

基于依存句法与词性标注的汉英翻译 分析与研究

陈佳仪

上海海事大学外国语学院, 上海

收稿日期: 2025年12月10日; 录用日期: 2026年1月4日; 发布日期: 2026年1月13日

摘 要

本论文通过构建语料库, 利用自然语言处理工具spaCy对语料进行词性标注和依存关系分析, 探讨汉译英翻译中的语言差异。研究表明汉语更倾向使用主动语态和名词短语, 而英语更依赖被动语态和介词短语表达复杂概念, 为翻译实践提供了新的见解, 也助于优化翻译策略及机器翻译系统。

关键词

依存句法, 词性标注, 翻译实践

A Study on Chinese-English Translation Based on Dependency Syntax and Part-of-Speech Tagging

Jiayi Chen

School of Foreign Language, Shanghai Maritime University, Shanghai

Received: December 10, 2025; accepted: January 4, 2026; published: January 13, 2026

Abstract

By constructing a corpus and using spaCy, a natural language processing tool, to conduct part-of-speech tagging and dependency relation analysis on the corpus, this paper explores the linguistic differences in Chinese-English translation. The research shows that Chinese tends to use the active voice and noun phrases, whereas English relies more on the passive voice and prepositional phrases to express complex concepts. This study provides new insights for translation practice and also helps to optimize translation strategies and machine translation systems.

Keywords

Dependency Syntax, Part-of-Speech Tagging, Translation Practice

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,我国对中华文化外译中华文化“走出去”予以高度重视,将文化外译作为推动的重要战略手段。希望通过翻译国内优秀作品,以增强中华文化的传播力和影响力增强国际社会对中国文化的认知与接受。而翻译实践中,汉英语翻译因其语言结构和表达习惯的巨大差异而备受关注。本研究旨在探讨汉译英翻译中的语言差异,特别是词序和词性功能方面的特点,以及这些差异对翻译质量的影响。

1.1. 研究背景

汉语和英语作为两种截然不同的语言体系,它们在语法结构、词汇使用和表达习惯上存在着显著差异。汉语是一种分析语言,其词序相对灵活,依赖于上下文来确定词义和语法关系;而英语则是一种综合语言,其词序较为固定,词形变化丰富,依赖于词性标记来表达语法关系。这些差异导致了汉译英翻译中的诸多挑战,如词义的精确选择、语序的恰当调整以及语义的准确传达等。因此,深入分析这些差异及其对翻译实践的影响,对于提高翻译质量具有重要意义。

词性标注和依存关系分析作为自然语言处理的两个核心领域,对于理解和处理语言差异至关重要。词性标注能够揭示词汇的语法属性,而句法依存分析是按照依存关系对句子中的词语进行分析,句子中的任意两个词之间都存在着某种支配和被支配关系[1]。该过程则能够挖掘词汇和句子的深层含义。在翻译研究中,这两种技术的应用有助于识别和解决因语言差异而产生的翻译问题,从而提升翻译的准确性和流畅性。

1.2. 研究意义

从母语译入外语确实很难,就汉译英而言,最大的挑战就是两种思维模式不一样,存在语言和文化的巨大差异,非母语译者很难真正把握目标语的地道表达,翻译出来的译文可能目标语读者看不懂,达不到国际传播的目的[2]。

本研究的意义在于通过深入分析汉译英翻译中的词性和依存关系问题,提高翻译质量,并发现翻译中的潜在规律。这对于翻译实践者来说,可以提供具体的指导和参考,帮助他们更好地理解和处理语言差异。此外,本研究的发现也将为翻译学和计算机科学等领域的交叉研究提供新的视角和数据支持。

1.3. 研究目标

本研究的目标是通过对官方译文与原文的比较分析,揭示两者在词性分布和语义表达上的特点与规律。通过这一分析,本研究旨在总结词性和语义对翻译效果的影响,并提出相应的改进建议。这不仅有助于优化翻译策略,还能够为机器翻译系统的开发提供理论支持和实践指导。通过这些研究,期望能够为汉译英翻译的质量和效率带来提升。

