

# 近百年西班牙语语料库建设与研究概述<sup>\*</sup>

赵冲 许家金

**摘要：**在语料库语言学发展史上，英语研究方面成果最为显著。然而，梳理近百年的西班牙语语料库发展史可以发现，西班牙语研究方面也是成绩斐然，并有其突出特点：首先，从发展进程看，西语电子化平衡语料库的研制早于英语，且以西语为对象的语料库研究在各个时期与英语相关研究总体同步；其次，在研究视角上，西语语料库研究始终关注地域、文体等方面的语言差异；最后，在机制建设上，当前的国家级西语语料库研制由权威语言学机构主导，可确保语言材料规范可信，平台系统维护良好，促进了语料库在学术研究中的广泛应用。这些可为我国语料库建设与研究提供有益参考。

**关键词：**语料库语言学；西班牙语；学术史

## 1 引言

一般认为，现代西班牙语始于 14 世纪后半叶（Abad Nebot, 2017）。1492 年伊比利亚半岛统一，西班牙语开始确立其主导语言地位。同年，欧洲第一本方言语法《卡斯蒂利亚语语法》（*Gramática de la lengua castellana*）出现，为西班牙语语言规范奠定了基础，也被视为西班牙语语言研究的开端（Nebrija, 1997）。1492 年，哥伦布远航美洲，西班牙语的使用范围也随之扩散。

20 世纪 90 年代以来，语料库语言学逐渐发展为语言学内部较有影响的一个分支。总体上，以英语为对象的语料库研究占据主导地位，而非英语语种为对象的研究则难以成为国际学术主流。然而，通过爬梳文献，我们注意到，在现代语料库语言学萌芽之初，以西语为对象的研究也有一些引领性的成果。

<sup>\*</sup> 本文系教育部人文社会科学重点研究基地重大项目“基于多语种语料库的外语及外语教育研究”（编号：22JJD740012）的阶段性成果。

最近三十年,西语语料库研究整体上进入快速发展期,建成了一系列不同种类、规模可观的西语语料库,同时还广泛采用前沿方法开展实证研究。从研究视角来看,西语语料库研究尤其关注西语在地域、文体等方面的语言差异。本文将概述西语语料库研究的特色及代表性成果,一方面为语料库语言学学科史研究补充史料,另一方面也为学界开展基于语料库的西语理论及应用研究或双语及多语对比研究提供历史性参考。

## 2 西班牙语语料库研究

本文所谈的语料库不以文本是否电子化为判定标准。正如 Francis (1992) 指出,在计算机发明之前,便已存在语料库。本文参考何中清和彭宣维 (2011)、许家金 (2017、2019) 等的观点,将所有基于大规模代表性文本开展的计量语言研究都视为语料库研究。

本文关注的是以西班牙语为研究对象的语料库研究,包括对西语本体和西语作为中介语、第二语言及传承语 (heritage language) 等的研究,也包括西语同其他语言的对比研究。本文不涉及西语国家学者以其他语言为对象的研究,也不讨论西语母语者的外语教学研究。以 Juilland & Chang-Rodríguez (1964) 这一西语最早的电子化语料库成果为界,本文将西语语料库研究划分为前电子化时期和电子化时期。

### 2.1 前电子化时期

同英语语料库研究类似,在“语料库”这一概念出现以前,西班牙语在宗教文献研究、词典学、语言教学研究等领域已展现出对真实文本中词语语境的关注 (Francis, 1992; Kennedy, 1998; Rojo, 2015; Subirats, 2022)。梳理西语语料库研究的主要成果,并比照同时期英语研究,不难发现,在前电子化时期,西语和英语语料库研究在起始时间上间隔不远,且在内容上较为接近,主要为含统计信息的词表、习语表和语法结构表。这一时期研究手动统计了自然文本中词语、短语或句法范畴的文本分布和出现频数,并以此为依据,确定了语言教学内容的优先级,这与语料库语言学基于真实语料而非单纯依赖内省、基于频数和分布提取语言特征的思路相吻合。

Keniston (1933) 的两份西语词频表是早期相关研究的代表性成果。两份词表均基于对真实语料的词汇统计,以词汇在文本中的分布广度为依据,共收常用词汇 1 322 个。第一份词表收录了在 80% 以上文本中出现的常用词汇;第二份词表收录了 30% 以上文本中出现的常用词汇。Keniston 创建词频表所用的语料文本主要为文学作品,也包含新闻、评论等文类。在 Keniston (1933) 之前, Jamieson (1924) 和 Cartwright (1925) 也基于词汇分布情况编制了西语教学大纲词表,但影响不及 Keniston (1933)。

