中国语言学研究 70 年：核心期刊的词汇增长

王珊
澳门大学人文学院中国语言文学系
珠海澳大科技研究院
shanwang@um.edu.mo

詹润哲
澳门大学科技学院计算机和信息科学系

姚双云
华中师范大学语言与语言教育研究中心

摘要

建国以来我国语言学经过 70 年的发展取得了瞩目的成就，已有研究主要以回顾主要历史事件的方式介绍这一进程，但尚缺少使用量化手段分析其历时发展的研究。本文以词汇增长为切入点探究这一主题，首次创建大规模语言学中文核心期刊摘要的历时语料库，并使用三大词汇增长模型预测语料库中词汇的变化。本文选择拟合效果最好的 Heaps 模型分阶段深入分析语言学词汇的变化，显示出国家政策的指导作用和特定时代的语言生活特征。此外，与时序无关的验证程序支撑了本文研究方法的有效性。

关键词：中国语言学；词汇增长；核心期刊；摘要；语料库；历时发展

70 Years of Linguistics Research in China: Vocabulary Growth of Core Journals

Shan Wang
Department of Chinese Language and Literature,
Faculty of Arts and Humanities,
University of Macau, Macau, SAR, China

Zhuhai UM Science & Technology Research Institute,
Zhuhai, China
shanwang@um.edu.mo

Runzhe Zhan
Department of Computer and Information Science,
Faculty of Science and Technology,
University of Macau, Macau SAR, China

Shuangyun Yao
Research Center for Language and Language Education,
Central China Normal University,
Wuhan, China

Abstract

Since the founding of P.R. China, linguistics in China has made remarkable achievements after 70 years of development. The existing studies have mainly introduced the development of linguistics by reviewing historical events, but no research has used quantitative means to analyze its longitudinal development. This article has explored this topic from the perspective of vocabulary growth. For the first time a large-scale diachronic corpus of abstracts from Chinese core linguistic journals is created. Subsequently, the analysis is conducted on this corpus with the help of three vocabulary growth models. Then the Heaps model with the best fitting effect is selected to further analyze the changes of linguistic vocabulary in different times, showing the guiding role of national policies and the characteristics of language life in specific era. Furthermore, a time independent validation procedure is performed, which supports the effectiveness of the proposed methodology of this study.

Keywords: Linguistics in China; Vocabulary Growth; Core Journals; Abstract; Corpus; Diachronic Development

1. 引言

建国以来，我国语言学研究经过 70 年的发展，从筚路蓝缕到开拓创新，取得了瞩目成

第二十一届中国计算语言学大会论文集，第1页-第13页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

就。已有的研究主要以回顾主要历史事件的方式介绍这段历史，对我们了解前辈学者和当代时贤对语言学的贡献起着重要作用。但是，尚无研究对这段时期语言学词汇的演变进行历时考察。汉语语言学核心期刊能够代表在不同时期中国语言学的最新发展动向和关注热点，而论文摘要是整篇文章中最具代表性的部分，具有牵引文章脉络的作用，展现了极高的信息密度，能够在一定程度上直观地反映语言学的发展。

词汇增长（Vocabulary Growth）模型体现了文本中独特性用词的比例。对于以时间序列组织的文本，该指标能够反映文本在一定时间范围内是否有新词加入、新词增加的比例或速度等特征。词种（types）与词例（tokens）数量的比值 TTR（type-token-ratio）刻画了一定长度的文本内非重复用词的比例，常被用作建模词汇增长问题的指标之一。在历时语料中，以时间序递进式扩大计算长度窗口，统计不同采样点的 TTR 能够反映出新词的增长率变化情况；词汇增长模型能够反映 TTR 的变化趋势，并预测在语料数量继续扩大情况下新词数量将以何种趋势增长。词汇增长模型的有效性已在不同类型的文本上得以证实（Savoy, 2015; 王珊、王会珍, 2019）。对于学术领域的文本，新词的出现或增长情况能够直观地反映出学科的发展，故选取词汇增长模型对学术文本的 TTR 变化特征进行建模拟合，能够体现某个学科的历时发展与演化特征。目前尚无采用词汇增长模型的方法探究学科领域演化的研究。本文选取建国以来的语言学中文核心期刊摘要，构建大规模历时语料库，进行词汇增长建模对约 70 年间语言学领域的词汇演化进行定量与定性分析，探究词汇的变化趋势与学科发展、社会因素之间的关系，从而进一步折射七十年来我国语言生活的变化情况。

2. 相关研究

2.1 词汇增长研究

词汇增长已用于判定作者身份、评定语言能力、分析施政风格等诸多研究中。Hoover（2003）对十二位作者的作品词种数量进行统计，发现词汇增长可以用于判断作者的身份。Yu（2010）指出词汇增长与写作和口语的质量在统计学上具有显著的正相关性，证明对于语言能力较高的人，其词汇增长指标也会相对较高，是考察学习者对一门语言掌握情况的指标。Mellor（2010）发现词汇增长可以用于衡量说话者与写作者的语言程度，语言程度越高，则使用的低频词较多。X. Wang（2014）分析了英语二语学习者的电子邮件中的词汇增长与写作熟练度之间的关系。Savoy（2015）分析了 1790 年到 2014 年历任美国总统发表的 225 篇演讲中的词汇增长情况，对比了 Heaps、Hubert-Labbe 两个词汇增长模型的适用性并对不同的词汇增长的变化作出了分析，作者将整个时间段分为低于模型预期值与高于模型预期值的时间段，联系历任总统的施政风格与对应时代的政治经济背景进行定性分析。王珊、陈钊、张昊迪（2021）利用词汇增长模型刻画十余年间澳门新闻报刊内容的词汇历时演变，结果表明词汇增长的倾向性与施政时期方针、人民生活的变化有极大关联性。

2.2 学术汉语的词汇研究

Swales（1985）提出了专门用途英语（ESP, English for specific purpose）与学术用途英语（EAP, English for academic purpose）的概念。学术词汇是学术语言的重要组成部分，是除了核心词汇外使用频率较高的词汇种类（Paquot, 2010）。学术用途英语研究取得了丰硕的成果，但目前对学术汉语词汇的研究仍旧十分稀少。涉及学术汉语的研究可分为三大类：一是文本特征研究，例如吴格奇、潘春雷（2010）参考 Hyland（2005）提出的立场分析框架，分析立场标记词汇以探究学术写作者的语用策略与身份建构；张赪、李加、申盛夏（2020）基于自建的多学科语料库，分析了不同学科的论文写作者对实词与虚词的使用特征；朱宇、胡晓丹（2021）基于自建的人文社科语料库，利用多维度分析法针对连词在论文中的语言功能进行了考察。二是英汉对比研究，其研究多以中文为母语者为考察对象，包括对立场信息一致性的对比（赵永青等, 2019）、身份指称的差异对比（李志君, 2014）、衔接用词的差异对比（胡芳、陈彧, 2005）等。三是学术词表的研制，刘锐、王珊（2017）用小规模知网学术论文语料构建学术词表，王笑然、王佶旻（2022）则通过自建经贸类的学术语料库，研制经贸类学术汉语词表。从现有研究来看，目前公开的大型综合性汉语语料库（如 BCC、CCL、Chinese

