美国语料库语言学百年 *

许家金

(北京外国语大学中国外语与教育研究中心,北京 100089)

摘 要: 一般认为, 语料库语言学盛于英国及欧洲大陆, 美国语料库语言学则等而下之。然而,《教师一万词词汇手册》(Thorndike 1921)、《美国英语语法》(Fries 1940a)、布朗语料库 (Francis & Kučera 1964)、MDA 多维语域变异分析法(Biber 1988)、COCA 语料库(Davis 2010)和多因素分析(Gries 2003)等成果的涌现,不禁会推翻此前的固有印象。美国语料库语言学几乎在各个时期都有领先世界的成果。美国语料库语言学的百年发展史给我国语料库研究带来一个重要启示:若仅把语料库研究视作一门技术或研究方法,其学科地位和理论贡献将难以保证。

关键词:美国语料库语言学;技术创新;理论贡献;学术史

中图分类号:H030 **文献标识码**:A **文章编号**:1005-7242(2019)04-0001-06

0. 引言

长久以来,语言学和诸多其他领域一样,有幸把前沿科技应用到研究当中。20 世纪 40 年代末至 50年代初, 意大利耶稣会士 Roberto Busa 就开始用计算机编写拉丁文《托马斯著述索引》(*Index Thomisticus*)。Busa 得到 IBM 公司技术支持,用计算机将 1100 万拉丁词 Saint Thomas Aquinas 作品编制成索引行(Busa 1950)。美国学者 Ellison(1957)也借助计算机编纂完成英文《尼尔逊校订标准版圣经完全索引》(*Nelson's Complete Concordance of the Revised Standard Version Bible*) ①。这两项早期语料库课题催生了所谓的 "数字人文学"(Digital Humanities)。而在 20 世纪 60 年代,语言学领域也出现了具有开创性的布朗语料库(Brown Corpus)(Francis & Kučera 1964)。这一取样严谨的美国书面英语语料库已成为现代语料库语言学的基石之一。

尽管美国语料库语言学起步较早,但学界(Simpson & Swales 2001:2)一般认为语料库语言学盛于英国及欧洲大陆,例如,英国的兰卡斯特大学、伯明翰大学、伦敦大学学院, 以及欧洲大陆的一些国家, 如比利时、德国、意大利、荷兰和斯堪的纳维亚半岛国家。美国语料库语言学则等而下之。美国语料库语言学未入主流,跟半世纪以来生成语言学的盛行不无关系。20 世纪 50 年代以来, 在美国流行生成语言学的同时,英国和欧洲的语料库语言学得到长足发展。然而,美国的语料库语言学在整个学科领域扮演了十分重要的角色,很多研究都极具开创意义。美国语料库语言学的历史比我们想象的要精彩得多。以下我们将对美国语料库语言学历史加以详述。

本文中我们不把电子化视作语料库语言学的必要条件。只要某项研究基于真实语言数据做了有代表性的取样和量化分析,我们就将其视作语料库研究。

1. 前电子化时代

本节将前电子化美国语料库语言分为三个阶段, 即 20 世纪一二十年代、20 世纪 20 年代到 40 年代、20 世纪 40 年代到 60 年代。

1.1 20 世纪一二十年代:Ayres(1913)拼写研究及其他词表研制项目

美国的量化文本分析可追溯到 Sherman (1893)的《文学分析法》(*Analytics of Literature*),该研究采用统计方法区分了散文和诗歌的文体特征。例如,Sherman (1893:259)统计了 6 位作家作品的平均句长(ASL):Fabyan(63.02)、Spencer(49.82)、Hooker(41.40)、Macaulay (22.45)、Channing(25.73)以及 Emerson(20.58),发现了某种"无意识的相似句长",即同一位作者的不同作品中平均句长大抵相似。因此,Sherman 指出平均句长可以作为辨别文体风格的有效指标。类似指标还包括:动词的使用、小句的衔接、从属和并列关系等。