率先将频率作为西语词表编制首要标准的学者是 Buchanan (1929)。受 Thorndike (1921) 英语词频表的启发, Buchanan (1929) 参考词语在文本中的绝对频数编写了西语词表。这一词表基于 120 万词文本,其中包含 7 种文体类型,75% 为文学作品,其余 25% 为新闻报刊、说明文和通俗读物。为减轻手工统计的工作量, Buchanan 的词表中没有统计功能词。Buchanan (1929) 词表共含 18 331 个词汇,并列出了词汇的频数和分布情况。此外, Keniston 开展了一系列采用计量分析的西语研究,包括短语频率表 (Keniston, 1929)、句法范畴频率分布手册 (Keniston, 1937a) 和卡斯蒂利亚语<sup>①</sup>语法统计分析 (Keniston, 1937b)。其中, Keniston (1937b) 聚焦 16 世纪处于变革中的西班牙语,是各语种中开展历时句法计量研究的早期典范。Keniston (1929、1937a) 为词频分析搭建的语料库也包含文学、报刊等多种文类,同时兼顾西班牙西语 (或半岛西语, *español peninsular*) 和美洲西语等不同地域变体,充分体现了早期研究者对文本取样代表性的关注。

针对特定地域变体西语词汇的计量研究也开始于同一时期。《卡斯蒂利亚语词汇研究》(*Investigaciones acerca de las palabras usadas en castellano*) (Céspedes, 1929, 转引自: Rodríguez Trujillo, 1980; López Morales, 2018) 基于巴拿马西语语料编制词表,共收 9 255 个词,全部按使用频数排序。Céspedes (1929) 所用语料来自各类书籍和报刊,共约 50 万词。与此前其他研究相比,这份词表并未提供词语的分布信息。

在语料采集上较具特色的两份重要词表是《西语词汇统计》(*Recuento de*

① 卡斯蒂利亚语 (英语称 Castilian, 西班牙语称 castellano) 是西班牙语的一种变体。Keniston (1937b) 中的卡斯蒂利亚语则是指自中世纪起在西班牙使用的西班牙语。

*Vocabulario Español*, Rodríguez Bou, 1952, 转引自: Ezquerro, 1974<sup>①</sup>) 和《普通词汇、常见词汇和基本词汇》(*Vocabulario usual, común y fundamental*, García Hoz, 1953) 两项研究。《西语词汇统计》中共列出 20 542 个词汇, 并按照不同的文类(广播节目、儿童口语和写作、成人口语和写作)给出词汇的分类频数以及总频数。Rodríguez Bou (1952) 为编写词表收集了 7 066 637 词的语言材料, 是同时代体量最大的西语语料库。其中文学作品的比例占 31%, 非文学文本如新闻报刊占比高达 41%, 余下 28% 的文本来自日常口语和联想词汇实验。

García Hoz (1953) 编写《普通词汇、常见词汇和基本词汇》所用语料库的库容约为 40 万词, 规模不及《西语词汇统计》。其取样方案及各类文本的占比与 Rodríguez Bou (1952) 语料库相似, 但 García Hoz (1953) 语料库中约 25% 的文本为农民或工人书写的家信, 共 620 封。García Hoz (1953) 认为, 下层民众的书信一定程度上能反映俚俗语言的使用情况, 因此专门收录了相应语料。以频数为依据, García Hoz (1953) 划分出约 13 000 个普通词汇、2 000 个常见词汇和近 200 个基础词汇。Rodríguez Bou (1952) 和 García Hoz (1953) 的语料采集方案反映了研究者控制文本多样性和代表性的意识, 与后来语料库语言学的取样原则不谋而合。

除书面语语料外, 成规模的西语口语语料收集工作几乎紧随英语研究开始。在英国学者夸克(Randolph Quirk)于 1959 年开展“英语用法调查”(Survey of English Usage)后不久, 墨西哥国立自治大学的布朗奇(Juan M. Lope Blanch)在 1964 年举办的美洲语言学和语言教学第二次研讨会(Segundo Simposio del Programa Interamericano de Lingüística y Enseñanza de Idiomas, PILEI)上提出了以访谈形式调查伊比利亚美洲<sup>②</sup>各大城市西语口语使用情况的构想, 也就是后来的“西班牙和伊比利亚美洲主要城市西班牙语雅语联合调查”(Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de España e Iberoamérica), 后更名为“布朗奇西班牙语雅语项目”(Proyecto de la norma culta hispánica Juan M. Lope Blanch)(Lope Blanch, 1969)。同英语用法

① Rodríguez Trujillo (1980)、Davies (2006)、López Morales (2018) 等文献也对 Rodríguez Bou (1952) 的研究作了介绍, Juilland & Chang-Rodríguez (1964) 和 Ueda (1987) 中亦列有此项研究, 足见其在早期西语词频表研制中的重要意义。

