Modern Linguistics 现代语言学, 2026, 14(5), 85-91
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145377

Hans汉斯

合作原则视角下提示词对人工智能清洗海事英汉混合语料的影响研究

——以DeepSeek清洗海事事故报告语料为例

张熙田

上海海事大学外国语学院，上海

收稿日期：2026年4月4日；录用日期：2026年4月27日；发布日期：2026年5月8日

摘　要

本研究聚焦于提示词设计对人工智能大语言模型清洗海事英汉混合语料效能的影响。针对海事语料专业性强、英汉混杂的特点，研究设计了“简单指令”、“角色 + 指令”、“角色 + 指令 + 约束”及“角色 + 背景 + 指令 + 约束”四种提示词框架，以DeepSeek模型为实验工具，对收集的海事事故调查报告语料进行清洗对照实验。结果显示，“角色 + 指令 + 约束”型提示词能够产生最符合预期的清洗结果，在去除格式噪音的同时最大程度保持原文的专业内容与结构。研究进一步依据格赖斯(H.P. Grice)的会话合作原则对结果进行了理论阐释，指出最优提示词框架在信息量、真实性、相关性和表达方式上均满足了有效人机交互的准则。本研究为利用大语言模型高效处理垂直领域混合语料提供了可复用的提示词设计框架，对推动人工智能与语言学研究方法的结合具有参考价值。

关键词

语料清洗，提示词，海事话语，会话合作原则

The Impact of Prompts on AI-Assisted Corpus Cleaning in Maritime Domain from the Perspective of Cooperative Principle

—A Case Study of DeepSeek in Processing Maritime Accident Reports

Xitian Zhang

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: April 4, 2026; accepted: April 27, 2026; published: May 8, 2026

文章引用：张熙田. 合作原则视角下提示词对人工智能清洗海事英汉混合语料的影响研究[J]. 现代语言学, 2026, 14(5): 85-91. DOI: 10.12677/ml.2026.145377

Abstract

This study focuses on the impact of prompt design on the efficacy of large language models in cleaning English-Chinese mixed maritime corpora. In response to the highly specialized and linguistically hybrid nature of maritime texts, the research designs four prompt frameworks: "Simple Instruction," "Role + Instruction," "Role + Instruction + Constraint," and "Role + Context + Instruction + Constraint." Using the DeepSeek model as the experimental platform, a controlled cleaning experiment was conducted on a collected corpus of maritime accident investigation reports. The results indicate that the "Role + Instruction + Constraint" prompt yields the most desirable cleaning outcomes, effectively removing formatting noise while maximally preserving the original professional content and structure. Furthermore, the study provides a theoretical interpretation of the findings based on H.P. Grice's Cooperative Principle, suggesting that the optimal prompt framework satisfies the maxims of quantity, quality, relation, and manner essential for effective human-machine interaction. This research offers a reusable prompt design framework for efficiently processing domain-specific mixed corpora using large language models, contributing to the integration of artificial intelligence and linguistic research methodologies.

Keywords

Corpus Cleaning, Prompt, Maritime Discourse, Cooperative Principle

1. 引言

语料清洗是指对用于机器翻译或自然语言处理的原始语料进行筛选、规范化处理和优化的过程；目的是剔除低质量或不相关的文本数据，确保训练数据的准确性和可用性[1]。语料清洗还涉及去除偏见性表达、对齐校验和敏感内容过滤，以确保数据的公正性和中性[2]。在语言学研究中，语料清洗是重要的前置环节之一。

由于海事领域的语言数据作为承载行业知识、支撑技术创新的核心载体；因此对海事话语的研究意义重大，而研究海事领域的语言就离不开高效、合理地进行语料清洗。海事领域有着海量多语种文本数据(如航海日志、船舶通信记录、港口规章、海事法庭案例等)，这些语料具有专业性高、术语密集、英语与汉语混合的特点(如中文语料中可能会有经纬度、船的名称等英文出现)。高效清洗海事语料，对良好推动语言学与海事领域结合，推动海事专门语言和语言学与海事的跨学科发展意义重大。

近年来，随着人工智能的发展，大语言模型(LLMs)的快速迭代为专业领域语料清洗提供了全新技术路径，其强大的自然语言理解与生成能力，能够快速处理复杂文本、识别语义歧义、规范专业表达，大幅优化语料清洗的效率，减少语料清洗的成本。提示词作为影响大语言模型输出什么样结果的关键因素，其质量直接决定了模型的任务执行效果。大语言模型凭借其强大的语义理解和上下文生成能力，已在通用领域语料清洗中展现出潜力，但其在垂直领域的应用效能仍受限于提示词设计的科学性。且部分模型在处理英汉混合的专门化文本时容易出现把部分原文内容忽视，或者无法良好识别关键信息与噪音的现象。