从 20 世纪第一个十年开始,许多成规模的(前电子化)美国语料库语言学项目开始蓬勃发展。这些早期语料库研究主要用于促进语言教学。

这一时期的代表性研究是 Ayres(1913)的词频调查。该研究选取了 2000 篇私人和商务信函,总语料规模达 110,160 词。据此,Ayres 得到常用词表,例如,"I (1080 次)""the(918 次)""and(697 次)""you(635次)""to(627 次)""your(585 次)……"。该词表的编制旨在用于指导拼写教学。Ayres(1913:7-8)甚至还提供了一些语用相关的词频表,例如,Ayres 将信函称呼语以词频降序列出:"Dear(520)""Dear Sir(490)""My Dear-(476)"

* 本文系教育部人文社会科学重点研究基地重大项目子课题"大数据视野下的外语及外语学习研究"(17JJD740003)的阶段性成果。

"Gentlemen (207)""DearMadam (168)""Miss-(28)""Dear Miss-(17)""Dear Friend(17)""Dear Sirs(14)"等。

Ayres(1913:10)发现美国全国教育协会(NEA)规定的414个拼写词汇和日常生活中人们实际使用的词之间有巨大的差异。在上述信函词频研究中,当时美国全国教育协会(NEA)的414个拼写词汇中有70%的词甚至"从未出现过"。因此,Ayres(1913:10)指出"我们不能坐在书桌前,单凭直觉就决定人们应掌握哪些词的拼写,我们必须知道人们日常需要掌握哪些词的拼写。"

此外,Cook & O'Shea(1914:226-227)对比了13位成人所写的20万字家书和专家编写的3本畅销英语拼写课本。他们发现拼写课本中只有70%的词汇出现在了家书当中。Cook和O'Shea认为,这些拼写课本显然没有把重点放在常人最需要掌握的词汇上。有趣的是,Cook和O'Shea同时还研究了13位成人书写用词的差异,家书和其他信函的用词差异,以及拼写词汇的性别差异。例如,"appetite""candy""apron""hair"这样的词多出现在女性书写的信函当中,而"argument""defeat""administration""convention"这样的词多出现在男性书写的信函当中。

Jones(1915:4-6)调查了2到8年级共1050名学生所写的1500万词作文语料,得到了4532个不同的单词,即今天语料库术语中的"类符"。Jones同时还对不同学业水平中涉及的拼写词汇进行了定级。

Andersen(1921)认为,自从20世纪伊始,教育学的科学研究取得了重大进展,对于教学成果的客观评价是其中一项重要成就。在所有相关课题当中,拼写实验的数量最多。他对20世纪前20年的词汇拼写项目做了全面回顾。Andersen对Ayres(1913)的取样标准持怀疑态度,于是在他的拼写研究课题中,他从更广泛的体裁中选取了361,184词语料,其中涉及的群体包括:专业人士、商务人士、家庭成员、混合群体、友人、医生、农民、银行要员和汽车零售商。

20世纪20年代以前的词频研究,多半受到Ayres(1913)信函研究的影响。这一阶段研究主要聚焦于拼写问题,关注的是中小学生的读写能力。在下一阶段,我们见到的研究不光聚焦于中小学生,也关注成人的语言学习。此外,与外语学习相关的语料库研究开始占据突出位置。

1.2 20世纪20到40年代:Thorndike(1921)、Fries(1940a)等

20世纪20年代,国际英语教学界兴起了"词汇控制运动"(参见Hornby 1953)。到了20世纪30年代,词汇控制的理念更广泛地应用于教学;除了用于教授拼写,还应用于教授阅读和写作、课程大纲设计、教材研发和评估。在理念上主张教授高频词汇的学者主要分为两类:偏爱主观方法的学者,其中包括Charles Ogden和I. A. Richards(在英国)、Harold Palmer和A. S. Hornby(在日本)、Michael West(在印度和加拿大)。这些学者多依赖个人直觉选取"基础词汇"(Basic Vocabulary)(Ogden 1930)。"基础词汇"中的"基础"(Basic)一词来源于"British, American, Scientific, International, Commercial"的缩写。类似的研究还包括"最低要求词汇"(West 1931)和通用词表(West 1953)。

另一类学者偏爱以客观视角开展词汇控制,即采用量化方法来寻找最低要求词汇。他们主要为美国学者。例如,Thorndike(1921)的1万词词汇手册便是这一时期基于大规模真实文本制作量化词表的先驱之作。此后,又出现了各类习语频率表、句法频率表和语义频率表。与20世纪20年代之前的研究相比,这一时期的量化研究在语言描写的各个层面和研究方法上都更胜一筹。