② 即使用西班牙语和葡萄牙语的美洲国家及地区。

调查类似,该语料库初期主要是以纸质方式出版(Esgueva & Cantarero, 1981; Pineda Pérez, 1983)。直到1998年,相关语料才以光盘形式发布(Samper Padilla et al., 1998)。该项目积累的数据为西语地域变体对比研究提供了重要的口语语料(Bernal Chávez & Hincapié Moreno, 2018)。

概言之,前电子化时期的西语语料库研究与英语同行的研究相比,起步并不算晚,且总体上早于其他语种的语料库研究。这一时期西语语料库研究呈不断扩展完善的趋势:词表统计依据由仅考虑文本分布向文本分布与词频结合转变,统计对象由词语向短语、句法范畴等成分扩展;同时,语料规模不断扩大,丰富程度不断增加:语料的地域分布由西班牙扩展至美洲,语体由笔语、准口语扩展至口语。同时,与英语相关研究类似,这一时期的西语语料库研究也都有比较明确的语言教学导向。

## 2.2 电子化时期

在早期的电子化语料库中,影响最大的是“布朗语料库”(the Brown Corpus),该语料库于1962年启动,并于1964年建成。然而,第一个电子化西语通用语料库项目《西班牙语频率词典》(*Frequency Dictionary of Spanish Words*)于1956年启动,于1963年建成(Juilland & Chang-Rodríguez, 1964),其诞生要先于布朗语料库。此后,西语语料库得到长足发展。以下我们将综述以英语和西语撰写的西语语料库建设与研究成果。

### 2.2.1 西语语料库建设成果

《西班牙语频率词典》项目开始于1956年,所用语料库规模为50万词,主要包含1920至1940年内出版的五类文本:戏剧、小说、散文、学术文献和报刊。其中,文学作品占总体量的80%,非文学文本占20%。该语料库仅收录半岛西语文本,未收美洲西语文本(Juilland & Chang-Rodríguez, 1964)。

此后,西班牙中世纪研究会(Hispanic Seminar of Medieval Studies)为编写《古西班牙语词典》(*Dictionary of Old Spanish Language*),于20世纪70年代将中世纪西班牙语的文本电子化。这些文本后来也成为西班牙皇家语言学院“历时西班牙语语料库”(Corpus diacrónico del español, CORDE)的一部分(Rojo, 2016),属于较早的西语电子文本库。

20世纪80年代初期,哥德堡大学也开发了两个西语语料库:一个是

“ONE71 语料库”，收集了 1951 至 1971 年出版的西语小说，规模约 100 万词（Mighetto, 1985）。另一个是“PE77 语料库”，收集了 3 000 余篇发表在西班牙 *El País* 报和 *Triunfo* 杂志上的文章，规模约 200 万词（Mighetto & Rosengren 1982、1983、1985）。1998 年，米格托（David Mighetto）基于上述两个语料库创建了西语在线语料库平台“西语在线”（Spanish Online）。

20 世纪 90 年代，取样较广泛的“鲁汶西语语料库”（corpus de Lovaina）建成。该语料库由（荷兰语）鲁汶大学的德考克（Josse De Kock）牵头建设，先后收集了 1922 至 1988 年间发表或出版的半岛西语和美洲西语文本共约 400 万词，于 1990 至 1992 年发布（De Kock 1990a、1990b、1991、1992；De Kock et al. 1991；De Kock et al. 1992）。

随着信息技术的进步和电子语料的富集，西语语料库在规模和类型上都有大幅提升。其中，通用大型语料库以西班牙皇家语言学院主持建设的“当代西班牙语参照语料库”（Corpus de referencia del español actual, CREA）、“西班牙语历时语料库”（Corpus diacrónico del español, CORDE）和“21 世纪西班牙语语料库”（Corpus del español del siglo XXI, CORPES）为代表。当代西班牙语参照语料库 CREA 包含来自所有西语国家或地区的口语和笔语材料，时间跨度在 1975 至 2004 年间。该库 2008 年建成时，库容超过 1.6 亿词。西班牙语历时语料库 CORDE 收录所有 1974 年前出版的西语文本，包括诗歌、韵文等多种文体，库容约 2.5 亿词。21 世纪西班牙语语料库 CORPES 由西班牙皇家语言学院创建和维护，语料持续更新。2021 年 6 月发布的 0.94 版包含来自 32.7 万个文本和录音的语言材料共 3.5 亿词，较上一个版本新增 1 800 万词。其语料的时间跨度为 2001 至 2021 年。在文体分布上，CORPES 虚构类文本（小说、影视剧本、故事等）超过 9 500 万词，非虚构类文本和媒体报道约 2.5 亿词；在地域分布上，半岛西语和美洲西语文本的体量分别占全库的 30% 和 70%（参见西班牙皇家语言学院网站）。这三个语料库建设框架合理，各类语法标注较为详尽准确，尤其是 CORPES 语料库依然在按照同样的取样框持续更新，能够与时俱进地反映目前西语使用的现状。