因此，本研究聚焦于探讨不同提示词对人工智能大语言模型清洗海事英汉混合语料的结果有什么影

响，以期为利用人工智能清洗海事领域的英汉混合语料提供可复用的提示词设计框架，推动人工智能在海事与语言学领域的深度应用。

2. 文献综述

2.1. 早期的人工清洗语料时期

语料清洗作为自然语言处理与计算语言学的基础环节，其方法论演进深刻反映了语言学研究范式与技术能力的协同变迁。从早期人工校对到现代人工智能驱动的自动化流程，语料清洗不仅在效率上实现了质的飞跃，更在清洗质量、语义保真度和任务适配性方面达到了前所未有的高度。

早期语言学的研究方法不太成熟，研究者缺乏系统、现代的研究方法训练。国内很多语言学领域的文章使用“思辨性”方法(举例论证)，还有很多的研究“不依赖数据”[3]。在20世纪60年代至80年代初期，语料清洗主要采用人工校对方法，研究者基于语言学理论知识对文本进行逐字逐句地审校。这一时期语料清洗相关的代表性工作包括通过手工方式完成分词、词性标注和句法结构分析，确保语料的语法正确性和语言学规范性。这种清洗方法虽然能够保证相对较高的准确性，但存在成本高昂、处理效率低下、可扩展性差等显著缺陷，且不同研究者之间的主观判断差异可能导致标注标准不一致，影响研究结果的可复现性。

2.2. 程序化和半自动化的语料清洗时期

国外使用计算机对语言学研究方法的创新起步早，国外的语料库语言学发展经历了手工收集、计算机化(未标注、已标注)等阶段。国内起步较晚，起步于20世纪80年代(以JDEST学术英语语料库为标志)，但发展迅速[4]。在计算机广泛被应用到国内各领域的学术研究中后，语言学的研究方法也依托计算机技术进行了广泛创新。

随着计算机技术在20世纪80年代末至90年代的快速发展，语料清洗开始向程序化和半自动化方向转变。研究者们利用正则表达式处理基础文本标准化任务，如去除HTML标签、统一空白字符格式、规范标点符号使用等。同时，有限状态自动机和上下文无关文法等形式化方法被应用于更复杂的语言处理任务。1995年左右，统计自然语言处理方法的兴起为语料清洗带来了新的技术路径，研究者开始将概率模型和信息论技术融入清洗流程，通过统计特征识别异常数据模式。这一阶段的清洗工具如WordSmith和AntConc等检索软件被广泛应用于语料库语言学研究中，提供了相对标准化的清洗功能模块。

2.3. 深度学习与全自动地处理语料时期

进入21世纪，特别是2010年代深度学习技术的突破性进展，彻底改变了语料清洗的技术范式。基于Python的NLTK库为研究者提供了更加灵活和丰富的语料处理方法，能够利用统一的数据标准避免不同类型数据转换的麻烦，并借助Python生态系统中的众多第三方库弥补传统工具的功能局限[5]。在专业术语识别方面，研究者开发了基于统计特征的自动术语识别方法，通过分析词汇的分布特征和组合模式来区分专业术语与普通词汇[6]。对于双语平行语料的清洗，研究者提出了机器辅助和全自动两种方法，前者适用于小规模语料库，后者则针对大规模语料处理，通过词对齐和句子长度特征检测翻译错误[7]。早期研究以定性分析和个案研究为主，侧重于对典型语例进行深入的描写和阐释。近年来，随着大数据和计算语言学技术的发展，定量研究方法的应用日益广泛。研究者利用语料库语言学方法，对大规模网络文本进行词频统计、关键词分析、共现网络分析等，以揭示流行语的使用频率、传播路径和语义韵。知识图谱可视化技术被用于描绘特定研究领域(如社交媒体话语研究)的知识结构、热点主题演变和学术共同体分布，使得研究结论更具客观性和系统性。同时，问卷调查、深度访谈等社会语言学方法被用于

探究不同年龄、性别、地域、教育背景的网民对网络流行语的认知、态度和使用差异，丰富了研究的维度和深度。

在现代研究中，大语言模型的出现为语料清洗带来了重大的技术突破。大语言模型可以快速高效地完成语料清洗任务，但其输出结果受限于提示词的质量；对大语言模型辅助语言学研究的相关方法仍在探索中。

