审校，而非被静默解决。

重要的是，该层级还编码了一个关键原则：**AI生成候选处于最低层级，而非最高层级**。这是对以下观察的直接回应：大语言模型可以产生流畅但无依据的译法，而仅凭提示词工程无法防止这一点。AI是候选生成者，而非权威。

 **关联方案** 【←方案 § 5.2 来源权威等级表】

6.2 第二层：语义对齐治理

第二层解决匹配应在什么粒度上进行的问题。以 *trunk* 为例说明的核心洞见是：匹配必须从词级转向义项级。框架提出以下原则：

1. **义项级匹配，而非词级绑定。** 对于任何多义词头，词典记录必须显式枚举义项，每个义项有其自己的目标语对等词。
2. **显式义项划定。** 每个义项必须附有简要定义或划定，使义项边界对人类审校者与下游翻译系统均可见。
3. **领域标签。** 每个义项标注其通常使用的领域（植物学、医学、日常语言等），实现领域感知检索。

4. 对等词与释义的区分。 词典记录必须区分真正的翻译对等词（目标语中在相关义项下替代源语词的词）与解释性释义（帮助用户理解源语词但非常规目标语对等词的短语）。

这一层操作化了社会术语学的洞见：词不是稳定单位，而是依赖语境、受共同体约束的词汇项。它还操作化了以人为本AI的可解释性原则：义项结构被显式化，而非隐含在模型的语境选择中。

 关联方案 【←方案 § 4 术语决策规则第二条】

6.3 第三层：术语评价机制

第三层解决候选术语如何被评价的问题。核心洞见是：仅凭来源权威是不够的；即使在单一权威来源内，也可能存在多个候选，而在它们之间做出选择需要多维度评价。框架提出五个评价维度：

- 1. 权威（来源权威性）： 候选来自的来源层级（见第一层）。
- 2. 语义匹配（语义匹配）： 候选捕捉源术语相关义项的程度，无过度翻译或欠翻译。
- 3. 领域接受度（领域接受度）： 候选在相关目标语实践共同体中（例如药剂师、植物学家或律师）实际使用的程度。
- 4. 使用稳定性（使用稳定性）： 候选随时间一致使用的程度，而非新近造词或短暂流行。
- 5. 用户可理解性（用户可理解性）： 候选对目标终端用户（例如普通词典用户 vs. 学科专家）可理解的程度。

这五个维度被组合为一个综合评价，产生一个推荐主术语与一组次要看别名，每个都标注其得分较高的维度。这恰恰是解决 [yarrow](#) 案例所使用的程序：蓍草因其高权威性、高使用稳定性与高可理解性被选为主术语，其他译法则作为领域标注的别名保留。

评价机制明确被设计为人类增强的：AI可以在每个维度上计算初步评分（例如通过在参考语料库中计数频率），但最终决策由人类专家做出，后者可以在语境中权衡各维度。

 关联方案 【←方案 § 3 标准工作流程SOP】

6.4 第四层：人机协同决策

第四层解决谁决策的问题。以以人为本AI理论（Floridi et al., 2018）为基础的核心洞见是：AI与人类专家应具有互补角色，而非重叠或竞争角色。框架提出以下劳动分工：

AI职责（候选生成与证据汇总）：

- 为给定源术语提出多个候选译法；
- 从权威来源检索并组织证据；

- 识别来源之间的冲突并标记；
- 准备五个维度的初步评价评分；
- 为多义词头建议义项划定。

人类职责（权威决策）：

- 确认或否定术语背后的概念；
- 从AI候选中选择主术语；
- 解决AI标记的冲突；
- 批准多义词头的义项结构；
- 对已发布术语承担最终问责。

这一劳动分工操作化了有意义的人类控制原则。AI不是输出"答案"的黑箱；它是一个汇总证据、提出候选的助手。人类不是简单盖章批准AI输出的审批者；他们是行使专家判断的决策者。这对于解决自动化偏差问题至关重要：通过将冲突识别显式分配给AI、冲突解决分配给人类，框架在结构上激励人类审视AI输出。