除西班牙皇家语言学院语料库系列以外，由 Subirats 和 Ortega（2012）建设的“当代西班牙语语料库”（Corpus del Español Actual, CEA）含有来自欧洲议会会议实录平行语料库（European Parliament Proceedings Parallel Corpus 1996-



2011)、维基百科和联合国多语平行文本中的西语文本,共约5.4亿词。由戴维斯(Mark Davies)主持开发的“西班牙语语料库”(El corpus del español)最初于2001年建成,包含13至19世纪的口、笔语文本约100万词(Davies, 2002)。此后,更多来自网络的语言材料被添加到该库之中,目前规模已超75亿词。戴维斯近年还在El corpus del español上线了73亿词次的西语新闻库、20亿词次的网络西语库等子库。此外,Sketch Engine平台中的esTenTen家族语料库包含的“西班牙语网络语料库2018”(esTenTen18)的库容已达到169亿词次(Kilgariff & Renau, 2013)。由互联网语料构成的数据库通常使用自动爬取技术,语料一般不作甄别,因此亦有学者认为将它们称为“档案库”更为合适(Rojo, 2016)。例如,esTenTen18语料库中混杂有本族语者原创西语、翻译西语和非本族语者西语,这在很大程度上会影响研究结论的可信度和可推广性。

除通用西语语料库外,西语的地域变体语料库、学习者语料库,以及学术西语库等专门语料库也日渐丰富。

地域变体语料库中,“智利西班牙语动态语料库”(Corpus Dinámico del Castellano de Chile, Codicach)由智利西语构成,库容近9亿词,其中绝大部分内容为笔语,也含有议会发言实录。文本均经过清洁、词形还原、词性标注和句法标注,可通过CQP网络界面在线检索(Sadowsky, 2006)。“瓦伦西亚西班牙语口语库”(Valencia Español Coloquial, Val.Es.Co.)最初包含341小时瓦伦西亚(西班牙)市区及周边的西语口语录音,及其中46段对话的转写,超12万词,可通过在线平台检索(Pons Bordería & Ruiz Gurillo, 2005)。<sup>①</sup>

学习者语料库方面,2021年10月更新的“西班牙语学习者语料库”(Corpus de aprendices de español, CAES)版本包含约76万词,囊括了《欧洲共同语言参考框架》(Common European Framework of Reference for Languages)中所有水平的学习者产出,学习者母语分别为阿拉伯语、中文、法语、英语、葡萄牙语和俄语。专门针对中国西语学习者的语料库有北京外国语大学西班牙语葡萄牙语学院何晓静和刘元祺主持开发的“中国西班牙语学习者语料库”(何晓静,刘元祺,2018)。该库已收集测试和非测试两类文本,其中测试类文本包括

① 从2.1版本起,Val.Es.Co.语料库的语料来源不再限于瓦伦西亚市及周边地区。目前最新版本的Val.Es.Co. 3.0含文字转写约25万词(Pons Bordería, 2021)。

全国西班牙语专业八级水平测试 2013—2015 年三年应届本科生的作文语料, 共计文本 3 842 篇, 约 80 万词; 非测试语料以命题作文的方式在高校西班牙语专业的学生中采集, 历时 4 年, 共计文本 957 篇, 约 28 万词 (何晓静, 刘元祺, 2018)。

学术文本方面, 由庞蒂菲西亚大学瓦尔帕莱索卡托利卡分校 (Pontificia Universidad Católica de Valparaíso) 主持建设的 “PUCV-2010 学术西语语料库” 包括来自物理、化学、生物、历史、文学和语言学等六个学科领域的学术文本 (Parodi, 2007、2010)。此外, 北京外国语大学刘元祺还建成 “西语国家新闻语料库” 和 “西语国家获奖作家小说语料库”, 二者都可通过北外语料库语言学多语种语料库平台在线检索。

西语文本的多语平行语料库包括 “联合国平行语料库” (Corpus paralelo de las Naciones Unidas) (Ziemski et al., 2016) 和 “西语 - 纳瓦特尔语平行语料库” (Corpus paralelo español-náhuatl Axolotl) (Gutierrez-Vasques et al., 2016)。纳瓦特尔语是墨西哥中部到哥斯达黎加西北部地区的原住民语言, 今天仍在墨西哥部分地区使用。这一平行语料库可在线检索使用。

可以看出, 在计算机技术不断发展, 高度普及的今天, 西语语料库建设非常活跃, 成果丰硕。既有由皇家语言学院等权威语言学术机构主持建设的大型通用语料库, 也有研究者个人根据不同研究目的自行建设的中、小型专门语料库; 既有纵向考察西语发展的历时库, 也有横向考察西语地域差异的方言库, 还有从社会语言学角度切入并引入说话人社会文化特征的口语库。丰富的语料库资源为当前西语语料库研究奠定了坚实的基础。