2.4. 提示词工程及其在特定 NLP 任务与 LLM 数据清洗中的应用

近年来，提示词工程已从经验性试探转向系统性研究。随着人工智能技术的发展，自然语言处理(NLP)和大规模语言模型(LLM)已成为研究热点。提示词工程在特定 NLP 任务中的应用主要体现在如何通过设计有效的提示词来改善模型的表现。

在现代研究中，针对特定 NLP 任务，提示词设计需高度适配任务特性。在数据清洗与预处理方面，利用 LLM 替代传统规则与小型模型已成为一种趋势。然而，现有工作多无法规范性运用 LLM，缺乏针对语料语言学特性(如方言变体、历时演变、语体差异，ESP 等)的深度优化。本研究高度关注此交叉缺口，旨在通过语言学驱动的提示词设计，促进人工智能在良好的提示词指引下较好完成语言学相关任务。

3. 研究方法

本研究的步骤分为以下几步。

3.1. 语料收集

在中华人民共和国海事局官网搜索海事事故调查报告(因海事事故调查报告是海事领域中典型的既含有汉语，又会有英文出现的语篇，所以本研究使用海事事故调查报告作为实验语料)，搜集不同类调查报告中最新的四篇，复制这四篇海事事故调查报告简介里正文以汉语为主但同时又有英文(如船的名字，经纬度等)出现的部分，将复制的内容粘贴在一个 txt 文档中，将此文档命名为“英汉混合语料”作为备用。收集语料的时间为 2026 年 3 月 15 日，收集好的语料共有 1718 个字符，空格换行等符号不做处理。

3.2. 设计提示词框架

Louie Giray 在其研究中提到角色设定为设计良好提示词的技巧，可用于辅助优化学术写作相关的输出内容[8]。但有时在执行专门化、多种语言混合的语料清洗任务时，即使输入的提示词包含了角色设定的表述，大语言模型仍然无法输出令人满意的结果(如在处理海事事故调查报告的语料中会把船的英文名等信息当作噪音)。在实际使用人工智能执行语言研究相关任务时，除了在提示词中加入角色设计外，还可通过在提示词中加入语料来源的背景，对输出结果的条件约束的指令来提高大语言模型清洗语料的质量。

依照以上思路，本研究设计了四种类型的提示词框架，第一种是“简单指令”型，即直接告诉人工智能大语言模型需要做什么；第二种是“角色 + 指令”型，即在提示词中先为人工智能大语言模型设立一个角色，随后告诉其需要做什么；第三种是“角色 + 指令 + 约束”型，即在提示词中先为人工智能大语言模型设立一个角色，随后告诉其需要做什么，再为其输出结果添加一定的约束条件。第四种是“角色 + 背景 + 指令 + 约束”型，即提示词中先为人工智能大语言模型设立一个角色，随后告诉其收到材料(如语料文本)的来源背景，需要做什么，再为其添加一定的约束条件。依照上述三种框架分别设计了四句提示词；简单指令型：“请你清洗这份语料。”角色 + 指令型：“假如你是一个资深的语言学研究者，请你对这份语料进行清洗。”角色 + 指令 + 约束型：“假如你是一个资深的语言学研究者，请你对这份语料进行清洗，使其符合语言学研究的标准。”角色 + 背景 + 指令 + 约束型：“假如你是一个资深

的语言学研究者，请你对这份来自中华人民共和国海事局的语料进行清洗，使其符合语言学研究的标准。”

3.3. 使用大语言模型清洗语料

在人工智能大语言模型中(本研究以 DeepSeek 为例，使用元宝 PC 端 V2.57.0 版本)建立四个新对话，把收集好的英汉混合语料上传到四个新对话的交互窗口，随后在四个对话框中分别输入设计好的四种提示词，打开“深度思考”选项(为了便于观察模型思考过程的不同)；各个对话中除提示词外其余设置均保持默认不变。

3.4. 结果评估与框架分析

对不同提示词产生的输出结果进行评估，使用语言学理论分析最优的提示词框架。

4. 研究结果

4.1. “简单指令型”的结果

在进行上述操作后，使用简单指令型提示词导致清洗后的语料丢失了大量信息，大语言模型在深度思考的过程中将“清洗”理解为了“整理文档内容，让它更清晰易读”。于是大语言模型分点归纳总结并改写了语料的关键信息，并且擅自去掉了原文中部分不完整的句子。显然，由于信息缺失，清洗后的语料背离原文，并不能满足语言学研究的要求。