在1931年至1944年间,Thorndike通过扩充语料将1921年的1万词手册增加到了2万词,随后又扩展到3万词。美国的其他教育研究者,受Thorndike的启发,编制了一系列其他语种的词频表,其中包括法语(Henmon 1924;Vander Beke 1929)、西班牙语(Buchanan 1927)、德语(Morgan 1928)和巴西葡萄牙语(Brown, Carr & Shane 1945)。学者们不仅把研究扩展到其他语种,同时开展了对法语(Cheydleur 1929)、西班牙语(Keniston 1929)、德语(Hauch 1929),以及巴西葡萄牙语(Brown & Shane 1951)的习语量化分析,并对西班牙语(Keniston 1937a)和法语(Clark & Poston 1943)句法范畴进行了统计。遗憾的是,并没有出现类似的英语短语和句法统计成果。

Eaton(1934,1940)基于现有的词频统计整理了英语、法语、西班牙语和德语中最常用的词汇,教授法语、西班牙语和德语的老师可以将这个词表用于教学材料的编制和课堂教学。Eaton(1934,1940)将这几种语言和世界语进行平行对齐,使得这份整合词表具有了语义词频表的价值。

Lorge(1937,1949)基于《牛津英语词典》(*The Oxford English Dictionary*)中的义项编制了英语语义频率表。Lorge的语义频率统计后来编入了广为人们引用的《通用英语词表》West(1953)。该通用词表后来被用作词汇分析软件Range和AntWordProfiler默认的基础词表。

Fife(1931:188-207)和Fries(1940b)曾对这一发展阶段做过详细回顾,讨论了这类研究项目对于教学的重要意义。这一阶段美国学者的相关研究已初步具备了现代语料库语言学研究的特征。

首先,几乎所有这一时期的研究都考虑到了语料的代表性。Keniston(1929:4-8)的研究就是其中之一。他选取的文章体裁包括戏剧、小说、各类杂文散文、报

刊、科技文本。这样的取样标准很容易使人们联想起布朗语料库的文本分类：新闻、通用、小说和学术。Keniston 的语料库同时还涵盖了西班牙语不同的区域变体，西班牙和拉丁美洲（约占 1/5）作家的作品都涵盖在内。有趣的是，尽管 Keniston 没有采用今天意义上的语料库语言学术语，却详细描述了真实性、人口学代表性、科学取样和体裁/情境多样性等概念。

其次，在文本筛选和统计当中，人们始终遵循分布原则。这项原则在许多文章的副标题中得以标明。在许多研究的导论部分，同频率（frequency）原则一起，常常会出现 “range”“distribution”“widely used”“units” “sources”等术语常常用于阐释分布原则。Keniston（1929：4）认为，在此前的词频表中，一些罕用词只出现在某个特定文本中，因此得到的频数并不能反映它们在语言中的全貌。因此，Keniston 把自己著作的副标题列为了“基于分布和频数”，而没有把自己的副标题列为“基于频数和分布”，正是为了强调分布比频数更重要。因此，在不同的篇章和文本体裁中，词语、习语和句法统计既包括（相对）频率，又涵盖了分布数据。此外，Fries（1940a：6–9）呼吁我们关注语言特征在历时、不同地域、书面语和口语、社会和阶层方面的量化差异。

另外，在语言层面，习语和句法的研究使得频数统计不再局限于词汇层面。这些习语既包括那些变化形式受限的习语、语义透明的表达和经常共现的词语组合。西班牙语、德语和法语的习语表均于 1929 年出版，时间远早于 Palmer 在 1933 年出版的英语搭配研究报告。此外，这些习语表基于大量的真实文本编制而成，并提供了统计数据。然而，Palmer 的《英语搭配中期报告第二辑》（*Second Interim Report on English Collocations*）仅提供了短语清单，并未说明这些短语是否取自真实文本，也没有提供量化信息。西班牙语和其他语种的频数统计涵盖了复合连接词、复合介词和不及物动词等语言项目，这些自然而然地过渡到了语法结构的量化描述。Keniston（1937a）的《西班牙语句法表》（*Spanish Syntax List*）和 Clark & Poston（1943）的《法语句法表》（*French Syntax List*）沿用了相同的分布和频率原则来对两种语言中的语法结构进行全面的量化分析。除了句法表，Stormzand & O’Shea（1924）还开展了学习者英语的对比研究，考察了成人和在校学生特定语法结构 “使用过多或使用不足的现象”（即语料库语言学术语中的 “多用” 和 “少用”）（Stormzand & O’Shea 1924:48）。Stormzand 和 O’Shea 还比较了不同学业水平的学生数据，从而评估他们的句法能力发展状况。Lorge 的创新之处就在于他的词频研究不仅限于语言的形式和结构，而且还深入到了词义层面。