### 2.2.2 电子化时期西语语料库研究

西语语料库的建设虽不比英语逊色太多, 但以西语为对象的语料库研究在学界的受关注程度却难与英语比肩。语料库研究近年来可以算是国内外语言学研究的一个热点, 但通过 Web of Science 数据库查询 2019 年至本文成稿时出版的 SSCI 期刊, 标题含有 “Spanish” / “español” 和 “corpus” 的语言学研究仅有 42 篇。同一时期, 在 SSCI 期刊上发表的标题含有 “corpus” 的论文共有 3 102 篇。西语语料库研究约为总数的 1%。

从 SSCI 上发表的论文来看, 近 30 年来较受关注的西语语料库研究大致集中在两方面: (1) 将英语学界新提出或已提出的理论和方法迁移至西语文本分析或特定语言特点分析, 以补充、完善相关理论的西语适用性, 例如, 话语



分析 (Romano, 2011; Sola Morales, 2012) 和语域或语体分析 (Parodi, 2009; Errázuriz Cruz, 2012) 等; (2) 将语料库技术及方法应用于更具体的领域, 如词汇、句法研究 (Coll-Florit, 2011) 和二语习得中的错误分析 (Campillos Llanos, 2014) 等方面。

为呈现西语语料库研究的具体样貌, 我们分别检索了 Web of Science 数据库中含 “corpus” 和 “Spanish” 关键词的英语文献和含 “corpus” 关键词的西语文献, 并将学科限定在语言学、文学、信息科学三个学科。我们选取了 1990 年至今每一个十年中引用数量最高的 20 篇西语语料库研究论文 (英语、西语各 10 篇), 一一列出并比对了它们的研究类型、关注的语言现象、研究方法、语料库信息等。其中, 1990 至 2000 年间, Web of Science 数据库只收录了两篇相关英语论文, 没有西语论文, 因此最终列出论文共 42 篇。

在这 42 篇论文中, 学术话语研究是一个较为突出的研究热点。22 篇英语论文中, 4 篇涉及学术西语; 20 篇西语论文中, 亦有 11 篇文章都以学术西语为研究对象。这一方面是西语学者对海兰 (Ken Hyland) 一系列成果 (如 Hyland, 2000) 的敏锐反应, 另一方面也是因为学术文本的收集相对便捷。除此以外, 语篇层面的研究多考察意义或形义关系问题, 其相关理论和研究方法容易在不同语言间迁移。采用语篇层面的理论框架, 如语步 (Venegas et al., 2016)、元话语标记 (Dafouz-Milne, 2008; Mur-Dueñas, 2011)、概念隐喻等 (Rojo & Orts, 2010), 在分析西语的语言材料时并不需要太多调整, 这也有助于相关西语研究的开展。

此外, 两个语种的论文都比较关注西语口语的研究, 包括双语者语码转换现象 (Herring et al., 2010; Fricke et al., 2016; Guzzardo Tamargo et al., 2016)、对特定区域口语的社会语言学调查 (Otheguy et al., 2007; San Martín & Guerrero, 2013)、口语委婉语 (Briz & Albelda, 2013) 等等。我们还注意到研究学习者偏误问题的论文 (Ferreira et al., 2014)。

高被引西语语料库研究的另一个特点是常使用先进的量化研究方法, 例如变项规则分析 (Otheguy et al., 2007)、逻辑斯蒂回归分析 (Erker & Guy, 2012)、混合效应线性回归 (Fricke et al., 2016; Guzzardo Tamargo et al., 2016)、多维度分析 (Parodi, 2004)、潜在语义分析 (Venegas, 2006) 和支持向量机 (Venegas, 2007) 等。

### 3 结语

本文梳理了以西语为研究对象的语料库研究，重点回溯了电子化之前和电子化早期被学界忽视的相关西语语料库研究。西语使用涉及欧洲、美洲的众多国家和地区，作为母语、继承语和第二语言的使用都十分广泛，强烈的研究需求推动了西语语料库的建设，也使西语在地域等层面的变异自然而然地成为研究者关注的焦点之一。也正因如此，西语语料库语言学研究的学者和机构并不局限于欧洲大陆，美洲地区的学者如智利的 Giovanni Parodi、美国的 Hayward Keniston 等也对学科发展做出了重大贡献。

诚然，语料库语言学在英语研究方面成果最为显著。然而，近百年的西班牙语语料库发展史也体现出其突出特点。首先，从学术源头来看，西语电子化平衡语料库的开发早于英语。其次，在研究的思路和方法上，西语语料库研究始终关注文本分布和地域分布，常用先进的量化统计分析方法。最后，在机制建设上，当前的国家级西语语料库研制由官方语言学机构主导，可以确保语言材料规范可信，平台系统维护良好，学术使用广泛普及。这些均可为我国语料库建设与研究提供有益参考。

---

### 参考文献

- Abad Nebot, F. 2017. *Historia general de la lengua española*[M]. Valencia: Tirant Humanidades.
- Bernal Chávez, J. A., & Hincapié Moreno, D. A. 2018. *Lingüística de corpus*[M]. Bogotá: Instituto Caro Y Cuervo.
- Briz, A., & Albelda, M. 2013. Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués. La base de un proyecto en común (ES. POR.ATENUACIÓN)[J]. *Onomazein*, 28: 288-319.
- Buchanan, M. A. 1929. *A graded Spanish word book (Rev. ed.)*[M]. Toronto: The University

of Toronto Press.