4.2. “角色 + 指令型”的结果

使用角色 + 指令型提示词产生的结果与上述简单指令型的结果情况类似，大语言模型依旧对文档进行了归纳总结与改写。即使大语言模型意识到了“最重要的是，我要基于文档本身的内容来操作，不能自己添加或修改事实信息。对于不完整的地方，只能标注出来，不能随意补充内容。清洗后的文本应该保持原意，只是在表达形式上更加规范和清晰。”并且有意地从语言学角度去处理文本，但生成的结果仍然有着信息缺失，句子成分被擅自改变，不忠实于源语料的情况。

4.3. “角色 + 指令 + 约束型”的结果

使用角色 + 指令 + 约束型提示词产生的结果基本符合语料清洗的规范，大语言模型移除了与语言结构无关的格式噪音，同时保持原文的专业内容和叙述结构，只是从格式上变得更加规范。这条提示词使大模型输出的结果基本可以满足语言学研究的需要。

4.4. “角色 + 背景 + 指令 + 约束型”的结果

使用角色 + 背景 + 指令 + 约束型提示词产生的结果偏离了原意，大语言模型给出了清洗与标准化处理方案，并且给出了语料标注相关的代码，与预期的目标不符。

5. 讨论

5.1. 合作原则与提示词设计的关联

本研究通过对照实验证实，以清洗语言学研究所需的英汉海事语料为目的向大语言模型输入提示词时，过多或过少的信息均可能导致输出结果显著偏离预期，而表层形式的微小差异(即使语义近似)亦能引发输出结果的显著分化。本章将立足格赖斯(H. P. Grice)的会话合作原则，对此现象进行深入的理论阐释，以期探讨出利用大语言模型进行语料清洗时的最佳提示词结构框架。

根据格赖斯的合作原则，语言交流的成功依赖于参与者之间的合作与协调。在与大语言模型的交互

中，提示词的设计可以被视为一种“人机协作”的语言行为，其效果受到合作原则的深刻影响。格赖斯的合作原则包括四个：量的准则、质的准则、关联准则和方式准则[9]。这些准则在提示词设计中具有重要的指导意义。

在大语言模型主导的生成语境中，提示语的控制效能并不单纯来自使用者的表达设计，还深刻依赖于模型的解析能力与生成机制。在提示语篇幅较长、结构复杂时，模型可能出现“长距离依赖”处理能力不足的问题，导致信息识别偏差，出现生成偏离[10]。

5.2. 基于合作原则对提示词的分析

量的准则要求所说的话应包含交谈目的所需的信息，且不应包含超出需要的信息。当提示词信息不足时，大语言模型模型无法满足获取足够的所需的信息这项要求。它必须进行大量的、不确定的填补和猜测。例如，在上述实验中仅输入“请你清洗这份语料。”这句提示词时，模型不清楚清洗的用途、以什么方式、面向什么受众执行任务，因此其输出内容可能包含随机的猜测与拓展。当提示词包含冗余细节、无关背景或次要目标时，模型会平等处理所有输入信息。过多的信息可能会覆盖核心指令，或使模型在多个目标间困惑，导致输出重点散漫、包含无关内容，甚至试图同时满足所有要求而产生混乱。比如上述实验在提示词中加入了语料的来源背景后，大语言模型输出的结果并没有因我们提示词的更细致而变得优质。

关联准则要求说话要有关联，即与当前话题相关。模型的注意力机制会评估提示词中所有信息单元与任务核心的关联性。在提示词中加入看似友好但与核心任务无关的语句，这些加入的内容可能会成为新的注意力焦点，不恰当地激活模型内部与之相关的知识网络，导致输出跑偏。这也解释了为什么上述实验中使用角色 + 指令 + 约束型提示词反倒比角色 + 背景 + 指令 + 约束型提示词输出的效果好。

5.3. 基于合作原则设计最优提示词框架

我们可以从格赖斯的会话合作原则的四个具体准则出发，来分析出适用于要求人工智能大语言模型进行海事英汉混合语料清洗的最佳提示词框架。量的准则要求提示词应提供适量的信息，既不能过于冗长，也不能过于简略。质的准则要求提示词应确保信息的准确性和真实性；不包括模棱两可的表述。关联准则要求提示词应与任务高度相关，避免引入无关信息。方式准则要求提示词应清晰、简洁，避免复杂的句型或模糊的表达。

因此，最佳的提示词框架应该具有句子长短合适、信息准确真实、与任务高度相关、表达方式简练的特点。我们经过上文的对比试验可以发现：“角色 + 指令 + 约束”型提示词为要求人工智能大语言模型进行海事英汉混合语料清洗的最佳提示词框架；此提示词框架长短相对适中，表达上无长难句或歧义，同时角色设立与约束条件的部分又保证了输出结果与任务的相关性，输出信息的准确性；因此，在提示词中先为人工智能大语言模型设立一个角色，随后告诉其需要做什么，再为其添加一定的约束条件的“角色 + 指令 + 约束”型提示词；为利用人工智能大语言模型进行海事英汉混合语料清洗的最佳提示词参照框架。