再有，这一时期几乎所有的量化研究，都有一定的应用语言学或语言学的研究目标。例如，几乎所有的频率表项目都尝试解决语言教学中的问题。此后还有许多语言教学上的应用。例如，最早的学习词典之一——《桑代克世纪中学生词典》（*The Thorndike–Century Junior Dictionary*）（Thorndike 1935）。Thorndike 将 20,000 词分成了 20 级，每个级别 1000 词，从第一个 1000 词（例如 be...1）到第二十个 1000 词（例如 authorization...20）进行标号。该词典的义项并非按照历史发展顺序安排，而是采用了由常用到罕见、由简单到困难的排列顺序（Thorndike 1935：iv）。值得注意的是，除了语言教学外，这一时期还有一些量化研究，重点关注语言本体特征。例如，Zipf（1935）对汉语、拉丁语、美式英语的音素、词频和词长、词义进行了统计和对比，从而得出了著名的“齐夫定律”。Keniston（1937b）基于 300,000 词的语料库对十六世纪卡斯提尔语（西班牙北部和中部的一种方言）句法进行了研究，开创了以量化方法研究历史句法的先河。Fries（1940a：154–159）的研究与 Keniston 的跨领域语法比较不同，描绘了英语的历时语法变化。例如，1560 年到 1920 年间，第一、二、三人称代词与 shall 和 will 的搭配用法。Fries（1940a：111）同时还比较了标准英语和方言英语中动词与介词/小品词的搭配状况。Fries（1952）还录制并转写了美国中北部居民的标准英语对话，总规模达 250,000 词。他基于这些真实口语语料编写的英语语法，要比 Quirk et al.（1972）的 “英语语法调查”（the Survey of English Usage）早了 20 年。

1.3 20 世纪 40 到 60 年代：从前电子化时代到电子时代的过渡

从 20 世纪 40 到 60 年代，量化语言研究延续了先前的研究范式。虽然这 20 年间美国并没有出现先前不同语种的词频项目，但是文体和语言风格的量化研究，以及 Harris 提出的具有语料库语言学特色的相关理论假设，特别值得关注。

Whatmough（1956）采用量化方法研究了希腊语和拉丁语中诗歌和科学文体中的风格差异。研究发现，词频和句长是区别作者风格的重要指标。Carroll（1960）在研究文体风格时较早采用了因子分析法。他基于 39 个语言特征和 29 个人们评价文体风格的指标计算出了评价散文文体风格的 6 个维度（总体风格评价、个人情态、修饰、抽象程度、文体正式程度和特征）。这在很大成度上是 Biber 多维分析的先声。

20 世纪 50 年代 Harris 提出了一系列理念和假说，为美国语料库语言学奠定了理论基础。遗憾的是，他的很多思想并没有得到后期语料库语言学家的广泛采纳。

Harris 很可能是最早把 corpus（Harris 1947：175）作为独立的语言学术语使用的学者：

When such comparisons are carried out for a large corpus, we obtain morphemic segments which are

repeated in various environments throughout the corpus.

在 Harris 的众多理论假设中，“分布假说”（the distributional hypothesis）对于语料库语言学有着重要的理论意义。Harris 曾在讨论形态理论时指出（Harris 1947：156–157）：

If we consider *oculist* and *eye–doctor* we find that, as our corpus of actually–occurring utterances grows, these two occur in almost the same environments (selection), ..., we say they are synonyms.