Gigaword 语料库等）中缺少学术领域的语料，进行有关研究仍需自建共时或历时语料库，所需时间代价与人力成本较高，影响了学术汉语研究的开展。

2.3 建国后的中国语言学发展

建国以来，我国的语言文字工作取得了瞩目的成就。建国初期国家就对文字改革提出了三大方向：简化汉字、推广普通话与制定推行汉语拼音方案，间接推动了语言学多个领域的发展，例如推广普通话所需要的调研任务为汉语方言学的研究打下了扎实的材料基础，在制定汉语拼音方案的过程中对普通话音位系统的描写奠定了汉语音系学研究的基础。尽管中间历经了文化大革命的动荡使语言学研究工作有所停滞，但改革开放后，在国家各行业逐步与国际环境接轨的大环境下，外来理论和方法的引入为汉语语言学研究带来更多新的视角；在吸收外来理论的基础上，现在我国汉语语言学研究工作正逐步迈入自主创新的阶段。

对于建国以来的语言学工作的开展，陆俭明（1999）从学科建设和学术发展的角度对 21 世纪之前的语言学工作与研究进行了梳理，刘丹青（2019）对我国语言学研究各个分支领域近 70 年内代表性的理论与应用进行了介绍，国家语言文字工作委员会（2019）则整理了纪年史料，侧重对党和国家在语言文字工作上的重要事件进行了汇编总结。这些研究回顾了语言学主要的历史事件，但尚未有借助量化统计方法对建国以来的语言学发展的分析。

综上所述，在大规模语料库基础上，对学术汉语中语言学的词汇增长进行考察，能够进一步完善现有的研究，丰富语言学发展的研究成果。语料库中的词汇增长情况能够使用数学模型进行展示，故对以时间关系组织的文本序列进行增长情况建模，能够得到目标文本历时的 TTR 变化信息。籍由此，我们能够进一步分析相应时间段之内和不同时间段之间的文本特征及其变化情况。本文选取自建国以来语言学中文核心期刊现存所有电子化收录的语料来创建语料库，主要采用语料库驱动、定量与定性相结合的研究方法，旨在解决以下问题：⑴如何构建具有代表性的语言学中文期刊语料库？⑵如何对语料库进行词汇增长模型建模并分析其反映的语言学发展特点？⑶如何验证词汇增长模型对语言学领域的词汇分析的准确性？

3. 创建语言学核心期刊语料库

现时对中文核心期刊的认定有认可度较高的索引，例如中文社会科学引文索引（Chinese Social Sciences Citation Index，CSSCI）、北京大学中文核心期刊要目总览与中国科学引文数据库（Chinese Science Citation Database，CSCD）等。其中 CSSCI 索引作为国家教育部的重点课题攻关项目，采取定量与定性评价相结合的方法筛选出人文社科领域的标志性期刊[1]，具有较大的影响力与公信力（马费成, 2000; 苏新宁, 2012; 邹志仁, 2000），被知网、万方等电子化文献数据库收录。CSSCI 数据库来源期刊的遴选工作遵循以下原则：公开、公平、公正；总量控制，动态调整；定量（文献计量指标）评价与定性（学科专家）评价相结合；质量优先，兼顾地区与学科平衡。所有入选期刊必须具备以下基本条件：刊载人文社会科学原创学术论文和学术评论等一次文献为主的中文学术期刊；中国大陆出版的期刊应具有 CN 号；按既定出版周期准时出版，符合期刊编辑出版规范，文献信息著录完整、规范。此外，CSSCI 索引不但针对人文社科研制，还对人文社科下属若干个子学科具有较完备的细化分类。

本文依据《CSSCI 来源期刊（2019-2020）目录》中“语言学”子类所收录的 24 个期刊，收集已电子化的论文元信息和摘要作为语料的来源。其中，论文的元信息包括期刊名、年份、期（卷）、作者、页码等信息。论文摘要高度凝练了学术论文的背景、动机、观点与结论，是整篇文章中最具代表性的部分（S. Wang, Liu, & Zhou, 2022）；就阅读过程而言，摘要具有牵引文章脉络的作用，展现了极高的信息密度；摘要中的词汇、句法均带有密集且丰富的特定领域话语特征，故本研究以论文摘要作为研究对象。

本研究构建的语言学中文核心期刊语料库收录了自 1957 年 6 月至 2020 年 8 月共计 71988 篇论文的信息。本文对所有能收集到的语料进行了预处理：第一，期刊发布的信息有一些不属于学术论文的内容，例如征稿通知、编委会信息、会议通知等与语言学的词汇增长情况并

[1] CSSCI 来源期刊遴选标准: https://cssrac.nju.edu.cn/gywm/lxbz/20200102/i64328.html

无关联，故本研究将其视其为噪声信息并采取关键词过滤的方式进行排除。第二，个别期刊的早期文献以繁体中文书写，例如 1957 年《外语教学与研究》中大部分论文以繁体中文收录，而在计算中，简繁词语字形不一致将被视作两个单位进行统计，本文使用 OpenCC 工具[2]将所有繁体字形式统一转换为简体中文，例如“英語詞彙学”转换为“英语词汇学”。第三，标点不属于词汇，因此统计时根据 Python 中的 zhon[3]与 string[4]库分别去除了中英文标点符号。

为了进行词汇信息的提取，预处理后的语料由 pkuseg 多领域中文分词工具[5]进行词语切分。该工具由北京大学语言计算与机器学习组研制（Luo, Xu, Zhang, Ren, & Sun, 2019），与 jieba、THULAC 等分词工具包相比，在细领域分词的 F-Score 与跨领域测试的平均分上均占据优势。默认分词模型在混合领域上训练，并支持在细领域上对预训练模型进行调优以取得更高的准确率。由于学术期刊涉及议题所含的领域较为复杂，语言学不仅仅涉及本体理论研究，更与各学科与社会现象紧密相关，故本研究采取该工具的默认分词模型进行词语切分工作[6]。