2. 文献综述

近年来,关于翻译中词性分布与语法结构的研究逐渐增多。李晓雨[3]在其研究中以中英词性标注集 CLAWS7 和 ICTPOS3.0 为出发点,探讨了两种不同标注集背后的名词使用差异。CLAWS7 是英语词性标注集,而 ICTPOS3.0 则是中文词性标注集。通过对这两个标注集的对比分析,李晓雨揭示了中英文在词性标注上的差异,并探讨了这些差异的合理性。该研究强调了词性标注的准确性对于语料库建设和自然语言处理的重要性,指出不断优化词性标注可以为这些领域打下良好的基础。

H. Fei 等人[4]的研究聚焦于跨语言语义角色标注(Cross-Lingual Semantic Role Labeling, SRL),提出了一种端到端的 SRL 模型,该模型整合了多种通用特征和迁移方法。研究显示,通过模型迁移和通用特征的帮助,可以实现跨语言的 SRL。实验结果表明,不同的跨语言特征对 SRL 性能有显著影响,尤其是金标准(gold-standard)语法特征与自动生成的语法特征相比,对跨语言 SRL 更为关键。此外,通用依赖结构特征能提供最佳帮助,而预训练的高阶抽象特征和上下文文化多语言词表示也能带来显著的性能提升。

Semenov 等人[5]的文章讨论了在 Ruzhcorp (俄罗斯-中文平行语料库)中处理中文文本时遇到的语言标注问题,特别关注了俄语借词的处理。文章一方面对中文文本处理的广泛标准进行了理论比较,另一方面描述了在包含许多音译词和借词的特定语料库数据上的词分割、字形到音素转换和词性标注三个领域的实验。最终,文章提出了将在 Ruzhcorp 中实施的中文文本预处理流程。Li [6]的研究通过平行语料库对比分析了中英文名词短语结构的差异,为理解两种语言的语法特征提供了实证数据。Manning [7]等人开发的工具包在自然语言处理领域被广泛应用,本研究中也采用了该工具进行词性标注和依存关系分析。

可以看到中英文词性标注集在理论和实践上都存在显著差异。这些差异不仅影响了词性标注的准确性,也对跨语言语义角色标注和翻译文本的语言标注提出了挑战。以上研究虽揭示了汉英在词性标注、名词短语结构等方面的表层差异,但缺乏从语言类型学、功能语法等理论层面系统解释差异的成因——例如汉语作为分析型语言的“意合”本质与英语作为综合分析型语言的“形合”特征[8],如何导致两者在词性分布、依存关系上的分化。同时,翻译过程中“源语言迁移”的认知机制,也未与依存句法特征的变化形成关联分析。本研究在对比词性分布与依存关系差异的基础上,进一步结合语序、功能语法及显化隐化等翻译理论,探究差异产生的语言本质与认知逻辑,为翻译策略优化提供更根本的理论支撑。同时为了提高词性标注的准确性和跨语言 NLP 任务的性能,需要深入研究和开发能够适应不同语言特性的标注工具和算法。此外,这些研究大多局限于特定语料库,且缺乏对依存关系的系统分析。本研究旨在为此方向的研究添砖加瓦,结合词性分布和依存关系分析汉英翻译中的语法特征。

3. 方法论

3.1. 研究问题

本研究旨在探索翻译活动作为一种认知过程如何影响翻译产出中的句法表现。句子作为自然语言文本的组成单元,其句子结构和句法关系的准确识别对于语义理解至关重要,可以通过句法分析来分析句子结构关系。词性标注可用于语言特征分析及中英文差异比较[9]。语义理解的目标是给定一段文本,能够识别出句子的主体、行为、状态、时间、地点等关键语义信息,并结合词义消歧技术可以解决自然语言中的歧义问题,进一步提升语义理解性能[10]。本文的研究意图在依存语法框架下调查语料库中的句法复杂性。试图解决以下研究问题:(1) 翻译英语的依存句法特征是什么?(2) 源中文文本与翻译英文文本之间的句法特征有何异同?为何异同?(3) 中文作为源语言在翻译过程中如何影响翻译英文的句法结构?