在 Harris 的分布理论框架中，“序列依存”（serial dependence）或“共现”（co–occurrence）原则类似于 Sinclair（1991）的习语原则；不同语言单位的“平行替代”（parallel substitutability）原则与 Sinclair（1991）的开放选择原则相似。

尽管 Harris 的分布假说及相关概念，并没有发展成为语料库语言学理论。但上述概念却奠定了“分布语义学”（distributional semantics）的基础，在自然语言处理中广泛应用。SketchEngine 中的同义词功能就是一个我们熟悉的应用实例。潜在语义分析、基于词向量模型的语义相似度分析和主题建模（topic modelling）等都基于这一假说发展而来。

2. 计算机时代的美国语料库语言学

2.1 语料库建设

20 世纪 60 到 70 年代，美国语料库语言学有所停滞。这很大程度上可能受到了迅速崛起的乔姆斯基生成语法的影响。乔姆斯基旗帜鲜明地反对把自然语料作为语言学的研究对象（Chomsky 1957：14–7，97）。

然而，在语料库建设方面，美国异军突起。由于计算机在美国学术界较早普及，第一个电子化通用英语语料库——布朗语料库——在布朗大学应运而生。该语料库存储在穿孔卡片（punched cards）上，研究者需要借助大型计算机方可使用。布朗语料库的出现使得美国语料库语言学自此得到国际认可。布朗语料库代表了 20 世纪 60 年代早期的美国书面英语。其取样方案（15 个文类，可细分为新闻、通用文章、小说和学术四大体裁，或信息型文本和想象型文本两大类）成为后来英国英语语料库 LOB、英国国家语料库 BNC 等项目取样的重要参照。

进入到 21 世纪，美国语料库研制的势头愈加迅猛。其中翘楚是杨百翰大学 Mark Davis 开发的 COCA 等大型免费在线语料库（Davis 2010），近年更是建成了超百亿词次的 iWeb 大数据语料库。杨百翰大学的系列语料库不仅规模大、界面友好，还特别关注语域、年代等的对比功能的实现。杨百翰大学语料库的网页界面最初的缩写是 VIEW（Variation in English Words and Phrases），正是为了突出这一特色。杨百翰大学系列语料库中最具影响的仍然是 COCA 语料库，其规模每年扩充 2000 万词，新增的文本等比取自口语、小说、流行杂志、报纸和学术文本。这使得 COCA 语料库成为第一个大规模历时平衡语料库，是研究语言变化的极好资源（Davis 2010：453）。在教学领域，Davis 等人基于 COCA 编纂了频率词典。参照 COCA 频率词典的做法，阿拉伯语、捷克语、荷兰语、法语、德语、日语、韩语、韩语、汉语、俄语、西班牙语、土耳其语频率词典相继问世。与此同时，大量专用语料库陆续建成。例如，托福 2000（TOEFL 2000 Spoken and Written Academic Language，简称 T2K–SWAL）语料库被用于描述和分析英语教学语境下的学术口语和书面语的特点（Biber et al. 2002）。另外值得一提的是密歇根学术英语口语语料库（Michigan Corpus of Academic Spoken English, MICASE）和密歇根高级学生作文语料库（Michigan Corpus of Upper–level Student Papers，MICUSP）。这两个语料库继承了 Charles Fries 在 20世纪四五十年代在密歇根大学基于真实口语和书面语料库所开展的语言描写和教学应用的实证研究。

2.2 技术创新

自计算机问世以来，美国学者为语料库技术进步做出了巨大贡献。例如，我们熟知的“语境中的关键词”（KWIC）索引行分析方法，就是由美国学者 Luhn（1960）提出的。平行索引分析工具则最先由美国计算语言学家 Church & Gale（1991）基于加拿大议会官方英法平行议会议记录研制。

另一个语言教育领域的重要在线文本分析工具 Coh–Metrix 3.0 提供 106 个词汇、句法、篇章和语义分析指标。该工具已广泛用于文本可读性分析和学习者写作评价（Crossley & McNamara 2011）。Kyle & Crossley（2015）研发了文本分析工具，几乎涵盖 Coh–Metrix 3.0 的各项指标，并添加很多新的功能。更重要的是，Kyle 开发的相关工具使得用户能够在本地电脑上批量处理语料文本。

