

基于语料库的英语学术论文摘要语言特征研究

张瀛文

北华大学, 吉林吉林, 132000;

摘要: 随着学术全球化进程的加速, 英语摘要已成为非母语学者参与国际学术对话的重要途径。本研究构建多学科平衡语料库, 采用词频统计、句法标注及语步分析技术, 系统探究英语学术论文摘要的语言特征, 特别是词汇选择、句法结构和信息组织方面的共性与差异。研究表明, 英语学术摘要在词汇、句法和语篇层面呈现出显著的学科特异性。自然科学倾向于使用高密度名词化结构与被动语态以强化客观性表达, 而人文科学则更注重理论框架构建与逻辑推演。这些发现不仅深化了对学术文本语言规律的理解, 也为非母语学者提供了宝贵的写作指导, 强调了掌握学科特定语言规范的重要性。

关键词: 学术摘要; 语料库分析; 词汇特征; 句法结构

DOI: 10.69979/3029-2735.25.1.087

学术论文摘要是研究成果的核心载体, 其语言特征直接影响知识传播的效度与信度。在学术全球化进程加快的背景下, 英语摘要已经成为非母语学者参与国际学术交流的必经之路。然而, 现有研究表明, 非母语学者在摘要写作中普遍存在语言规范偏差, 如第一人称过度使用、时态混乱及信息密度不足等问题, 这些偏差可能削弱学术文本的权威性与可理解性。与此同时, 学科之间语言的共性和差异还没有得到系统地揭示, 传统的定性研究很难全面把握语言的复杂性。在此背景下, 语料库语言学以其大规模文本量化分析的优势, 为解构学术摘要的深层语言规律提供了方法论支持。

1 研究现状

1.1 学术语篇特征研究

学术语篇特征的系统性研究始于 20 世纪后期, 随着学术英语 (English for Academic Purposes) 作为独立研究领域的兴起, 学者们对学术文本的修辞结构和语言形式特点产生了关注。Swales (1990) 提出的 CARS (Create a Research Space) 模型在此领域具有里程碑意义, 这一模型分析了学术论文引言部分的语步结构, 揭示了学术写作中“建立研究空间”的普遍逻辑框架。CARS 模型不仅为学术语篇的宏观结构分析提供了理论工具, 更启发了后续研究对摘要、讨论等特定章节的功能性探索。然而, Swales 的研究主要聚焦于全文结构, 对摘要这一高度程式化文本的微观语言特征关注有限。

1.2 语料库研究方法进展

语料库语言学在学术英语研究中的应用, 标志着该领域从直觉驱动向数据驱动的模式转型。早期研究多依赖手工检索与频率统计, 如计算学术文本中的词簇 (lexical bundles) 分布, 此类工作虽能揭示特定语言形式的出现规律, 但未能深入阐释其功能动因。多维度分析法 (MD Analysis) 的引入突破了这一局限, 该方法通过因子分析提取语域特征的核心维度 (如“信息密度与叙事性”“交互性与非交互性”), 进而实现语篇类型的系统性聚类。在学术英语研究中, 成功地识别了学科之间在抽取程度、论证方式和作者介入程度方面的显著差异。例如法学论文呈现高频率的逻辑连接词与条件从句, 而化学文本则表现出极端的名词化倾向与低人际互动特征。

2 研究方法

本研究采用语料库驱动的研究路径, 通过定量分析与质性阐释相结合的方法, 系统探究英语学术论文摘要的语言特征。研究设计涵盖语料库构建与多维度语言分析两个核心环节, 旨在突破传统单一学科或单一分析维度的局限, 揭示学术摘要的跨学科共性与差异。

2.1 语料库构建

研究语料选自 Springer 与 ScienceDirect 两大权威学术数据库 2015 至 2022 年间发表的同行评议论文摘要, 遵循多学科平衡原则, 最终构建包含 500 篇摘要的专用语料库。学科分布上, 自然科学 (含物理学、生物学、工程学) 占 40%, 社会科学 (含经济学、教育学、

社会学)占30%，人文科学(含文学、历史学、哲学)占30%，该比例设置参考了Scopus学科分类标准，同时兼顾学科群的语言特征差异性(Hyland & Jiang, 2020) [2]。为确保语料时效性与学术规范性，入选论文均发表于JCR Q1/Q2区期刊，并排除非实证研究类文本(如评论、书评)。

2.2 分析维度

研究从词汇、句法、语篇三个层面构建分析框架，整合自动分析与人工标注技术，形成多维交叉验证机制。

表 1 分析维度

分析层面	具体指标	工具
词汇特征	高频词、词簇、词汇密度	AntConc
句法特征	被动语态、时态分布、名词化现象	Sketch Engine
语篇特征	语步结构、信息焦点	人工标注+UAM CorpusTool

词汇特征分析聚焦学科间共性特征，采用 AntConc 3.5.9 进行高频词统计与关键词提取，计算标准化的词频分布(每百万词频数)，并通过 LLR 对数似然比检验识别学科间显著差异项(Anthony, 2022) [3]。词簇(lexical bundles)分析以4词序列为基本单元，过滤停用词组合后，依据Biber (2006)的功能分类框架，识别结构型(如“the results of the”)与立场型(如“it should be noted that”)词簇的分布规律。词汇密度(lexical density)计算采用Halliday (1985)公式，即实词占总词数的比例，用于评估摘要信息压缩程度[4]。