表 1 列出了“语言学中文核心期刊摘要语料库”（1957-2020）的信息，包括期刊名称、所搜集到的期刊的时间跨度、记录总条数、有摘要的文献数量，以及词例数、词种数、字数等统计信息。其中时间跨度、记录总条数、摘要数量是在经上文所述预处理方法清洗后的文本上得出的，词例数、词种数、字数的统计方法为：基于所获得的摘要，统计含中外文的词例、词种、字数，例如，分词后的文本“HSK 四 级”，计为 3 个词例、3 个词种、5 个字。该语料库共涵盖 24 个语言学核心期刊的 65791 个摘要，语料规模高达 5949428 词例，198241 词种，10813606 字。

表 1 语言学中文核心期刊摘要语料库概况

期刊	时间跨度	记录总条数	摘要数量	词例数	词种数	字数
外语教学与研究	1957-2020	3783	3443	308826	23640	573421
当代语言学	1962-2020	2310	1969	196766	18505	385596
现代外语	1978-2020	2648	2575	245678	19943	462929
外国语	1978-2020	3567	3355	316703	28601	648638
方言	1979-2020	1824	1702	126130	18454	224487
上海翻译	1979-2020	2976	2795	221340	20206	413212
中国翻译	1979-2020	5509	5171	468122	39394	878834
外语教学	1979-2020	4294	4117	394671	25443	715272
民族语文	1979-2020	3181	2783	179859	19534	324393
外语电化教学	1979-2020	3784	3614	321875	19196	588461
语言教学与研究	1979-2020	2852	2677	207521	15503	359776
外语教学理论与实践	1979-2020	2174	2151	217189	15612	395126
外语界	1980-2020	2786	2533	226961	13476	416330
汉语学习	1980-2020	3980	3488	351913	26654	597033
语文研究	1980-2020	2076	1973	171294	19724	290294
当代修辞学	1982-2020	5578	5034	532995	43538	907909

[2] https://github.com/BYVoid/OpenCC，本文选取“t2s.json Traditional Chinese to Simplified Chinese 繁体到简体”

[3] https://pypi.org/project/zhon/

[4] https://docs.python.org/3/library/string.html

[5] https://github.com/lancopku/pkuseg-python

[6] pkuseg 性能评估：https://github.com/lancopku/pkuseg-python/blob/master/readme/comparison.md

外语与外语教学	1984-2020	4664	4298	377824	24639	698950
世界汉语教学	1987-2020	2059	1592	158581	14522	311380
古汉语研究	1988-2020	2617	2433	175187	25101	293994
语言文字应用	1992-2020	2733	2315	187747	13808	333344
中国语文	1994-2020	2692	2176	213085	24228	367221
语言科学	2002-2020	1493	1243	125248	12876	219472
中国外语	2004-2020	1615	1588	156500	11401	292117
汉语学报	2004-2020	793	766	67413	8214	115417
总计	/	71988	65791	5949428	198241[7]	10813606

4. 以词汇增长建模语料库中的词汇历时变化

4.1 三种词汇增长模型

词汇增长模型假设词例与词种之间存在着某种函数关系，该关系能够预测在不同词例下的词种的数量。通过分别分析词种的预测值与观测值之间的差值与 TTR，能够反映不同时间段下词汇丰富度的变化情况。词汇增长模型主要有以下三种：

Guiraud（1954）提出对语料库的词汇增长进行建模，即转换为词种对词例数量的比值之间的关系。词种数量V与词例数量N的平方根的比值为常数c。若将该法则转换为预测模型，预测词种数量V'与当前词例数量n之间的关系可被表示为：

$$V'(n) = c \cdot \sqrt{n}$$

Heaps（1978）则提出，在双对数空间内，预测词种数量V'与当前词例数量n之间存在线性关系，并可通过如下关系式表达：

$$V'（n） = k \cdot \ n^b$$

其中，对一般的英文语料库来说，典型的k, b值处于以下区间内（Manning, Schütze, & Raghavan, 2008）：$30 \leq k \leq 100, b \approx 0.5$。例如，在 Reuters-RCV1 数据集（含有约10^5个词例）下，拟合得到的参数值为$k = 44, b = 0.49$。

Hubert and Labbe（1988）等人将词频因素作为建模的考量，提出了另一种基于词频高低的增长模型（Hubert & Labbe, 1988; Labbé, Labbé, & Hubert, 2004）。给定当前语料占总语料的比例u，则预测词种数量V'与参数u之间存在如下关系：

$$V'(u) = p \cdot u \cdot V + （1 - p）[V - \sum_{i=1}^{i=freq} V_i \, (1-u)^i]$$

其中，V_i代表词汇出现的频次为i的词汇的数量， p为预估参数，代表出现频次较低的词所占的比例。p的值受到文本的词汇多样性影响较大，能够体现不同文本的风格，Savoy（2015）应用该模型分析了不同时期美国总统演讲的语料。

上述三种模型均用于本文的实验以进行词汇增长拟合效果的对比。

4.2 词汇增长模型设置

本节对上面提到的三种模型使用非线性最小二乘法（Non-linear Least Squares）进行拟合。给定一组词汇增长采样点$\{(x_1, y_1), (x_2, y_2), \ldots, （x_n, y_n）\}$，用于拟合增长的模型$y = f(x; \boldsymbol{\theta})$,其中$\boldsymbol{\theta}$为若干模型参数集合$\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$，拟合过程中采用的残差为$r_i$，迭代过程中根据残差平方和进行拟合，形式化如下：

[7] 因不同期刊之间的词种有重复，因而此处不是每个期刊的词种数量相加。其他表格相同。

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{r_i^2}{2\sigma_i^2}$$

$$r_i = y_i - f\ (x_i; \boldsymbol{\theta})$$

实验通过拟合得到Heaps模型中参数 $k = 24.04, b = 0.58$，Guiraud模型中参数$c = 81.26$, Hubert-Labbe 模型中参数 $p = 0.00$（拟合参数触碰至搜索下界）。在下文中所得出的数据均基于这些参数。

在文本处理中，我们依时间对语料库中的文本每隔 1000 个点进行采样。此外，为了分析特定时间段内的词汇增长，我们在特定时间段内进行了新词检测实验。本文的新词是指在下一个时间段未出现在之前所有时间段的词，例如“新冠”一词首次在 2020 年的摘要中出现，但在 2020 年之前的摘要中没有这个词。新词检测由当前时间段的累积词汇集合减去之前所有的时间段的累积词汇集合得出，其中检测出的新词按照词汇出现频次进行排序。由于各区间段结果数量分布不一致（新词数量如表 4 所示），分析时统一筛选了排行前 50 的词语。