20 世纪 90 年代以来，计算语言学开始出现统计转向（参见 Armstrong 1993），自然语言处理领域陆续出现能为语料库研究所用的技术和资源。其中包括词性赋码工具、句法分析器、命名实体识别工具、词汇网络（WordNet）、框架网络（FrameNet）、情感分析工具谷歌神经机器翻译系统等。这些技术最早都由美国的计算语言学者开发，并对语料库研究产生重大影响。

2.3 理论贡献

就理论创新而言，Biber（1984，1988）基于语料库研究语域变异的多维分析法，改变了美国语料库语言学的面貌。他的研究方法已成为美国语料库语言学中最流行的分析方法，成为了美国语料库语言学的标识。语域变异研究在教学方面的应用突出表现为《朗文口笔语语法》（*Longman Grammar of Spoken and Written English*）（Biber et al. 1999）是对 Quirk 语法在语料库时代的更新换代之作，已成为十分重要的教学语法参考书。该书的

最大特色是以图表形式展示了语法项目在不同语域中的使用差异。多维分析法的提出很大程度上受到了社会语言学变异研究和多元统计分析的影响。Biber (1984)借助布朗语料库和LLC 语料库尝试了语域变异的多维分析，并通过《口语和书面语间的变异》(*Variation across Speech and Writing*)(Biber 1988)一书，迅速确立了多维语域分析的国际地位。Biber 学术生涯中大部分时间在亚利桑那州弗拉格斯塔夫市的北亚利桑那大学度过，因此其弟子将其开创的研究传统称作“弗拉格斯塔夫学派”(Flagstaff School)(Cortes & Csomay 2015: xv)。多维语域分析法广泛应用于学术写作和口语话语、网络语域、学习者话语、课堂互动话语等。

美国语料库语言学的另一项重要进展是将多因素分析(multifactorial analysis)运用于认知语言学研究，其中最具代表性的学者是 Stefan Gries。在研究当中，Gries 等人基于词法、句法、话语、语用等多个层面的语言特征，利用多因素分析法研究了不同因素对于特定语言构式选择的影响。其中常见研究选题包括构式交替现象(例如，与格交替、属格交替、动词补语交替)，语法标记的选择(例如，补语标记词 that 使用与否的问题，can 和 may 选用问题)，以及语言特征位置变化的动因分析(例如，英语当中多个形容词的排列顺序问题，时间副词从句置于主句前还是主句后的问题)。这一研究方法与基于用法的语言学研究和构式语法密切关联，既可用于理论语言学研究(参见 Bresnan & Ford 2010)，也适用于二语习得研究(参见 Gries & Ellis 2015)。

3. 结语

本文梳理了美国语料库语言学的百年历史，特别是其早期历史。从中可以看出，一个世纪以来，语料库研究者的研究目的始终未变，即借助大规模真实语言素材揭示语言使用的概率性特征，从而回答语言本体问题或指导语言教学实践。真正发生变化的是存储和提取语料信息的工具和方法。例如，从最初的纸质索引卡片，发展到穿孔卡片格式的电子语料库，再到微型计算机语料库，乃至如今大数据背景下的云端语料库。

由前文综述可见，与英国和欧洲语料库语言学相比，美国语料库语言学在各个历史时期都是语料库创新技术的引领者。同时，美国语料库语言学方法创新也不逊色。其创新领域主要体现为：语料库研制、统计算法、计算语言学工具和各种基于语料库的实证语言研究方法。例如，学者们经常把布朗语料库视作第一个电子化的英语语料库；当前全球范围内引用最多的语料库是 COCA 语料库；Biber 所倡导的多维语域变异分析法是目前基于语料库的体裁分析中最有效的研究路径。

与此同时，在理论建树上，美国语料库语言学也贡献显著。如今的美国语料库语言学主体源于后布龙菲尔德结构主义语言学(post-Bloomfieldian structuralism)。譬如，Harris 提出的很多具有语料库语言学理论高度的概念和假说，以及 Fries 倡导的基于语料库的语言教学理念，都是结构主义路径的直接表现。而 Biber的多维分析法则根植于 Hymes 等学者开创的社会语言学和话语分析思想，并发展成体裁和语域研究领域举足轻重的理论和方法；Gries 的多因素分析思路，力图将功能语言学和认知语言学理论与多变量统计方法充分结合，在复杂统计方法和深厚语言学理论之间取得最佳平衡。