 **关联方案** 【←方案 § 2 组织架构与角色职责】

6.5 第五层：术语全生命周期管理

第五层解决术语如何随时间演变的问题。核心洞见是：术语不是静态客体，而是一份活的记录，随新来源出现、用法转变与错误发现而变化。框架提出六阶段生命周期：

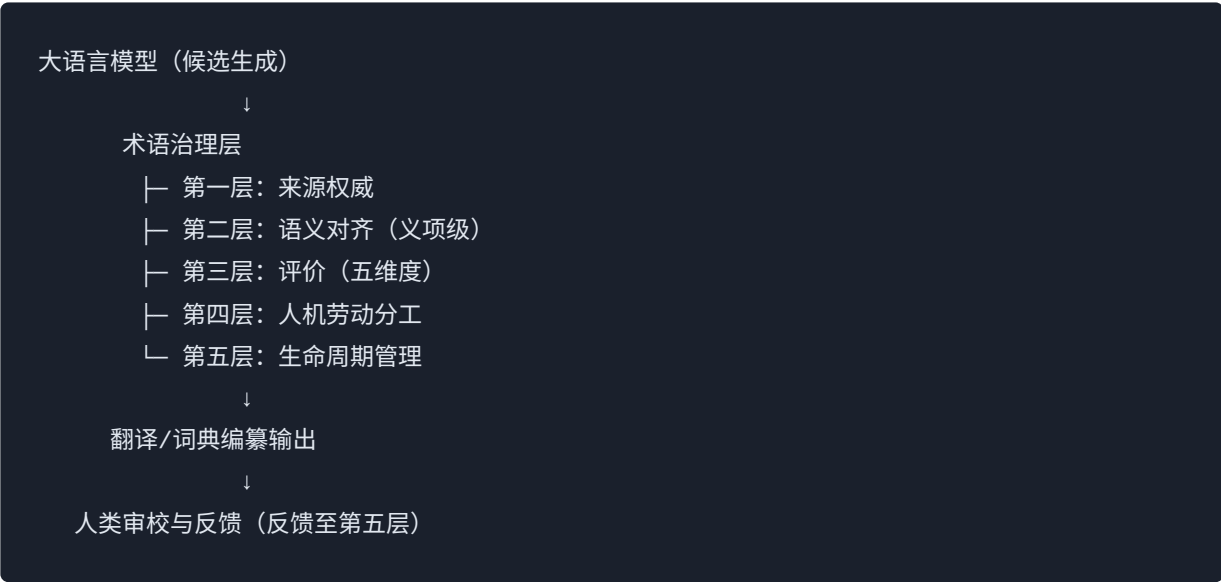
1. **创建（创建）**：新候选术语被录入，附带来源与初步评价。
2. **审核（审核）**：候选由人类专家按五个维度审核。
3. **发布（发布）**：术语获得批准并录入术语库，附完整来源归因。
4. **修订（修订）**：当新证据出现时（例如新的权威来源、用法转变），术语被更新。
5. **废弃（废弃）**：当术语不再适当时（例如标准命名变更），术语被废止。
6. **归档（归档）**：已废止术语保留在档案中以供审计与可追溯，记录废止时间与原因。

这一层操作化了AI治理的问责与可追溯原则。每一个术语决策都是一条记录：它有日期、决策者、一组来源与理由。它可以被重新审视、修订或废止，历史被保存。这是对生成式AI语境的直接回应，其中"决策"分布在模型版本、提示词、检索语境与人类审校者之间：只有显式的生命周期记录才能使如此分布式的决策可审计。

 关联方案 【←方案 § 9 版本管理与归档规范】

6.6 术语治理层作为系统组件

五层合在一起构成了我们所称的术语治理层（Terminology Governance Layer, TGL）：一个显式的中介组件，位于大语言模型与翻译输出之间，示意如下。



本文的核心主张是：提示词工程无法替代这一层。 一个说"请查阅权威词典"的提示词仅在最表面的意义上是治理层的替代物：它产生更好的初稿，但它不创建裁决冲突、使义项结构显式、评价候选、分工或追踪生命周期的机制。这些机制需要一个显式的、结构化的层，而非一条提示词。

7. Discussion

7.1 Why "governance" rather than "management"?

The distinction between management and governance is not merely terminological. In classical terminology management, the object of management is the term base: a relatively stable, human-curated resource. The manager's task is to keep it consistent, up to date, and retrievable. In terminology governance, the object of governance is the [terminological decision-making process](#) itself, which now involves a generative AI that produces candidates at scale, a set of heterogeneous sources that disagree, a human expert who must adjudicate, and a lifecycle that extends over years.