4.3 语言学核心期刊摘要语料库的模型拟合

不同词汇增长模型对语料实际增长情况的预测模式不一，但均无法处理由于现实语言生活因素带来的增长突变。在本例中，Heaps 模型对前期增长拟合较好，但在后期存在高估的趋势，Guirauds 模型则反之；Hubert-Labbe 模型与前两者相比整体趋势为低估增长，原因在于拟合得出的超参数p触碰到了搜索下界 0 值，使原本模型中低频词的部分失效，从而造成了较差的拟合表现。图 1 显示了不同模型对实际观测值（采样点）的拟合情况，曲线代表词种与词例数量之间的增长关系，残差r_i总和显示 Heaps 模型对现实增长拟合最好。在增长起始与中间阶段（20 万词例与 30 万词例间），各个模型的拟合性均出现了较大的偏差。图 2 则将图 1 中 Heaps 模型拟合得到的增长曲线与关键的年份时间点结合进行可视化，关键年份点的划分方法与依据将在本文第 5 节进行详细阐述。

图 1 不同模型拟合与现实词汇增长的曲线

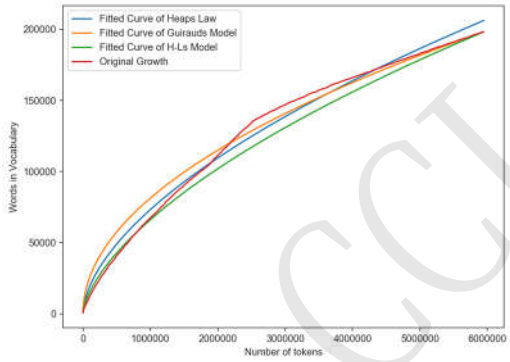


图 2 现实词汇增长情况在年份区间上的映射

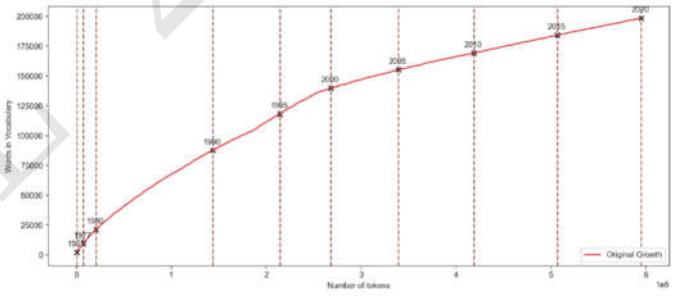
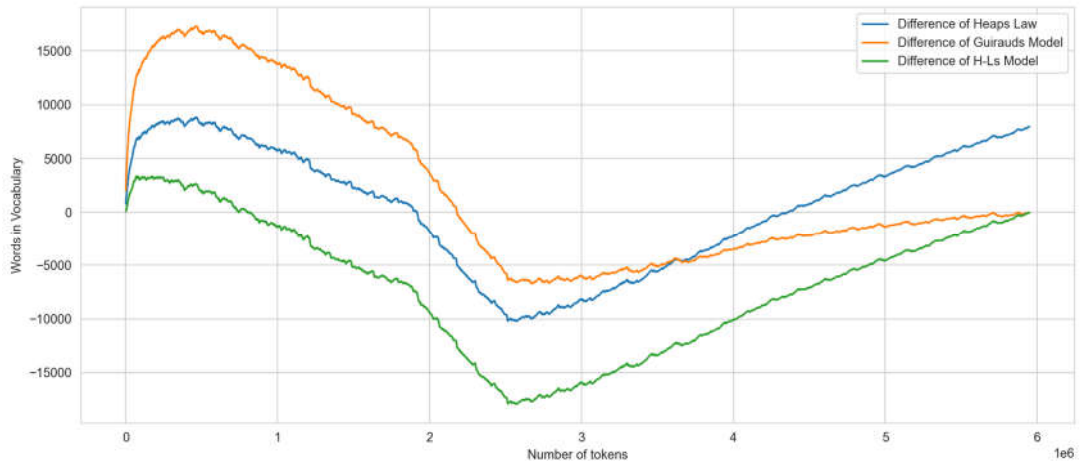


图 3 显示了模型在词汇增长预测中与实际观测值的偏差情况。从总体上看，增长起始阶段大部分模型预测值高于现实文本的词汇增长，现实增长的词种无法达到模型预期，这可能是由于学科发展缓慢造成的。而在中期预测拟合值则低于实际观测值，显示该阶段现实语言生活的中的重大事件可能较多，学科的快速发展引入的新词数量远高于模型预测。后期二者之间的偏差趋于稳定。可能为学科发展稳定亦或是特殊事件导致的语言生活各方面发展稳定。总体上看，利用三大模型的实验结果显示，Heaps 模型的拟合效果最好。下文将结合该模型的结果，对语言学发展的各个时期进行分析。

图 3 不同词汇模型的预测误差曲线



5. 语言学核心期刊摘要的词汇增长的历时分析

学科的发展具有一定的阶段性，若使用每一年作为分析的周期，容易受该年份随机性的影响，结果常具有偶然性。现有的研究也都采用分期的方式，例如刘丹青（2019）基于理论与应用研究发展历程，将中国语言学工作的发展总结为三段，即“封闭自足—对外开放—自主创新”；郭熙（2019）在考察政策、文献、年份代表性词语、语言产品的基础上对 70 年来中国的语言生活总结为两个阶段：前三十年（1949～1978 年）和后四十年（1979～2018 年）。本文先利用词汇增长模型对大规模语料得出词汇历时演化的基本趋势（第 4.3 节），再针对宏观趋势结合时代关键节点分析影响演化趋势的主要原因。

本研究采用的分期情况如下：（1）将 1957~1980 年视作特殊的学科发展起始阶段。一方面，建国后各项科学文化工作百废待兴，加之 1966~1976 年间文化大革命的爆发令学科发展停滞，大部分期刊停刊或未保有出版记录，可供观测分析的样本较少，1978 年前的期刊情况如表 2 所示。另一方面，改革开放初期大量刊物创刊并开始出版，1978-1980 年的期刊情况如表 3 所示。（2）1980 年后期刊多样且数据较为平衡，该段以十年为周期进行分段，分为 1980 年代（即 20 世纪 80 年代）、1990 年代、2000 年代、2010 年代，具有整体性与稳定性。语言学是一门与社会关联性紧密的学科，这样分期便于指称不同时代的特点。每阶段的观测值与 Heaps 模型的预测值数据如表 4 所示，预测值与观测值的差值变化及在年份上的映射情况则如图 4 所示（与 Hubert-Labbe 模型相对比），整体趋势同 4.3 中所述一致。

罗常培（1989）指出，“（88 页）语言的内容足以反映出某一时代社会生活的各方面的侧影响。社会的现象，由经济生活到全部社会意识，都沉淀在语言里面”。故透过计量得到对预测影响较大的词汇，从语言与社会的关系出发，进行语言生活中的重要事实的分析，是探究词汇增长模型预测差异的恰当的切入点，而学术文献也可折射语言学这一学科发展的不同阶段。