3.2. 实施方法

为了研究上述问题，本研究首先构建了一个汉英语料库，其语料来源于由上海外国语大学语料库研究院胡开宝教授领衔的团队研发的智能化多语种教学与科研平台，经过严格的清洗和预处理，其代表性和质量有所确保。以下为具体的语料信息(见表 1)。

Table 1. Basic information of the Chinese-English corpus
表 1. 汉英语料基本信息

语料	句子数量	字数
中文	83	5105
英文	89	3749

其次，安装下载 spaCy，在 python 中加载模型，输入指令后，使用 spaCy 自动进行词性标注，并提取依存关系信息。随后在自动标注的基础上结合人工检查校对确保其准确性。最后进行进一步的句法统计，进行分析。同时，为确保该研究过程可以被复现，所有的代码指令和语料库均来源于网络开源平台。Nivre 提出的基于感知器的依存句法分析方法在本研究中被用来处理和分析翻译文本[11]。

4. 结果与分析

4.1. 词性分布

使用 spaCy 进行自动词性标注，得到详细的词性结果，随后进一步利用可视化工具，将中文和英文的词性分布分别以条形图的形式直观展现出来。通过对比图 1 和图 2，可以明显观察到，不同词性在两种语言中的分布频率存在显著差异。

从数量上来看，英文词性中，名词占据主要数量，超过 1000 个，随后为介词(ADP)和限定词(DET)，均接近 500 个，形容词(ADJ)与副词(ADV)的数量也占据显著数量，助动词(AUX)和从属连词(SCONJ)的数量也相对丰富。

在中文文本中，名词的使用接近 1200 个，与英文接近，动词为使用频率第二高的词性类别，达 600 余个，多于英文中使用的动词数量。然而，剩余词性的分布均相对较少，彼此之间数量差异也较小。与英文的相比，有显著下降。

从占比上来看，汉语中名词占比约为 48%，动词为 25%，而在英文中，两者数量虽然接近，但名词比例下降至 26%，动词比例骤降为 10%左右。

4.2. 依存关系

英语中的大多数修饰性依存关系为正向，修饰成分通常后置于语义核心，名词短语中的修饰依存关系与言语相比显著增加，然而，汉语的修饰性依存关系倾向于负向，前置修饰的数量有限且结构较为扁平，依存距离较短，以动词为核心节点的句法结构依存关系占比较高。

4.3. 分析

以“高质量发展是全面建设社会主义现代化国家的首要任务”中英文的依存关系为例：

如图 3 所示“高质量发展”作为句子的主语，其中，“高”修饰“质量”表示优质的属性，明确发展的品质维度；“质量”与“高”一起构成复合名词，修饰“发展”，聚焦“发展”的核心特征；“发展”作为核心语素，构成“高质量发展”这一主语成分，整体指向句子陈述的核心对象。“是”作为句子的