美国语料库语言学的百年发展史给我国语料库研究带来一些重要启示：我们应兼顾技术创新和理论深化，两者不可偏废。若仅把语料库研究视作一门技术或研究方法，其学科地位必将难以稳固。

注释：

① Ellison(1957)出现的时间早于布朗语料库，但并非服务于语言学目的的通用英语语料库。美国第一个非英语的、取样均衡的电子化语料库是 Juilland & Chang-Rodriguez(1964)50 万词的西班牙语语料库，该语料库始建于 1956 年。

参考文献：

Andersen, W. 1921. *Determination of a Spelling Vocabulary Based upon Written Correspondence*[M]. Iowa City: University of Iowa.

Armstrong, S. (ed.). 1993. *Using Large Corpora*[C]. Cambridge, Massachusetts: The MIT Press.

Ayres, L. 1913. *The Spelling Vocabularies of Personal and Business Letters*[M]. New York City: Division of Education, Russell Sage Foundation.

Biber, D. 1984. *A Model of Textual Relations within the Written and Spoken Modes*[D]. Ph.D. Dissertation. University of Southern California, Los Angeles, CA.

Biber, D. 1988. *Variation across Speech and Writing*[M]. Cambridge: Cambridge University Press.

Biber, D., S. Conrad, R. Reppen, P. Byrd & M. Helt. 2002. Speaking and writing in the university: a multi-dimensional comparison[J]. *TESOL Quarterly* 36(1): 9-48.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Spoken and Written English Grammar*[M]. London: Pearson.

Bresnan, J. & M. Ford. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English[J]. *Language* 86(1): 186-213.

Brown, C. & M. Shane. 1951. *Brazilian Portuguese Idiom List: Selected on the Basis of Range and Frequency of Occurrence*[M]. Nashville: Vanderbilt University Press.

Brown, C., W. Carr & M. Shane. 1945. *A Graded Word Book of Brazilian Portuguese*[M]. New York: F. S. Crofts & Co., Inc.

Buchanan, M. 1927. *A Graded Spanish Word Book*[M]. Toronto: The University of Toronto Press.

Busa, R. 1950. Complete index verborum of works of St. Thomas[J]. *Speculum: A Journal of Medieval Studies* XXV(1): 424-425.

Carroll, J. 1960. Vectors of prose style[C]//T. Sebeok (ed.). *Style in Language*. Cambridge, Massachusetts: The MIT Press: 283-292.