表 2 1957-1977 年考察期刊概况

期刊	时间跨度	文献数量	词例数	词种数	字数
外语教学与研究	1957-1977	568	45204	6390	78861
当代语言学	1962-1977	268	26328	4779	49651
总计	/	836	71532	9057	128512

表 3 1978-1980 年考察期刊概况

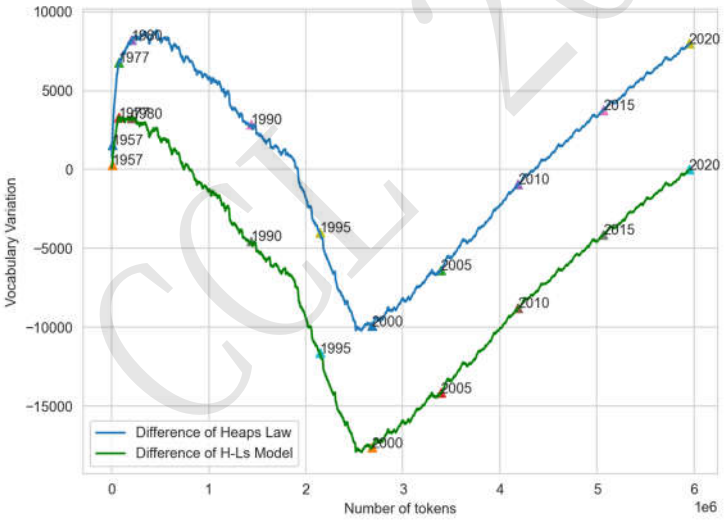
期刊	时间跨度	文献数量	词例数	词种数	字数
语文研究	1980-1980	15	1787	781	3073
方言	1979-1980	54	8104	2529	14440
外语教学理论与实践	1979-1980	90	9633	2423	17244
外语界	1980-1980	24	1952	921	4052

外语电化教学	1979-1980	90	8953	2454	16072
外国语	1978-1980	206	19206	4465	36232
汉语学习	1980-1980	87	13589	3099	22111
外语教学与研究	1978-1980	164	12200	3187	22662
中国翻译	1979-1980	76	6660	2451	13371
现代外语	1978-1980	149	9872	3202	21157
语言教学与研究	1979-1980	110	6158	1845	10469
当代语言学	1978-1980	279	24638	5065	48818
民族语文	1979-1980	99	6901	2153	12423
外语教学	1979-1980	53	4270	1644	8154
总计	/	1496	133923	16894	250278

表 4　每阶段现实观测值与 Heaps 预测值

时间段	累积词例	累积词种	Heaps 预测值	Heaps 预测值-观测值/累积词种	观测值 TTR	预测值 TTR	新词语数量（在之前所有阶段未出现的词）	累积文献数量
1957-1977	71532	9057	15833	6776	0.12661	0.22134	9057	836
1957-1980	205455	21001	29215	8214	0.10222	0.14220	11944	2332
1957-1990	1436173	87482	90352	2870	0.06091	0.06291	66481	15279
1957-2000	2679564	139687	129779	-9908	0.05213	0.04843	52205	30211
1957-2010	4187052	169098	168171	-927	0.04039	0.04016	29411	48353
1957-2020	5949428	198241	206222	7981	0.03332	0.03466	29143	65791

图 4 词汇增长模型预测误差曲线在年份区间上的映射



在初创阶段（1949~1980 年），区间终点采样点的模型预测值高于实际观测值 8214 个词种。由于建国初期百业待兴、经济基础薄弱，全民文化教育水准也普遍较低，文盲率高达 80%，该阶段我国较为重要的语言生活事件集中在国家所开展的文字改革，因此 50 年代我国语言学科发展较为缓慢。由于普通话、汉语拼音方案的推广，该阶段的学科研究热点关注汉语语音研究，例如 1958 年全国人大批准颁布《汉语拼音方案》（中华人民共和国第一届全国人民代表大会, 1958）前，语言学界已于 1957 年发表《北京语音音位简述》（徐世荣, 1957）对普通话音位系统给出了理论描述，为普通话音位系统深化研究奠定了基础。从计量结果来看，“变调”作为词频最高的新词出现，其他出现次数较高的词语还包括“声”、“韵”、“调”、“音标”、

“上声”、“阳平”、“阴平”等。但随之而来的十年文化大革命扰乱了语言生活的正常秩序，更令刚起步的学科受到重创，该区间段的高频新词中出现了“四人帮”，间接表明文化大革命以及社会因素对该阶段学科发展的影响之大。另一方面，在语言生活的国际交流层次与前苏联关系密切，使得语言学研究领域分类格局等受到苏联影响较大，早期文献（B·A·阿尔乔莫夫、郑昌荣、周毅, 1958; 林学洪, 1958; 刘世沐, 1958; 王鸿斐, 1958） 中出现的如“资产阶级”、“社会主义”、“苏联”、“斗争”等，印证了该时期语言生活的意识形态与语言学发展受到苏联等社会因素的制约。

1981~1990 年间，区间终点模型预测值高于观测值 2870 个词种，且从趋势上看模型的预测正偏差开始急剧地向负偏差增长，表明该阶段学科的发展开始恢复，并具有增长的趋势。相较于之前 30 年间政治运动因素是影响社会语言生活的重要因素（郭熙, 1999），改革开放带来的思想开放与新概念、新事物的引入使得语言生活中的政治成分下降，所涉及的主题逐渐丰富。首先，语言生活学术环境极大改善，文革后期刊的复刊（如《中国语文》）与雨后春笋般出现的新刊物（如《方言》等，见表 3）为语言学工作者构造了良好的空间，“辞格”、“喻体”、“义项”、“借代”等新增词占据高频榜单，代表着语言生活中对学术研究探求的正常化。而该阶段新增高频词中“测试”、“考生”、“试卷”、“托福”等说明了语言教育生活的正常化及进阶化，尤其是 1986 年《中华人民共和国义务教育法》的通过使得越来越多儿童能够接受正规的语言教育，且恢复高考后高等教育体系的不断完善使得中国语言学的硕博士点不断增加，进而为语言学科发展输送了大量人才。与此同时，学界对外关闭的大门得以敞开，不少论文介绍西方语言学理论，使语言学的理论基础与研究范式得以提升，丰富了汉语研究方法，研究的视角不再只聚焦于印欧语系的语言特点，例如该阶段徐烈炯对生成语法的介绍（徐烈炯, 1988）。此外，新的研究方法与视角也不断涌现，例如 80 年代初在《中国语文》期刊上对析句方法的讨论（华萍, 1981; 陆俭明, 1981）。 “国家教委”是该阶段后五年出现频率最高的词，体现了核心期刊选题方向依然以国家重要机构的指导政策作为参考，但同时语言学的发展同时也受改革开放的时代因素影响。一方面，人们语言生活的接触面增加，例如“录像机”、“录像带”等词在后五年文献语料中的关注度极高；另一方面，语言生活的国际接触面变广，催生了语言生活中新的教育需求，“HSK”、“二语”等作为新增高频词出现在后五年的文献中，这与汉语国际教育的兴起有直接联系。众多大学开设汉语国际教育学院，并逐渐形成为一个独立的学科，有《世界汉语教学》、《语言教学与研究》等期刊作为高质量学术交流平台，涌现出大量的基础性工作，奠基了该领域理论研究的基础。