The shift from management to governance parallels similar shifts in other domains. In corporate information management, the shift from "records management" to "information governance" reflected the recognition that, once information is distributed

across digital systems, the task is no longer to store records but to steer the processes that produce, use, and dispose of them. In AI ethics, the shift from "technical safety" to "AI governance" reflected the recognition that safety cannot be achieved by technical fixes alone but requires institutional, procedural, and accountability mechanisms. The argument here is analogous: in the GenAI context, terminology work cannot be managed as if it were a stable resource; it must be governed as a dynamic, distributed knowledge asset.

7.2 Prompt engineering is not governance

One of the most common responses to the problems identified in this paper is to argue that better prompts will solve them. We argue that this is a category error. Prompt engineering is a technique for eliciting better output from a given model; it is, by nature, ephemeral and ad hoc. A prompt is not a source authority hierarchy; it does not enforce sense-level matching; it does not implement a five-dimension evaluation; it does not codify a human–AI division of labour; and it does not track a lifecycle. A prompt can invite the model to behave in a governed way, but it cannot enforce that behaviour, and it cannot create the auditable record that governance requires. The TGF is a structural complement to, not a substitute for, good prompt engineering.

7.3 The role of the human expert in the GenAI era

The case studies and the framework together imply a re-theorization of the human expert's role. In classical TM, the terminologist is a curator: they collect, define, and record terms. In the GenAI era, the terminologist remains necessary, but their role shifts from curator to governor: they set the authority hierarchy, they adjudicate conflicts, they evaluate candidates, and they own the lifecycle. This is not a diminution of the human role; it is an elevation of it. The expert's judgment is no longer a matter of looking up the right term; it is a matter of making and being accountable for a decision in a context of distributed, AI-generated uncertainty.

7.4 The specificity of English–Chinese bilingual terminology

The framework was developed in the context of English–Chinese bilingual work, and it is worth noting what is specific to this language pair. Chinese, as a writing system that does not mark word boundaries in the same way as alphabetic languages, and as a language with a long and continuous lexicographical tradition, presents particular challenges for sense-level matching. The distinction between a standard name (a formally designated Chinese name for a plant, animal, or mineral) and a descriptive translation (a gloss that explains the meaning) is particularly salient in Chinese, because the Chinese scientific nomenclature system has historically been more centralized and prescriptive than the

corresponding English system. This makes the *yarrow* case—where multiple standard names compete—more visible in Chinese than it might be in another language pair. We note this specificity as a boundary condition: the framework's transferability to other language pairs is an empirical question, but the underlying logic (authority, granularity, evaluation, division of labour, lifecycle) is general.

七、讨论

7.1 为何用"治理"而非"管理"?

管理与治理之间的区分不仅仅是术语性的。在经典术语管理中，管理的对象是术语库：一种相对稳定、人工策展的资源。管理者的任务是保持其一致性、时效性与可检索性。在术语治理中，治理的对象是术语决策过程本身，现在它涉及一个大规模生成候选的生成式AI、一组存在分歧的异质来源、一位必须裁决的人类专家，以及一个延续数年的生命周期。

从管理到治理的转变与其他领域的类似转变并行。在企业信息管理中，从"记录管理"到"信息治理"的转变反映了这样一种认识：一旦信息分布在数字系统中，任务不再是存储记录，而是引导产生、使用与处置记录的过程。在AI伦理中，从"技术安全"到"AI治理"的转变反映了这样一种认识：安全不能仅靠技术修复来实现，还需要制度、程序与问责机制。此处的论证是类似的：在生成式AI语境中，术语工作不能被当作稳定资源来管理；它必须被当作一种动态的、分布式的知识资产来治理。