Cheydleur, F. 1929. *French Idiom List: Based on a Count of 1,183,000*

Running Words[M]. New York: The Macmillan Company.
Chomsky, N. 1957. *Syntactic Structures*[M]. The Hague: Mouton.
Church, K. & W. Gale. 1991. Concordances for parallel texts[R]. Paper presented at the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Oxford, England.
Clark, R. & L. Poston. 1943. *French Syntax List: A Statistical Study of Grammatical Usage in Contemporary French Prose on the Basis of Range and Frequency*[M]. New York: H. Holt and Company.
Cook, W. & M. O'Shea. 1914. *The Child and His Spelling*[M]. Indianapolis: The Bobbs-Merrill Company.
Cortes, V. & E. Csomay. (eds.). 2015. *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*[C]. Amsterdam: John Benjamins.
Crossley, S. & D. McNamara. 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing[J]. *International Journal of Continuing Engineering Education and Life Long Learning* 21(2-3): 170-191.
Davis, M. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English[J]. *Literary and Linguistic Computing* 25(4): 447-464.
Eaton, H. 1934. *Comparative Frequency List*[M]. New York City: International Auxiliary Language Association.
Eaton, H. 1940. *Semantic Frequency List for English, French, German, and Spanish*[M]. Chicago: University of Chicago Press.
Ellison, J. 1957. *Nelson's Complete Concordance of the Revised Standard Version Bible*[M]. New York: Thomas Nelson & Sons.
Fife, R. 1931. *A Summary of Reports on the Modern Foreign Languages*[M]. New York: The Macmillan Company.
Francis, W. & H. Kučera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-day Edited American English*[M]. Providence, Rhode Island: Brown University.
Fries, C. 1940a. *American English Grammar*[M]. New York: D. Appleton-Century-Crofts.
Fries, C. 1940b. *English Word Lists: A Study of Their Adaptability for Instruction*[M]. Washington, D.C.: American Council on Education.
Fries, C. 1952. *The Structure of English: An Introduction to the Construction of English Sentences*[M]. New York: Harcourt, Brace and Company.
Gries, S. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*[M]. London: Continuum.
Gries, S. & N. Ellis. 2015. Statistical measures for usage-based linguistics[J]. *Language Learning*, 65 (Supplement 1): 1-28.
Harris, Z. 1947. Structural restatements: II[J]. *International Journal of American Linguistics* 13(3): 175-186.
Hauch, E. 1929. *German Idiom List: Selected on the Basis of Frequency and Range of Occurrence*[M]. New York: The Macmillan Company.
Henmon, V. 1924. *A French Word Book Based on a Count of 400,000 Running Words*[M]. Madison: University of Wisconsin.
Hornby, A. 1953. Vocabulary control: history and principles [J]. *ELT Journal* 8(1): 15-21.
Jones, W. 1915. *Concrete Investigation of the Material of English Spelling*[M]. Vermillion: The University of South Dakota.
Juilland, A. & E. Chang-Rodriguez. 1964. *Frequency Dictionary of Spanish Words*[M]. The Hague: Mouton & Co.
Keniston, H. 1929. *Spanish Idiom List: Selected on the Basis of Range and Frequency of Occurrence*[M]. New York: The Macmillan Company.
Keniston, H. 1937a. *Spanish Syntax List: A Statistical Study of Grammatical Usage in Contemporary Spanish Prose on the Basis of Range and Frequency*[M]. New York: H. Holt and Company.
Keniston, H. 1937b. *The Syntax of Castilian Prose: The Sixteenth Century*[M]. Chicago: The University of Chicago Press.
Kyle, K. & S. Crossley. 2015. Automatically assessing lexical sophistication: indices, tools, findings, and application[J]. *TESOL Quarterly* 49(4): 757-786.
Lorge, I. 1937. The English semantic count[J]. *Teachers College Record* 39(1): 65-77.
Lorge, I. 1949. *The Semantic Count of the 570 Commonest English Words*[M]. New York City: Teachers College, Columbia University.
Luhn, H. 1960. Keyword-in-context index for technical literature[J]. *American Documentation* 11(4): 288-295.
Morgan, B. 1928. *German Frequency Word Book: Based on Kaeding's Häufigkeitswörterbuch der Deutschen Sprache*[M]. New York: The Macmillan Company.
Ogden, C. 1930. *The Basic Vocabulary: A Statistical Analysis*[M]. London: Kegan Paul, Trench, Trubner & Co., Ltd.
Palmer, H. 1933. *Second Interim Report on English Collocations*[M]. Tokyo: The Institute for Research in English Teaching, Department of Education.
Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1972. *The Grammar of Contemporary English*[M]. London: Longman.
Sherman, L. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*[M]. Boston: Ginn & Company.
Simpson, R. & J. Swales (eds.). 2001. *Corpus Linguistics in North America*[C]. Ann Arbor: University of Michigan Press.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*[M]. Oxford: Oxford University Press.
Stormzand, M. & M. O'Shea. 1924. *How Much English Grammar*[M]. Baltimore: Warwick & York, Inc.
Thorndike, E. 1921. *The Teacher's Word Book*[M]. New York City: Teachers College, Columbia University.
Thorndike, E. 1935. *The Thorndike-Century Junior Dictionary*[Z]. New York: D. Appleton-Century-Crofts.
Vander Beke, G. 1929. *French Word Book*[M]. New York: The Macmillan Company.
West, M. 1931. Notes, news and clippings[J]. *The Modern Language Journal* 15(8): 638-647.
West, M. 1953. *A General Service List of English Words*[M]. London: Longmans, Green.
Whatmough, J. 1956. *Poetic, Scientific and other Forms of Discourse*[M]. Berkeley: University of California Press.
Zipf, G. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*[M]. Boston: Houghton-Mifflin Company.

收稿日期:2019-02-09

作者简介:许家金,博士,教授。研究方向:话语研究,二语习得,语言对比与翻译,语料库语言学。

(**责任编辑**:梁婧玉)