- Campillos Llanos, L. 2014. Análisis de errores pragmático-discursivos en un corpus oral de español como lengua extranjera[J]. *Círculo de lingüística aplicada a la comunicación*, 58: 23-59.
- Cartwright, C. W. 1925. A study of the vocabularies of eleven Spanish grammars and fifteen Spanish reading texts[J]. *The Modern Language Journal*, 10(1): 1-14.
- Céspedes, A. T. R. 1929. *Investigación acerca las palabras usadas en castellano*[M]. Panamá: Panamá Star and Herald.
- Coll-Florit, M. 2011. Aproximación empírica a los modos de acción del verbo: Un estudio basado en corpus[J]. *Revista signos*, 44(77): 233-250.
- Dafouz-Milne, E. 2008. The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse[J]. *Journal of Pragmatics*, 40(1): 95-113.
- Davies, M. 2002. Un corpus anotado de 100.000. 000 palabras del español histórico y moderno[J]. *Procesamiento del Lenguaje Natural*, 29: 21-27.
- Davies, M. 2006. *A frequency dictionary of spanish: Core vocabulary for learners*[M]. New York: Routledge.
- De Kock, J. 1990a. *Gramática española: enseñanza e investigación V: 1. Concordancia alfabética de 19 textos (solo consultable en forma de listado)*[M]. Salamanca: Universidad de Salamanca.
- De Kock, J. 1990b. *Gramática española: Enseñanza e investigación V: 2. Concordancia alfabética de 20 textos (solo consultable en forma de listado)* [M]. Salamanca: Universidad de Salamanca.
- De Kock, J. 1991. *Gramática española: enseñanza e investigación IV Índices: 1. Índices de 19 textos*[M]. Salamanca: Universidad de Salamanca.
- De Kock, J. 1992. *Gramática española: enseñanza e investigación IV Índices: 2. Índices de 20 textos*[M]. Salamanca: Universidad de Salamanca.
- De Kock, J., Gómez Molina, C., García Mouton, P., & Delbecque, N. 1992. *Gramática española, enseñanza e investigación III Textos: 2. 20 texto*[M]. Salamanca: Universidad de Salamanca.

- De Kock, J., Verdonk, R., & Gómez Molina, C. 1991. *Gramática española, enseñanza e investigación III Textos: I. 19 textos*[M]. Salamanca: Universidad de Salamanca.
- Erker, D., & Guy, G. R. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish[J]. *Language*, 88(3): 526–557.
- Errázuriz Cruz, M. C. 2012. Análisis del uso de los marcadores discursivos en argumentaciones escritas por estudiantes universitarios[J]. *Perfiles Educativos*, 34(136): 98-117.
- Esgueva, M., & Cantarero, M. (Eds.) 1981. *El habla de la ciudad de Madrid. Materiales para su estudio*[M]. Madrid: Consejo Superior de Investigaciones Científicas.
- Ezquerro, R. 1974. Los diccionarios de frecuencia en español[J]. *Boletín de la Asociación de Profesores de Español*, 10: 43-54.
- Ferreira, A., Elejalde, J., & Vine, A. 2014. Análisis de errores asistido por computador basado en un corpus de aprendientes de español como lengua extranjera[J]. *Revista Signos*, 47(86): 385-411.
- Francis, N. 1992. Language corpora B. C.[M]. In J. Svartvik (Ed.), *Directions in corpus linguistics*. Berlin: Mouton de Gruyter: 17-32.
- Fricke, M., Kroll, J. F., & Dussias, P. E. 2016. Phonetic variation in bilingual speech: A lens for studying the production–comprehension link[J]. *Journal of Memory and Language*, 89: 110-137.
- García Hoz, V. 1953. *Vocabulario usual, común y fundamental: Determinación y análisis de sus factores*[M]. Madrid: Consejo Superior de Investigaciones Científicas, Instituto “San José de Calasanz” .
- Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. 2016. Axolotl: A web accessible parallel corpus for Spanish-Nahuatl[M]. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association (ELRA): 4210-4214.
- Guzzardo Tamargo, R. E., Valdés Kroff, J. R., & Dussias, P. E. 2016. Examining the relationship between comprehension and production processes in code-switched language[J]. *Journal of Memory and Language*, 89: 138-161.
- Herring, J. R., Deuchar, M., Parafita Couto, M. C., & Moro Quintanilla, M. 2010. “I saw