7.2 提示词工程不是治理

对本文所识别问题的最常见回应之一，是主张更好的提示词将解决这些问题。我们认为这是一种范畴错误。提示词工程是一种从给定模型引出更好输出的技术；它本质上是短暂的与临时性的。一条提示词不是来源权威层级；它不强制执行义项级匹配；它不实施五维度评价；它不法典化人机劳动分工；它也不追踪生命周期。一条提示词可以邀请模型以受治理的方式行事，但它不能强制这种行为，也不能创建治理所要求的可审计记录。TGF是良好提示词工程的结构性补充，而非替代物。

7.3 人类专家在生成式AI时代的角色

案例研究与框架共同暗示了对人类专家角色的重新理论化。在经典TM中，术语学家是策展人：他们收集、定义并记录术语。在生成式AI时代，术语学家仍然是必要的，但他们的角色从策展人转向治理者：他们设定权威层级、裁决冲突、评价候选并拥有生命周期。这不是对人类角色的贬低；而是提升。专家的判断不再是查阅正确术语的问题；而是在分布式的、AI生成的不确定性语境中做出决策并为之负责的问题。

7.4 英汉双语术语的特殊性

框架是在英汉双语工作语境中发展起来的，值得指出这一语言对的特殊性。汉语作为一种不以拼音文字相同方式标记词边界的书写系统，以及作为一种具有悠久连续词典编纂传统的语言，对义项级

匹配提出了特殊挑战。标准名（为植物、动物或矿物正式指定的中文名）与描述性翻译（解释意义的释义）之间的区分在汉语中尤为显著，因为中国科学命名系统在历史上比相应的英语系统更为集中和规定。这使得 *yarrow* 案例——多个标准名竞争——在汉语中比在其他语言对中更为可见。我们将这一特殊性注意为边界条件：框架向其他语言对的可迁移性是一个实证问题，但底层逻辑（权威、粒度、评价、分工、生命周期）是普遍的。

8. Theoretical Contributions

This paper makes four theoretical contributions.

First, it introduces the conceptual distinction between Terminology Management and Terminology Governance. This distinction is not merely rhetorical; it identifies a real structural change in the object of terminological practice, driven by the arrival of generative AI. The paper articulates what the new object is (a dynamic, distributed knowledge asset) and what questions it raises (authority, granularity, evaluation, division of labour, lifecycle) that the older concept of management did not address.

Second, it integrates five hitherto separate theoretical perspectives—terminology management theory, socioterminology, translation technology studies, human-centred AI, and AI governance—into a single framework. Each perspective has been developed in isolation, and none has yet addressed the specific configuration that generative AI introduces. The TGF shows how these perspectives can be productively combined: the classical theory provides the conceptual vocabulary, socioterminology problematizes authority, translation technology studies situate the practice, human-centred AI provides the normative division of labour, and AI governance provides the accountability structure.

Third, it grounds the framework in two structurally distinct case studies—*yarrow* (authority conflict) and *trunk* (sense mismatch)—that illustrate the two fundamental challenges of GenAI-era terminology work. The cases show that the framework is not merely abstract; it corresponds to concrete, recurring problems that practitioners actually face.

Fourth, it reframes the role of the human expert in the GenAI translation ecosystem. The prevailing narrative in popular discussion is that AI will either replace translators or augment them in a simple productivity sense. The TGF implies a more nuanced picture: the human expert's role shifts from curator to governor, and the epistemic core of the role—judgment under uncertainty, accountability for decisions—becomes more, not less, important.