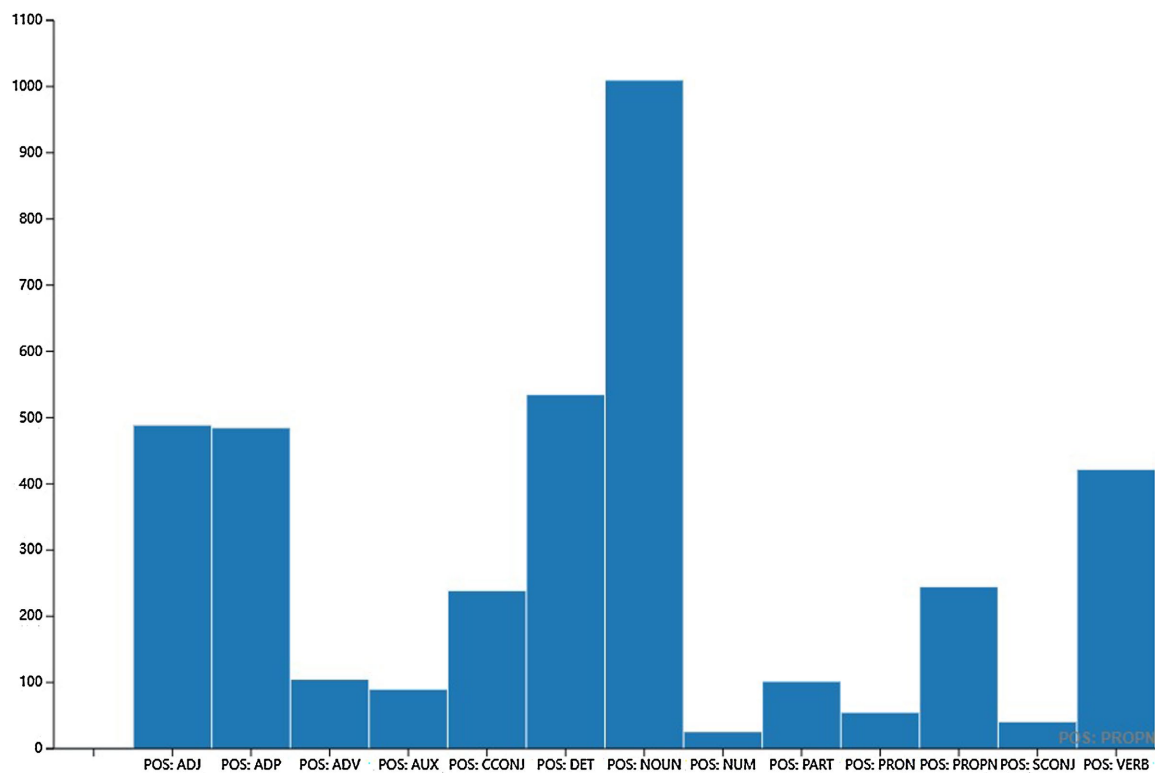


Figure 1. Distribution of parts of speech in English

图 1. 英文词性分布

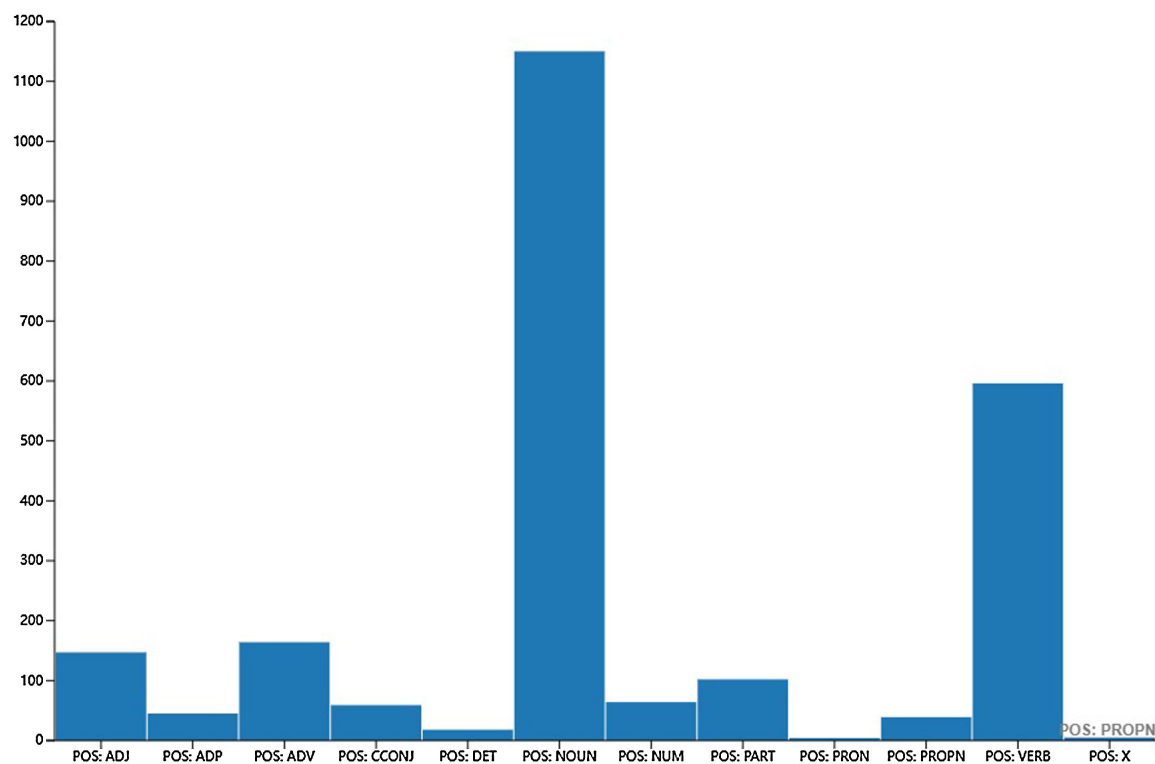


Figure 2. Distribution of parts of speech in Chinese

图 2. 中文词性分布

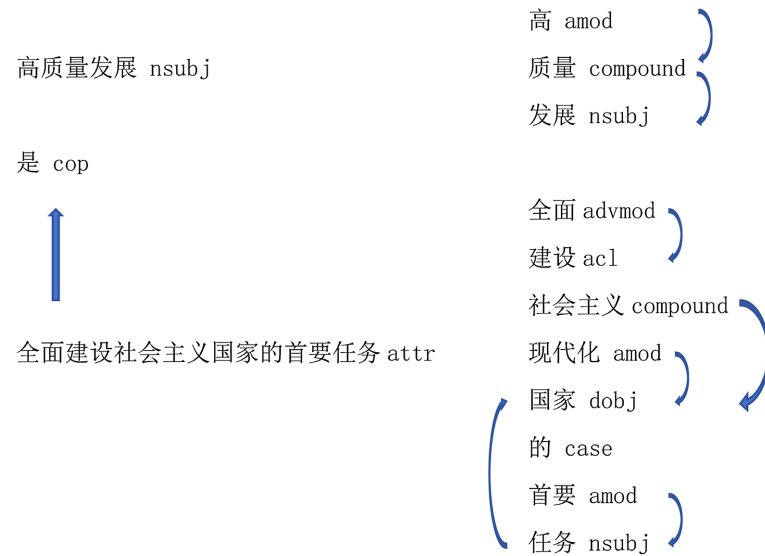


Figure 3. An example of dependency relations
图3. 依存关系示例

系动词核心，承担连接主语与表语的语法功能，体现“主语→表语”的陈述逻辑。“全面建设社会主义现代化国家的首要任务”作为句子的表语，其中“全面”修饰“建设”表示建设的范围是全方位、无遗漏的；“社会主义”修饰“国家”表示国家的制度属性；“现代化”修饰“国家”表示国家的发展阶段；“国”作为“建设”的宾语，明确建设的具体对象；“的”用于连接修饰成分与中心语“任务”，体现偏正结构的语法关系；“首要”修饰“任务”表示任务的优先级为第一位，最终通过“任务”聚焦表语的核心，明确“高质量发展”的核心定位。

该例子相对的英文译文为“High-quality development is the primary task of building a modern socialist country in an all-round way.”其中“High-quality development” (nsubj): 句子的主语，与中文中的“高质量发展”相对应，通过连字符“-”将形容词“High-quality”与名词“development”连接，形成复合名词短语。“is” (cop): 句子的系动词，对应中文的“是”，承担连接主语与表语的功能。“the primary task” (attr): 表语的核心成分，与中文中的“首要任务”相对应，定冠词“the”明确“task”的特指属性。“of building a modern socialist country” (prep): 介词短语，修饰“the primary task”，通过介词“of”明确“任务”与“建设社会主义现代化国家”的从属关系，对应中文中“建设社会主义现代化国家的”这一修饰成分。“in an all-round way” (advmod): 介词短语作状语，修饰动词“building”，对应中文中的“全面”，明确“建设”的方式与范围。