- the *madre*”: evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data[J]. *International Journal of Bilingual Education and Bilingualism*, 13(5): 553–573.
- Hyland, K. 2000. *Disciplinary discourses: Social interactions in academic writing*[M]. London: Longman.
- Jamieson, E. I. 1924. A standardized vocabulary for elementary Spanish[J]. *Modern Language Journal*, 8(6): 325–333.
- Juilland, A., & Chang-Rodríguez, E. 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. 1933. *A basic list of Spanish words and idioms*[M]. Chicago: University of Chicago Press.
- 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: Henry Holt and Company.
- Keniston, H. 1937b. *The syntax of Castilian prose: The sixteenth century*[M]. Chicago: The University of Chicago Press.
- Kennedy, G. 1998. *An introduction to corpus linguistics*[M]. London: Addison Wesley Longman.
- Kilgariff, A., & Renau, I. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish[J]. *Procedia-Social and Behavioral Sciences*, 95: 12–19.
- Lope Blanch, J. M. 1969. Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica[M]. In Universidad Nacional Autónoma de México. *El simposio de México, enero de 1968: Actas, informes y comunicaciones*. México D.F.: PILEI: 222–233.
- López Morales, H. 2018. *Estudios sobre el español de Cuba*[M]. Madrid: Editorial Verbum.
- Mighetto, D. 1985. *ONE71. Banco de datos de once novelas españolas 1951–1971*[M]. Gotemburgo: Göteborgs Universitet.
- Mighetto, D., & Rosengren, P. 1982. *Banco de datos de Prensa española 1977:*



- Concordancia lingüística y texto fuente*[M]. Gotemburgo: Göteborgs Universitet.
- Mighetto, D., & Rosengren, P. 1983: *PE77. Palabras gráficas españolas: Lista y frecuencias en Prensa Española 1977*[M]. Gotemburgo: Göteborgs Universitet.
- Mighetto, D., & Rosengren, P. 1985: *Diccionario reverso*[M]. Gotemburgo: Göteborgs Universitet.
- Mur-Dueñas, P. 2011. An intercultural analysis of metadiscourse features in research articles written in English and in Spanish[J]. *Journal of Pragmatics*, 43(12): 3068-3079.
- Nebrija, E. A. 1997. Prólogo a la gramática de la lengua castellana[M]. In I. Guzmán Betancourt & E. Nansen Díaz (Eds.), *Memoria del Coloquio La Obra de Antonio de Nebrija y su recepción en la Nueva España, quince estudios nebrisenses (1492-1992)*. México D.F.: Instituto Nacional de Antropología e Historia: 199-202.
- Otheguy, R., Zentella, A. C., & Livert, D. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community[J]. *Language*, 83(4): 770-802.
- Parodi, G. 2004. Specialized texts and discourse technical-professional communities: A computerized corpus-based approach[J]. *Estudios Filológicos*, 39: 7-36.
- Parodi, G. 2007. Specialized written discourse at university and professional domains: Composition of a corpus[J]. *Revista Signos*, 40(63): 147-178.
- Parodi, G. 2009. Corpus, discurso y géneros: Español en contextos académicos y profesionales[M]. In A. Vera Luján & I. Martínez Martínez (Eds.), *El español en contextos específicos: enseñanza e investigación*. España: Fundación Comilla y ASELE: 65-88.
- Parodi, G. 2010. Multisemiosis y lingüística de corpus: Artefactos (multi) semióticos en los textos de seis disciplinas en el corpus PUCV-2010[J]. *RLA. Revista de Lingüística Teórica y Aplicada*, 48(2): 33-70.
- Pineda Pérez, M. Á. (ed.) 1983. *Material de encuestas para el estudio del habla urbana culta de Sevilla*[M]. Sevilla: Universidad de Sevilla.
- Pons Bordería, S. (dir.) 2021. *Corpus Val.Es.Co 3.0*[OL]. [2023-02-24]. <http://www.valesco.es/>.
- Pons Bordería, S., & Ruiz Gurillo, L. 2005. Corpus para el estudio de la conversación coloquial. El corpus Val.Es.Co. (Valencia, Español Coloquial)[J].