6. 总结与展望

6.1. 总结

本研究通过设计“简单指令”、“角色 + 指令”、“角色 + 指令 + 约束”及“角色 + 背景 + 指令 + 约束”四种提示词框架，在 DeepSeek 大语言模型上进行了海事英汉混合语料清洗的对照实验。实验结果表明，提示词的设计质量显著影响大语言模型在垂直领域语料清洗任务中的输出效能。其中，“角

色 + 指令 + 约束”型提示词(例如：“假如你是一个资深的语言学研究者，请你对这份语料进行清洗，使其符合语言学研究的标准。”)能够产生最符合预期的清洗结果，在移除格式噪音的同时，较好地保持了原文的专业内容与叙述结构。

本研究从格赖斯(H.P. Grice)的会话合作原则理论视角对上述现象进行了深入分析。研究发现，最优的“角色 + 指令 + 约束”型提示词框架恰好满足了合作原则的四项准则：在“量”上提供了适中且必要的信息，避免了信息不足导致的随机猜测或信息过载引发的焦点分散；在“质”上通过明确的角色(资深语言学研究者)和约束(符合语言学研究标准)确保了任务的准确性；在“关联”上紧密围绕语料清洗这一核心任务，未引入无关信息(如语料来源背景)；在“方式”上表达清晰、简洁，无歧义。

因此，本研究的核心结论是：在利用大语言模型进行海事等专业领域的英汉混合语料清洗时，“角色 + 指令 + 约束”型(即在提示词中先为人工智能大语言模型设立一个角色，随后告诉其需要做什么，再为其输出结果添加一定的约束条件。)的结构简明、指令清晰、包含角色设定与任务约束的提示词框架最为有效。该框架为利用大语言模型高效地清洗专门领域的英汉混合语料，高效开展语言学研究，推动人工智能与语言学研究方法相结合提供了思路。

6.2. 本研究的不足

本研究仍存在一定局限性，如语料的样本规模相对较小、使用的大语言模型模型相对单一，设计的提示词框架仍可进一步细化等。

6.3. 展望

利用语言学理论(如合作原则)来设计提示词框架的拓展潜力巨大。在语言学垂直领域，可深入方言学、历史语言学及社会语言学，通过设计捕捉音系、句法历时演变或社会变体特征的提示词，实现对非标准语料的分析与标注。其次，可探索更复杂的约束组合，例如将事实性约束、风格控制与多步推理链相结合，以优化机器翻译、文本生成等需要平衡多重目标的复杂任务。

参考文献

[1] 戴光荣, 郑宇. 机器翻译的数据与算法偏见规避策略研究[J]. 外语教学, 2025, 46(6): 51-57.

[2] 陈秋娜, 徐彩华, 孙素宇. “中文+”视域下职业汉语词表的研制——工程机械技术汉语分级词表示例[J]. 南宁职业技术大学学报, 2025, 33(1): 63-72.

[3] 桂诗春, 宁春岩. 语言学研究方法[J]. 外语教学与研究, 1997(3): 17-23, 83.

[4] 陈钊. 国内外语料库语言学发展研究概述[J]. 辽宁教育行政学院学报, 2021, 38(3): 83-87.

[5] Wang, M. and Hu, F. (2021) The Application of NLTK Library for Python Natural Language Processing in Corpus Research. *Theory and Practice in Language Studies*, **11**, 1041-1049. https://doi.org/10.17507/tpls.1109.09

[6] Kageura, K. and Umino, B. (1996) Methods of Automatic Term Recognition. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, **3**, 259-289. https://doi.org/10.1075/term.3.2.03kag

[7] Srivastava, J., Sanyal, S. and Srivastava, A.K. (2019) An Automatic and a Machine-Assisted Method to Clean Bilingual Corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **19**, 1-19. https://doi.org/10.1145/3342351

[8] Giray, L. (2023) Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, **51**, 2629-2633. https://doi.org/10.1007/s10439-023-03272-4

[9] 胡壮麟. 语言学教程[M]. 第五版. 北京: 北京大学出版社, 2017: 173-180.

[10] 刘华, 陈凯艺. 从表达到调度: 提示语驱动的人机协同与语言能力再理解[J]. 湖南师范大学社会科学学报, 2025, 54(5): 138-147.