八、理论贡献

本文做出四项理论贡献。

第一，它引入了术语管理与术语治理之间的概念区分。这一区分不仅是修辞性的；它识别了由生成式AI到来驱动的术语实践客体的真实结构变化。本文阐述了新客体是什么（一种动态的、分布式的知识资产），以及它提出了哪些旧有管理概念未能处理的问题（权威、粒度、评价、分工、生命周期）。

第二，它将五个迄今为止相互分离的理论视角——术语管理理论、社会术语学、翻译技术研究、以人为本的AI与AI治理——整合为一个单一框架。每个视角都是孤立发展的，尚无一处处理过生成式AI引入的特定配置。TGF展示了这些视角如何被富有成效地组合：经典理论提供概念词汇，社会术语学将权威问题化，翻译技术研究定位实践，以人为本的AI提供规范性分工，AI治理提供问责结构。

第三，它将框架扎根于两个结构不同的案例研究——*yarrow*（权威冲突）与 *trunk*（义项不匹配）——说明生成式AI时代术语工作的两大根本挑战。案例表明，框架并非仅仅是抽象的；它对应于从业者实际面对的具体、反复出现的问题。

第四，它重新界定了人类专家在生成式AI翻译生态系统中的角色。大众讨论中流行的叙事是AI要么将取代译者，要么以简单的生产力意义增强译者。TGF暗示了一幅更为微妙的图景：人类专家的角色从策展人转向治理者，而该角色的认识论核心——不确定性下的判断、对决策的问责——变得更加而非更少重要。

9. Practical Implications

The framework has practical implications for several audiences.

For translation practitioners and post-editors, the TGF implies that terminological decisions should be made explicit, attributable, and revisable, rather than left to the implicit choices of the LLM. Practitioners should be aware that a fluent AI output may hide unresolved conflicts, and that part of their professional value lies in recognizing and adjudicating those conflicts.

For localization and translation-technology vendors, the TGF suggests that future CAT and translation-management systems should incorporate an explicit Terminology Governance Layer: a structured component that encodes the source authority hierarchy, enforces sense-level matching, implements the five-dimension evaluation, codifies the human-AI division of labour, and tracks the lifecycle. This is a concrete product-development direction that goes beyond the current paradigm of "integrate a term base into the CAT tool."

For bilingual lexicographers and dictionary publishers, the TGF supports the explicit separation of standard names from descriptive translations, the documentation of competing renderings with their sources, and the maintenance of a clear sense structure. It also suggests that AI-assisted dictionary projects should design their workflow around governance principles from the outset, rather than retrofitting governance onto a system that was designed for single-answer retrieval.

For AI researchers and model developers, the TGF implies that the terminological reliability of an LLM is not primarily a function of the model's raw capability but of the governance layer through which its output is filtered. This suggests that model evaluation should include terminological governance metrics (e.g., does the model correctly identify when sources conflict? does it hallucinate sources?) rather than only fluency and accuracy metrics.

For standards bodies and policymakers, the TGF is relevant to the emerging domain of AI governance for language technologies. It suggests that terminology is a paradigmatic case of a "knowledge asset" that requires explicit governance mechanisms, and that existing standards (e.g., ISO 704) may need to be extended to address the generative-AI context.

 **Cross-reference** 【← Governance Plan § 6: Dispute Resolution; Appendices A–C: Form Templates】

九、实践启示

框架对若干受众具有实践启示。

对翻译从业者与译后编辑者，TGF意味着术语决策应被显式化、可归因、可修订，而非留给大语言模型的隐含选择。从业者应意识到，流畅的AI输出可能隐藏未解决的冲突，而其专业价值的一部分在于识别并裁决这些冲突。

对本地化与翻译技术供应商，TGF建议未来的CAT与翻译管理系统应纳入一个显式的术语治理层：一个结构化组件，编码来源权威层级、强制执行义项级匹配、实施五维度评价、法典化人机劳动分工并追踪生命周期。这是一个具体的产品开发方向，超越了当前"将术语库集成到CAT工具中"的范式。

对双语词典编纂者与词典出版商，TGF支持标准名与描述性翻译的显式分离、竞争译法及其来源的文档化，以及清晰义项结构的维护。它还建议AI辅助词典项目应从一开始就围绕治理原则设计 workflow，而非将治理改造到一个为单一答案检索而设计的系统上。

对AI研究者与模型开发者，TGF意味着大语言模型的术语可靠性主要不是模型原始能力的函数，而是其输出所经过的治理层的函数。这表明模型评价应包括术语治理指标（例如，模型是否正确识别来源何时冲突？它是否幻觉来源？），而不仅仅是流畅度与准确性指标。