- Oralia*, 8: 243-263.
- Rodríguez Bou, L. 1952. *Recuento de vocabulario Español*[M]. Río Piedras: Universidad de Puerto Rico.
- Rodríguez Trujillo, N. 1980. Listas de frecuencias de palabras: Una revisión de la literatura en español y de sus posibles usos en investigación[J]. *Lectura y Vida*, 1(4): 1-6.
- Rojo, A. M., & Orts, M. Á. 2010. Metaphorical pattern analysis in financial texts: Framing the crisis in positive or negative metaphorical terms[J]. *Journal of Pragmatics*, 42(12): 3300-3313.
- Rojo, G. 2015. Sobre los antecedentes de la lingüística de corpus[M]. In A. Álvarez Menéndez et al. (Eds.), *Studium grammaticae. Homenaje al Profesor José Antonio Martínez*. Oviedo: Universidad de Oviedo: 675-689.
- Rojo, G. 2016. Los corpus textuales del español[M]. In J. Gutiérrez-Rexach (Ed.), *Enciclopedia lingüística hispánica*. Abingdon and New York: Routledge: 285-296.
- Romano, M. B. 2011. Análisis gramatical, discursivo y crítico de noticias acerca del escándalo político en prensa escrita argentina[J]. *Revista electrónica de estudios filológicos*, 21: 1-29.
- Sadowsky, S. 2006. *Corpus Dinámico del Castellano de Chile (Codicach)*[OL]. [2023-02-24]. <http://sadowsky.cl/codicach.html>.
- Samper Padilla, J. A., Hernández Cabrera, C. E., & Troya Déniz, M. 1998. *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (MC-NLCH)*[CD]. Las Palmas de Gran Canaria: Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria-ALFAL.
- San Martín, A., & Guerrero, S. 2013. Una aproximación sociolingüística al empleo del discurso referido en el corpus PRESEEA de Santiago de Chile[J]. *Revista Signos*, 46(82): 258-282.
- Sola Morales, S. 2012. ¿Víctima o heroína? Un análisis crítico de los blogs “Mujeres” y “En Femenino” [J]. *Discurso & Sociedad*, 6(4): 815-849.
- Subirats, C. 2022. Lingüística de corpus y semántica cognitiva en español[M]. In G. Parodi et al. (Eds.), *Lingüística de corpus en español*. Abingdon and New York: Routledge: 205-222.
- Subirats, C., & Ortega, M. 2012. *Corpus del Español Actual*[OL]. [2023-02-24]. <http://>

spanishfn.org/tools/cea/spanish.

Thorndike, E. 1921. *The teacher's word book*[M]. New York City: Teacher College, Columbia University.

Ueda, H. 1987. *Frecuencia y dispersión del vocabulario español*[M]. Tokyo: Universidad de Estudios Extranjeros de Tokio.

Venegas, R. 2006. La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente[J]. *Revista Signos*, 39(60): 75-106.

Venegas, R. 2007. Clasificación de textos académicos en función de su contenido léxico-semántico[J]. *Revista Signos*, 40(63): 239-271.

Venegas, R., Zamora, S., & Galdames, A. 2016. Hacia un modelo retórico-discursivo del macrogénero Trabajo Final de Grado en Licenciatura[J]. *Revista Signos*, 49(supl.1): 247-279.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. 2016. The United Nations parallel corpus v1.0[M]. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA): 3530-3534.

何晓静, 刘元祺, 2018. 中国西班牙语学习者语料库 (CACE): 规划与展望 [J]. 语料库语言学 5 (2): 98-108.

何中清, 彭宣维, 2011. 英语语料库研究综述: 回顾、现状与展望 [J]. 外语教学 32 (1): 6-15.

许家金, 2017. 语料库研究学术源流考 [J]. 外语教学与研究 49 (1): 51-63, 159.

许家金, 2019. 美国语料库语言学百年 [J]. 外语研究 (4): 1-6.

---

## 作者简介:

赵冲, 北京外国语大学中国外语与教育研究中心在读博士生。研究方向: 西班牙语语言学, 语料库语言学。电子邮箱: zhao\_chong@bfsu.edu.cn

许家金，教授，博士生导师，北京外国语大学中国外语与教育研究中心 / 人工智能与人类语言重点实验室。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。电子邮箱：xujiajin@bfsu.edu.cn

*Red Chamber*, discuss the role of culture in translation practice, and highlight the challenges for translators in the new era.

**Keywords:** *Dream of the Red Chamber*; “Won-Done Song”; cultural translation theory; Italian

## A Centenary Review on the Spanish Corpus Construction and Research

Zhao Chong, Xu Jiajin

**Abstract:** In the history of corpus linguistics, English studies have been taking the leading position. A historical overview shows that Spanish corpus studies also figure prominently in the following respects. Firstly, regarding scholarly provenance, the compilation of the first electronic balanced Spanish corpus predates its English counterpart. In general, Spanish corpus studies are mostly paralleled to the English corpus studies in each period of time. The second point concerns the methodology. Spanish corpus research has always prioritized textual and regional distribution and variation. Lastly, some of the biggest corpus projects of Spanish are coordinated by authoritative linguistic institutions, which ensures a more reliable text source and a well-maintained public access to the corpora, facilitating a wide application of the corpora in research. The review of Spanish corpus construction and research will offer insights into China’s corpus linguistics.

**Keywords:** corpus linguistics; Spanish; historiography