可以看出中文句子中的修饰成分如“高”“全面”“社会主义”“现代化”均以前置修饰的方式附着于核心名词或动词，通过复合名词结构和偏正关系直接组合语义，无需额外虚词衔接；而英文中则通过定冠词“the”“a”明确名词的特指与泛指，通过介词“of”“in”构建短语结构，将中文的前置修饰转化为后置的介词短语修饰，形成层级更清晰的名词短语结构。在两个句子中，“是”和“is”都作为系动词承担连接功能，但中文的“的”作为结构助词仅起连接作用，英文则通过介词短语的语法功能实现同等语义关联。不难看出，中文句子更倾向于使用前置修饰、复合名词和结构助词“的”来表达复杂概念，体现“意合”的语言特征；而英文则通过连字符、冠词和介词短语来构建详细描述，体现“形合”的语言特征，这与汉英两种语言的语法结构和表达习惯差异高度契合。

上述差异的产生，本质上源于汉英两种语言的类型学属性差异。汉语作为典型的分析型语言，缺乏

形态变化,语法关系依赖词序和语义关联,因此倾向于通过名词并列(如“中央纪律检查委员会常务委员会委员”)和主动语态直接组合语义,形成“前置修饰、扁平结构”的依存特征。这是汉语“修饰语-核心语”(Head-final)语序原则的必然结果,无需介词、冠词等形态标记即可实现成分关联。而英语作为综合-分析型语言,词形变化丰富且依赖“核心语-修饰语”(Head-initial)语序,必须通过介词短语(如 of、in charge of)、定冠词 the 等形态手段明确成分从属关系,因此形成了“后置修饰、层级结构”的依存特征,这一差异是两种语言类型学选择的系统性体现。

从功能语法视角来看,汉英差异服务于不同的信息传递策略。汉语注重“话题优先”,将复杂的身份信息整合为厚重主位,述位仅用简洁动词完成陈述,符合高语境语言“依赖语境省略冗余标记”的交际需求;而英语遵循“主语优先”,主位简洁,通过述位的介词短语补充细节,体现低语境语言“形态显化确保语义明确”的策略。这种主位-述位结构的差异,直接导致了两者在名词短语复杂度、介词使用频率上的分化。

翻译理论中的显化和隐化的编码机制进一步解释了差异的转换逻辑。汉语以“意合”为核心,通过语义隐化省略语法标记(如无冠词、少介词),属于“高语境编码”;英语以“形合”为核心,需通过介词、连词显化成分关系,属于“低语境编码”。因此,案例中中文的“负责同志”(名词并列隐化动作关系)在英语中必须显化为“in charge of the relevant parties”(介词短语显化动作关系),这种转换并非单纯的结构调整,而是语言编码方式的适应性适配。英文倾向于使用被动语态和介词短语来构建句子,而中文则更侧重于主动语态和名词短语的直接组合。这种差异体现了不同语言在组织信息和表达思想时的不同方式。

此外,这种对比分析有助于翻译者在实践中做出更恰当的转换策略,避免直接逐字翻译,而是更注重传达原文的深层含义和语境。这对于提高翻译的准确性和自然度至关重要。同时,它也促进了跨文化交流,帮助翻译者和语言学习者更好地理解不同文化背景下的语言使用。

5. 结论

5.1. 主要发现

本论文通过对汉译英文本的词性标注及依存关系的分析,解释了两者在结构和表达方式上存在的差异。数据显示,英文中的修饰性依存关系多为正向,修饰成分通常后置于语义核心;中文则倾向于负向,前置修饰的数量有限,结构较为扁平。因此在翻译过程中,英文的句法结构极易受到中文的影响,尤其在词序调整和语义传达等方面。