对标准制定机构与政策制定者，TGF与语言技术AI治理这一新兴领域相关。它表明术语是一种"知识资产"的典型案例，需要显式的治理机制，而现有标准（例如ISO 704）可能需要扩展以应对生成式AI语境。

 **关联方案** 【←方案 § 6 争议处理；附录A—C 表单模板】

10. Conclusion, Limitations, and Future Research

10.1 Conclusion

This paper has argued that the arrival of generative AI requires a conceptual upgrade from Terminology Management to Terminology Governance. Management, in the classical tradition, treats terminology as a stable resource to be stored, retrieved, and kept consistent. Governance, as articulated here, treats terminology as a dynamic, distributed knowledge asset whose authority, granularity, evaluation, human–AI division of labour, and lifecycle must be explicitly steered. The five-layer TGF—Source Authority, Semantic Alignment, Evaluation, Human–AI Collaboration, and Lifecycle Management—operationalizes this upgrade, and the two case studies (yarrow and trunk) show that the framework corresponds to real, recurring, structurally distinct problems.

The central message is not that AI is unreliable or that humans should reject AI assistance. It is that the **form** of human–AI collaboration in terminology work must change: from a model in which the AI retrieves an approved term and the human checks it, to a model in which the AI generates candidates and assembles evidence, and the human governs the decision. Prompt engineering, however sophisticated, cannot substitute for this structural governance.

 **Cross-reference** 【← Governance Plan § 1– § 11 and All Appendices】

10.2 Limitations

The paper has several limitations that should be acknowledged. First, the empirical material is drawn from a single English–Chinese bilingual dictionary project; the framework's transferability to other language pairs, domains, and professional contexts is an open question. Second, the framework is proposed and illustrated but not yet experimentally validated: we do not report a controlled comparison of AI-assisted translation with and without the TGF, nor a Delphi study of expert consensus on the five evaluation dimensions. Third, the source authority hierarchy and the five evaluation dimensions are proposed on the basis of practical reasoning and theoretical synthesis; their empirical calibration (e.g., what weights should the five dimensions receive? how should the source tiers be operationalized in specific domains?) requires further work. Fourth, the positionality of the authors as participants in the dictionary project is a reflexive consideration, even though the analysis has been conducted in good faith.

10.3 Future research agenda

We identify four directions for future research.

1. **Empirical validation of the TGF.** A Delphi study with subject-matter experts, terminologists, and professional translators could refine the five evaluation dimensions and their relative weights. A controlled experimental study could compare terminological consistency, reliability, and user satisfaction in AI-assisted translation with and without the governance layer.
2. **Cross-language and cross-domain transfer.** The framework should be tested in other language pairs (e.g., English–Spanish, English–Arabic) and in other domains (e.g., legal translation, medical translation) where terminological governance has distinct stakes.
3. **Technical implementation of the TGL.** The five-layer framework should be operationalized as a concrete software component—an extension to existing CAT tools or a standalone governance layer—and its performance should be evaluated in real translation workflows.
4. **Extension of terminology standards.** The framework should be engaged with the ongoing revision of ISO 704 and related standards, with the goal of incorporating governance principles (authority, lifecycle, human–AI division of labour) into the next generation of terminology standards.

The study of terminology in the age of generative AI is not, ultimately, a study of words. It is a study of how human communities govern the knowledge that words encode. The TGF

is a first step toward a theoretically grounded, practically actionable, and ethically accountable account of that governance.