1991~2000 年间，模型预测呈现负偏差，即预测值相对于语言学领域期刊的实际词汇增长情况低 9908 个词种。该段时间的语言生活可视作上一阶段的领域持续深化，主要为学术生活发展与教育需求强化。该时间段汉语语言学理论研究与国际语言学前沿接轨，部分基础理论的介绍已跟进国外较新的研究成果。例如上一时间段已有对生成语法的介绍，“Chomsky”在该年份间就出现在新词检测的结果中，进一步的数据调阅则显示已有基于 Chomsky 提出的最简程序（Chomsky, 1995）解释汉语中“得”字句的工作（杨寿勋, 1998）。同时改革开放深化的影响持续体现在语言生活中，“市场经济”、“韩国”、“流行语”位列前五年词频前三的新词，体现语言生活中的新现象与大环境的改变；后五年中，“因特网”、“互联网”等极富时代特点的词出现，表现出新的沟通方式扩展了语言生活的媒介。同时 1993 年《中国教育改革和发展纲要》、《关于重点建设一批高等学校和重点学科点的若干意见》等教育领域的重大文件相继发布，随即这一时代的新词中包括“CET-4”、“EFL”等与我国高等教育密切的词，高校培养新时代高等人才的迫切需求推动了在大环境下对英语二语教育相关的研究，表明语言生活中国民高等教育的进一步发展。

2001~2010 年，模型预测值比观测值低 927 个词种，预测误差开始出现由负转正的趋势，现实观测值增长速率有所放缓，但模型预测值依旧偏低。主要原因是语言生活与时俱进使语言生活的内容形式不断丰富，学界发展仍有诸多新话题可供探究，但经过前阶段的介绍和引入，整体学科体系与基础理论趋于完善。随着加入世贸组织带来的更多机遇与挑战，语言生活中的诸多社会现象受到语言学界关注，如“WTO”、“农民工”分别作为前、后五年词频最高的新词出现，从“女性主义”、“以人为本”等词也同样可见一斑。而随着语言生活的沟通媒介不断丰富，“手机”、“博客”、“聊天室”等关键词的出现表明已有一部分工作将新时代的语言

模式作为研究对象。另一方面，就语言生活的研究方法而言，现代化分析手段被引入，如“SPSS”、“体验式”、 “语料库”与“PPT”等。与语言教学特别是二语教学相关的研究也进一步发展，尤其是标准语言测试的主题增加，英语等级考试（CET）相较于前十年词频上升，并得到持续关注，“TEM-4”作为新的专业英语等级考试也开始被取材为研究对象；二语领域新增词汇术语也频繁出现在该时代的文献中，如 “二语词”、“ 二语朗读”等。

2011~2020 年，该阶段的发展相较于前两个十年的增长速率放缓，模型的预测高于观测值 7981 个词种，可能是由于前二十年较多概念与观点的介绍，使我国语言学研究打下了理论基础。从这个时期出现的新词来看，此时语言生活相伴发展的时代特征较为突出，透过新词频率排行可发现，该阶段新媒体媒介的使用对语言生活的影响较大，例如“翻转”、“MOOC”、“微课”等教学工具的引入体现了新时代对语言教学方法的思考、而“大数据”、“微博”等词则体现了对于大数据时代对语言生活新环境下的网络用语关注及研究方法上对语料库范式的利用。而随着语言生活领域的扩展，政治、医疗等领域开始重视语言问题及探究语言生活与该领域的关联与影响，例如“习近平”、“十九大”、“领导力”等词关注了国家政治人物和政治用语，新型冠状病毒即使作为 2020 年的医疗新事件，在该时间段的文献中“新冠”也作为新词出现（12 次），在本语料库中，该词最早出现于《世界汉语教学》（田列朋, 2020）。

综上所述，模型能够从理论上预测一定时间段内语料库的词汇增长趋势，但现实中词汇如何增长会受到受到现实因素影响而呈现不同的趋势，语言学的词汇变化也是受语言生活与国家语言规划政策的影响而变化的通过预测偏差分析能够体现这这一进程，便于我们找出整体词汇增长趋势的关键节点，如图 5 所示。一方面，因为国家的语言政策直接影响着语言生活与语言的时代特征，对语言文字的研究活动也属于语言生活的范畴（李宇明, 1997），故语言学研究会受到国家政策的影响；另一方面，这种影响并非是单向的，在语言学的研究中，对语言生活的研究亦会影响国家的语言政策。“就语言生活为语言生活而研究语言和语言生活”的学术主张（李宇明, 2016）促成了中国语言生活派的形成，语言生活派的研究成果与关注点也为和谐语言生活的建立与语言学研究带来了不可忽视的影响。李宇明（2010a）对外语规划和公民外语素养的提升提出了思考和建议，在对应年份的文献高频词中便有二语教育的内容，正如前文中 2001～2010 阶段中新词内容所体现的那样，2013 年教育部亦启动《大学英语教学指南》的研制工作；李宇明（2010b）指出需要关注城市化进程的语言问题，并从语言规划层面着重考虑，同样是 2006～2010 年份区间段的研究热点。此外，《中国语言生活状况报告》（又称“中国语言生活绿皮书”）的逐年发布也能够为国家政策提供参考（李宇明, 2007），为学界、社会与国家语言规划之间建立起了桥梁。由此可见，词汇增长模型的误差分析对语言学发展的刻画不但能够大致勾勒所分析的区间段的热点，也能够体现出国家语言规划等对语言学研究的指引。