综上,汉英词性分布与依存关系的差异,是类型学属性、功能语法策略与语言编码机制共同作用的结果:汉语的“名词+动词”高占比、前置修饰特征,源于其“意合”的语言本质;英语的“介词+限定词”高占比、后置修饰特征,源于其“形合、主语优先”的编码需求。而翻译过程中的源语言影响,本质是汉语“意合逻辑”向英语“形合结构”的迁移与适配——译者需通过显化语法标记、调整修饰语位置、重构主位述位结构,实现两种语言编码逻辑的等效转换,这也是本研究对翻译实践的核心启示。

5.2. 局限与展望

由于数据和语料量的限制,本研究无法覆盖尽量多领域的文本,可能存在一定的偏差,同时本研究选取的语料仅为官方文本及译文,并未与其他翻译语料进行对比研究。对于数据的分析,尽管在得到自动标注的结果后进行了人工处理,仍可能存在细微差错。由于大多数翻译实践研究都集中于某一特定类型的语料库[12],未来可进一步扩展语料来源,少有研究能结合不同类型的语料库辅助翻译实践,并结合其他自然语言处理技术以发现潜在的问题,提高翻译质量和效率。此外,由于翻译技术还需要受翻译规律和语言规律等客观因素的制约,必须处理好“语言与技术”“翻译与技术”“语言与文化”等之间的关

系[13]。可进一步研究对文本处理模型的优化提出针对性的建议, 提高精度。

参考文献

- [1] 杨旭华, 金鑫, 陶进, 等. 基于图神经网络和依存句法分析的文本分类[J]. 计算机科学, 2022, 49(12): 293-300.
- [2] 陶友兰, 赵田园. 中国文化外译实践与研究的现状、挑战与发展趋势——陶友兰教授访谈录[J]. 山东外语教学, 2023, 44(2): 1-6.
- [3] 李晓雨. 英汉对比视角下的中英文词性标注集分析[J]. 海外英语, 2017(3): 191-192.
- [4] Fei, H., Zhang, M., Li, F. and Ji, D. (2020) Cross-Lingual Semantic Role Labeling with Model Transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 2427-2437. <https://doi.org/10.1109/taslp.2020.3016127>
- [5] Semenov, K.I., Titizian, A.K., Piskunova, A.O., Korotkova, Y.O., Tsvetkova, A.D., Volf, E.A., et al. (2021) Linguistic Annotation of Translated Chinese Texts: Coordinating Theory, Algorithms and Data. *Journal of Linguistics/Jazykovedný casopis*, **72**, 590-602. <https://doi.org/10.2478/jazcas-2021-0054>
- [6] Li, X. (2017) A Comparative Study on Chinese and English Noun Phrase Structure Based on Parallel Corpus. *Journal of Foreign Languages*, **42**, 45-59.
- [7] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014) The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, 22-27 June 2014, 55-60. <https://doi.org/10.3115/v1/p14-5010>
- [8] 侯凌霄. 跨境电商学生译者译后编辑能力的培养与实践探索[J]. 科教文汇, 2024(14): 129-132.
- [9] 王朝晖, 余军. 韩素音国际翻译大赛语料库构建及翻译教学应用[J]. 中国翻译, 2019, 40(5): 72-79.
- [10] 杨丹. 基于句法分析的文本语义理解研究[D]: [硕士学位论文]. 西安: 西安工业大学, 2024.
- [11] Nivre, J. (2013) Dependency Parsing with a Perceptron. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, 4-9 August 2013, 349-358.
- [12] 刘秉栋, 曹灵美. 基于语料库的本科生翻译课程教学模式建构研究[J]. 外语电化教学, 2021(5): 68-73, 10.
- [13] 王鼎, 王汝发. 自然语言处理中机器翻译发展态势的哲学思考[J]. 绥化学院学报, 2023, 43(11): 78-80.