十、结论、局限与未来研究

10.1 结论

本文论证了生成式AI的到来要求从术语管理到术语治理的概念升级。经典传统中的管理将术语视为一种稳定资源，予以存储、检索并保持一致。如本文所阐述的治理则将术语视为一种动态的、分布式的知识资产，其权威、粒度、评价、人机分工与生命周期必须被明确引导。五层TGF——来源权威、语义对齐、评价、人机协同与生命周期管理——操作化了这一升级，两个案例研究（yarrow与trunk）表明框架对应于真实的、反复出现的、结构不同的问题。

核心信息并非AI不可靠，也非人类应拒绝AI辅助。它是术语工作中人机协同的形式必须改变：从AI检索已批准术语、人类检查的模式，转向AI生成候选并汇总证据、人类治理决策的模式。无论提示词工程多么精巧，都无法替代这一结构性治理。

 关联方案 【←方案 § 1— § 11及全部附录】

10.2 局限

本文有若干应承认的局限。第一，实证材料取自单一英汉双语词典项目；框架向其他语言对、领域与专业语境的可迁移性是一个开放问题。第二，框架被提出和说明但尚未经过实验验证：我们未报告有无TGF的AI辅助翻译的受控比较，也未报告关于五个评价维度专家共识的德尔菲研究。第三，来源权威层级与五个评价维度是基于实践推理与理论综合提出的；它们的实证校准（例如，五个维度应获得什么权重？来源层级如何在特定领域中操作化？）需要进一步工作。第四，作者作为词典项目参与者的立场性是一个反身性考量，即使分析是诚信进行的。

10.3 未来研究议程

我们确定了四个未来研究方向。

1. **TGF的实证验证。** 一项由学科专家、术语学家与专业译者参与的德尔菲研究可以精炼五个评价维度及其相对权重。一项受控实验研究可以比较有无治理层的AI辅助翻译在术语一致性、可靠性与用户满意度方面的差异。
2. **跨语言与跨领域迁移。** 框架应在其他语言对（例如英西、英阿）与其他领域（例如法律翻译、医学翻译）中检验，这些领域中术语治理具有不同的利害关系。
3. **TGL的技术实现。** 五层框架应被操作化为一个具体的软件组件——现有CAT工具的扩展或独立治理层——并在真实翻译工作流中评估其性能。

4. 术语标准的扩展。 框架应与ISO 704及相关标准的持续修订接轨，目标是将治理原则（权威、生命周期、人机分工）纳入下一代术语标准。

生成式AI时代对术语的研究，归根结底不是对词语的研究。它是对人类共同体如何治理词语所编码知识的研究。TGF是迈向一种有理论根基、实践可操作且伦理可问责的治理论述的第一步。

References

- Antia, B. E. (2015). *Terminology and language planning: With special reference to South Africa*. Palgrave Macmillan.
- Busch-Lauer, I. A. (2018). Terminology management in the age of big data. *Journal of Terminology Science*, 8, 1–18.
- Ensink, T., & Sauer, C. (2003). *Framing and discourse in language studies*. John Benjamins.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Taddeo, M. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- García, I. (2019). Machine translation and the profession. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 225–240). Routledge.
- Gaudin, F. (1993). *Socioterminologie: Des termes et des contextes dans les spécialités*. Presses Universitaires de Rouen.
- ISO 704:2022. (2022). *Terminology work and terminology science—Principles and methods*. International Organization for Standardization.
- Kageura, K. (2002). *The dynamics of terminology: A descriptive theory of term formation and term evolution*. Continuum.
- Kenny, D. (2019). Translation in the age of machine learning. *Translation Studies*, 12(2), 153–169.
- O'Hagan, M. (2013). *The manipulation of machine translation by professional translators: A sociological approach* (Unpublished doctoral dissertation). Dublin City University.
- O'Hagan, M. (2019). *The Routledge handbook of translation and technology*. Routledge.
- Pym, A. (2014). Translation skill in the age of neural machine translation. Summit on Translation, The Seoul International Conference on Translation.
- Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins.

Sanz Rosa, V. (2019). Terminology management systems in the era of cloud computing. *Localisation Focus*, 17, 45–58.

Temmerman, R. (2000). Towards new ways of terminology description: The sociocognitive-approach. John Benjamins.

Wright, S. E., & Budin, G. (Eds.). (1997). *Handbook of terminology management* (Vols. 1–2). John Benjamins.

Wüster, E. (1974). *The general theory of terminology and linguistics*. (Translation of 1968 German original). Infoterm.

Word count target: $\geq$ 8,000 words (excluding references).