图 5 词汇增长模型预测误差分析框架图

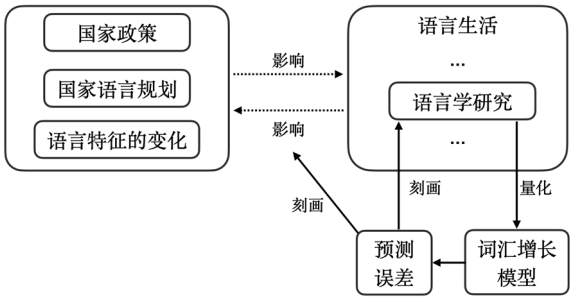
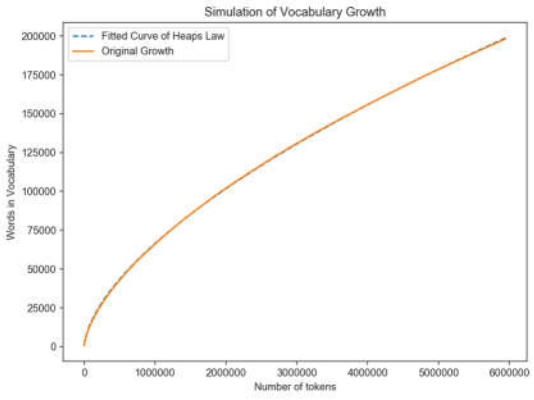


图 6 随机文本增长与预测曲线



6. 词汇历时变化的建模有效性验证

上一节在 Heaps 模型预测结果的基础上，从定量结合定性的角度分析了词汇实际增长情况，阐释了模型拟合的差异受社会发展、语言生活变化等因素的影响。上述分析基于模型的

理论预测值与现实观测值的误差进行分析，因此合理的模型理论预测值是保障研究结论正确的一大重要前提。本节将从定量角度进一步验证所使用模型的理论正确性，参考 Savoy（2015），本文采用随机文本的方法进行验证：即将文本以词语为单位随机打乱，使其失去原有的词汇分布特征和时间信息，从而得到一个新的学术文本语料库，该语料库既不具备时间序列上的增长关系，也不具备词语之间语义分布的关联，能够在消除外部影响因素的情况下，仅从语料库计量的角度验证模型是否能够成功拟合重新组织后的文本。对乱序后的文本，我们采用与 4.2 节一致的设定，每隔 1000 个词例进行采样，对采样点集合采用 Heaps 模型进行拟合，得到的参数为$k = 13.64, b = 0.61$，拟合曲线如图 6 所示，显示出 Heaps 模型在随机文本上较高的拟合精度。为了进一步探究该拟合结果，采用 Z-Score 指标进行拟合差异的评估，公式如下:

$$Diff（n） = V'(n) - V(n)$$

$$Z_{score} = \frac{Diff(n) - \mu_{Diff}}{\sigma_{Diff}}$$

图 7 和图 8 分别显示了随机文本与现实文本下不同采样点的 Z-Score 分布情况，两个文本的标准评分均处于（-2，2），即数据均在可接受的范围$[-3\sigma, 3\sigma]$内，符合正态分布，证明拟合模型所带来的差异是可以接受的。此外，随机文本中的预测误差显示出一定的随机性，而现实文本的预测误差 Z-Score 分布与上述年份的分析变化保持一致。就总体而言，现实文本差异的变化范围较随机文本大，而这种差异可能是由于现实因素所造成的，而非拟合模型的失误。

图 7　随机文本的预测误差 Z-Score 分布

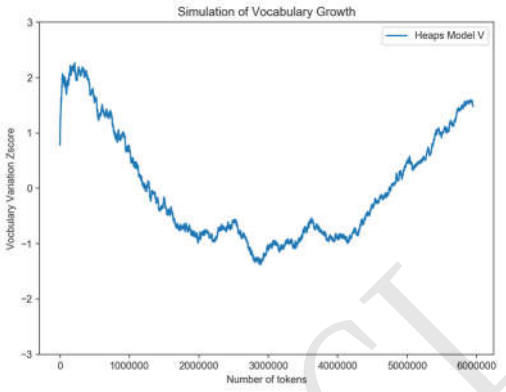
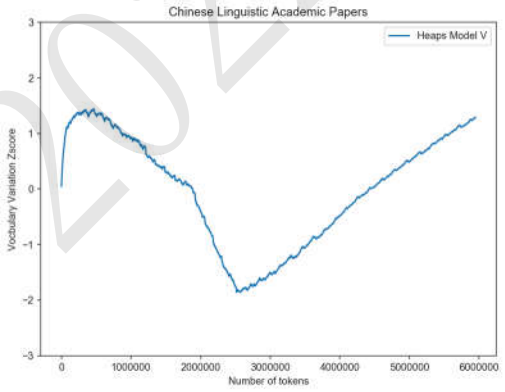


图 8　随机文本的预测误差 Z-Score 分布



通过在随机文本下的拟合实验，一方面证明前面实验中选择的模型能够较好地对语言学领域文本的词汇增长进行拟合，同时拟合的鲁棒性受到重大政治事件、经济变化、热点社会问题等现实因素的影响较小。另一方面，在实际拟合过程中由外部因素与学科发展所造成的预测误差是不可避免的，而无论定性还是定量角度均支撑了这些误差存在的合理性。

7. 结论

建国以来我国语言学历经 70 年的发展取得了瞩目成就，已有研究对语言学发展史上的重要事件进行了回顾和介绍，但尚未有使用量化手段对其历时演变的分析。为探究建国以来中国语言学的发展，首先，本文搜集 CSSCI（2019-2020）收录的中文核心期刊于 1957~2020 年间的摘要，构建了 5949428 词例、198241 词种、10813606 字的大规模语料库。第二，在该大型语料库基础上，本文采用三种词汇增长模型作为建模手段进行拟合，其中 Guiraud 模型（Guiraud, 1954）利用语料库中词种与词例的平方比值关系进行建模，Heaps 模型（Heaps, 1978）则基于双对数空间下二者的线性关系，Hubert-Labbe 模型（Hubert & Labbe, 1988; Labbé *et al.*, 2004）在 Heaps 模型（Heaps, 1978）基础上融入了对词频高低的考量。第三，本文选取拟合误差较小的 Heaps 模型对拟合结果以每 10 年为周期，借助量化统计结果与不同阶段的新词进行分析，研究结果体现出语言学发展与社会发展的关联性。根据模型的预测差值的变化趋势不同，可以看到我国语言学的发展经历了三阶段：起步阶段封闭停滞、改革开放后蓬勃发展、近十年逐步迈入自主创新阶段。这一结果符合实际发展趋势，从定性角度支撑了

本文所采取的研究方法的可行性。第四，本文打乱语料库的词序，使之失去时序信息，在随机文本上进行的验证程序证实了 Heaps 模型在探究中文词汇演化这一研究主题上的有效性与合理性。在未来研究中，我们将开展以下工作：1）根据本文收集的历时语料探索能够对未来学术文本增长趋势以及潜在话题进行预测的模型；2）根据现有研究框架进一步对语言学之外的人文社科其他领域以及科学领域核心期刊的词汇增长进行历时考察；3）利用已建立的学术汉语语料库，将提取语言学摘要的学术词表、核心词表等，为学术写作的研究和教学提供资源；4）构建英语学术论文摘要语料库，开展汉英学术论文摘要写作的对比研究，为一语和二语学术写作提供语料库驱动的教学建议。

致谢

本研究受国家语委（项目号：YB135-159）和澳门大学（项目号：MYRG2019-00013-FAH）科研项目资助。

参考文献

B·A·阿尔乔莫夫, 郑昌荣, 周毅. 1958. 苏联外语教学心理学四十年(1917—1957). 《外语教学与研究》(02), 129-140.