3 结果与讨论

3.1 词汇特征分析

词汇分析表明，学术摘要表现出明显的动词集群化和信息密度两极化特征^[6]。如表2所示，高频动词 TOP10 中，研究过程动词(如investigate, analyze, examine)占据主导地位，其总占比达45.7% ($\chi^2=32.15$, $p<0.01$)，反映出学术摘要聚焦研究行为描述的核心功能。值得注意的是，demonstrate 与 propose 在自然科学摘要中的频率分别比人文科学高38%与27%，暗示学科间对“知识呈现方式”的认知差异：自然科学侧重于

实证发现的可验证性，人文学科则侧重于建构性的理论架构。

表 2 高频动词分布 (单位: 每千词出现频次)

动词	自然科学	社会科学	人文科学	总体频率
investigate	14.2	11.6	10.3	12.7
analyze	10.1	9.8	8.0	9.3
demonstrate	9.5	7.2	5.1	8.6
propose	8.3	7.1	5.4	7.9
examine	6.9	6.2	5.0	6.5

零冠词现象在方法描述语步中占比高达68%(如“data were collected”)，显著高于结果(24%)与结论部分(8%) ($F=19.34$, $p<0.001$)。这种语法压缩策略与被动语态高频共现($r=0.72$)，共同为方法学命题的客观表述服务。此外，词汇密度分析显示，自然科学摘要的实词占比(72.3%)显著高于人文科学(65.1%) ($t=4.28$, $p<0.05$)，说明硬学科更多地依赖于名词化结构来有效地封装信息。

3.2 句法结构特征

句法分析揭示，被动语态与时态系统构成学术客观性的核心语法机制。如表3所示，被动语态在方法语步中使用频率达52%，其功能聚焦于施事者隐退(如“samples were analyzed”)，而在结论部分骤降至12%，此时主动语态与第一人称复数(如“we argue”)的使用率上升至41%，体现作者立场的策略性显现。

表 3 被动语态分布与时态选择 (单位: %)

语步	被动语态频率	现在时	过去时	现在完成时
方法	52	18	74	8
结果	28	63	29	8
结论	12	81	9	10

时态选择呈现高度语境化特征：过去时在方法语步占74%，用于标示已完成的研究动作；现在时在结论部分达81%，强调研究发现在理论上的普适性。值得注意的是，社会科学摘要的现在完成时使用率(15%)显著高于自然科学(6%) ($\chi^2=21.07$, $p<0.01$)，常用于建立研究背景与现有文献的关联(如“scholars have debated...”)。名词化现象反映了学科的特殊性：自然科学摘要中“动词派生名词+介词”结构占比达38%

(如“application of this method”),而人文科学更偏好“形容词派生名词”(如“significance”)($t=5.13$, $p<0.001$)。

3.3 语篇模式

语篇分析验证了 Bhatia (1993) 四语步结构的主导地位,但是,不同学科之间存在着明显的组织策略差异。全语料库中,背景—方法—结果—结论的完整结构占比 61%,其中自然科学摘要省略背景语步的比例(23%)显著高于社会科学(46%)($\chi^2=34.12$, $p<0.001$)。究其原因,可能是因为硬学科写作的范式更侧重于研究创新,而不是理论上的对话[6]。信息焦点分析表明,自然科学摘要的“结果”部分平均包含 3.2 个数据指标(如“increase by 15%”),而人文科学更侧重理论推演,逻辑连接词(如“therefore, consequently”)频率比自然科学高 57%。

4 结语

研究表明,英语学术论文摘要的语言特征在词汇、句法和语篇层面均呈现出显著的学科特异性。未来工作应进一步探索如何利用大规模语料库分析技术,提升非母语学者在国际学术交流中的表达效果。通过开发针对性的写作辅导工具,有助于作者更好地把握学科特有的语言规范,提高科研成果的传播力和影响力。随着机器学习、自然语言处理等技术的不断发展,有助于实现对

学术写作规范演化的动态跟踪,从而实现教学资源的实时更新,更好地支撑学术写作教育的发展。最终,这些努力将促进全球学术共同体内部更加高效、平等地知识共享与对话。

参考文献

- [1]Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text[J]. arXiv preprint arXiv:1903.10676, 2019.
- [2]Hyland K, Jiang F. Text-organizing metadiscourse: Tracking changes in rhetorical persuasion[J]. Journal of Historical Pragmatics, 2020, 21(1): 137-164.
- [3]Hyland K, Wang W, Jiang F K. Metadiscourse across languages and genres: An overview[J]. Lingua, 2022, 265: 103205.
- [4]Boginskaya O. Evaluative Stancetaking in English-Medium Academic Prose: A Study of Research Article Abstracts by Russian and Chinese L2 Writers[J]. Journal of Language and Education, 2024, 10 (3): 40-52.
- [5]金常心,李加■. 汉英学术论文摘要语步对比研究[J]. 宁波工程学院学报, 2023, 35 (01): 84-90.
- [6]葛迪. 科技英语论文摘要的构词特点及翻译研究[J]. 校园英语, 2020, (24): 237-238.