郭熙. 1999. 《中国社会语言学》，南京: 南京大学出版社.

郭熙. 2019. 七十年来的中国语言生活. 《语言战略研究》(04),14-26.

国家语言文字工作委员会. 2019. 《新中国语言文字事业发展 70 年纪事》，北京：语文出版社.

胡芳, 陈彧. 2005. 英汉学术期刊论文摘要衔接手段的对比研究. 《武汉科技大学学报(社会科学版)》(03), 83-86.

华萍. 1981. 评“暂拟汉语教学语法系统”. 《中国语文》(02).

李志君. 2014. 汉英学术语篇中作者身份指称语使用的调查与分析. 《外国语言文学》31(02), 81-89.

林学洪. 1958. 苏联的翻译事业. 《外语教学与研究》(04), 434-439.

刘丹青. 2019. 《新中国语言文字研究 70 年》，北京：中国社会科学出版社.

刘世沐. 1958. 必须坚决地清除语言学研究中的资产阶级立场观点和方法——评李赋宁同志“英国民族标准语的形成与发展”一文. 《外语教学与研究》(03), 261-265.

陆俭明. 1981. 分析方法刍议——评句子成分分析法. 《中国语文》(03).

陆俭明. 1999. 新中国语言学 50 年. 《当代语言学》(04), 3-5.

李宇明. 1997. 语言保护刍议. 《双语双方言》（五)，香港: 汉学出版社.

李宇明. 2007. 关于《中国语言生活绿皮书》，《语言文字应用》(1), 12-19

李宇明. 2010a.《关注中国城市化进程中的语言问题》.《中国语言生活状况报告 2009》，北京：商务印书馆。

李宇明. 2010b. 中国外语规划的若干思考. 《外国语（上海外国语大学学报)》(1).

李宇明. 2016. 语言生活与语言生活研究. 《语言战略研究》（3), 15-23.

刘锐, 王珊. 2017. 《两岸中文学术常用词对比研究》.《第三届国际汉语教学研讨会论文集》, 香港教育大学，香港.

罗常培. 1989. 《语言与文化》，语文出版社.

马费成. 2000. CSSCI 与社会科学评价. 《南京大学学报(哲学.人文科学.社会科学版)》(04), 155-160.

苏新宁. 2012. 中文社会科学引文索引(CSSCI)的设计与应用价值.《中国图书馆学报》38(05), 95-102.

田列朋. 2020. “战疫语言服务团”用语言利器助力抗疫. 《世界汉语教学》 34(02), 231.

王鸿斐. 1958. 科学研究一定要政治挂帅. 《外语教学与研究》(04), 422-423.

王珊, 陈钊, 张昊迪. 2021. 《近十年来澳门的词汇增长》. 《第 20 届中国计算语言学大会论文集》, 内蒙古大学，内蒙古.

王珊, 王会珍. 2021. 中文词汇增长研究. 《中文信息学报》, 35(1), 17-24.
王笑然, 王佶旻. 2022. 经贸类本科专业学术汉语词表研究. 《语言教学与研究》(04), 9-19.
吴格奇, 潘春雷. 2010. 汉语学术论文中作者立场标记语研究. 《语言教学与研究》(03), 91-96.
徐烈炯. 1988. 《生成語法理論》, 上海外語教育出版社.
徐世荣. 1957. 北京语音音位简述. 《语文学习》(08), 22-24.
杨寿勋. 1998. “得”的生成语法研究. 《现代外语》(01), 52+51+53-73.
张赫, 李加, 申盛夏. 2020. 学术汉语的词汇使用特征研究. 《语言教学与研究》(06), 19-27.
赵永青, 薛舒云, 邓耀臣, 徐建伟, 丁科家.（2019). 同一期刊论文英汉摘要作者立场信息一致性研究. 《解放军外国语学院学报》 42(03), 73-81+160.
中华人民共和国第一届全国人民代表大会. 1958. 《中华人民共和国第一届全国人民代表大会第五次会议关于汉语拼音方案的决议》. http://www.moe.gov.cn/ewebeditor/uploadfile/2015/03/02/20150302165814246.pdf
朱宇, 胡晓丹. 2021. 汉语连词在不同学术语域的聚合：多维度定量分析. 《语言教学与研究》(02), 57-69.
邹志仁. 2000. 中文社会科学引文索引(CSSCI)之研制、意义与功能. 《南京大学学报(哲学.人文科学.社会科学版)》(04), 145-154.
Chomsky, N. 1995. *The minimalist program*. Cambridge: The MIT Press.
Guiraud, P. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*: Presses universitaires de France.
Heaps, H. S. 1978. *Information retrieval, computational and theoretical aspects*: Academic Press.
Hoover, D. L. 2003. Another perspective on vocabulary richness. *Computers and the Humanities, 37*(2), 151-178.
Hubert, P., & Labbe, D. 1988. A model of vocabulary partition. *Literary and Linguistic Computing, 3*(4), 223-225.
Hyland, K. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies, 7*(2), 173-192.
Labbé, C., Labbé, D., & Hubert, P. 2004. Automatic segmentation of texts and corpora. *Journal of Quantitative Linguistics, 11*(3), 193-213.
Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. *CoRR, abs/1906.11455*.
Manning, C. D., Schütze, H., & Raghavan, P. 2008. *Introduction to information retrieval*: Cambridge university press.
Mellor, A. 2010. *Automatic essay scoring for low level learners of English as a second language.*（Ph.D.). Swansea University（United Kingdom), Ann Arbor.
Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis*: Bloomsbury Publishing.
Savoy, J. 2015. Vocabulary growth study: an example with the State of the Union addresses. *Journal of Quantitative Linguistics, 22*(4), 289-310.
Swales, J. 1985. *Episodes in ESP: A source and reference book on the development of English for science and technology*（Vol. 1): Pergamon.
Wang, S., Liu, X., & Zhou, J. 2022. Readability is decreasing in language and linguistics. *Scientometrics*. doi:https://doi.org/10.1007/s11192-022-04427-1
Wang, X. 2014. The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*.
Yu, G. 2010. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics, 31*(2), 236-259.