第三章 语料库方法论

在第二章中我们已经对话语的定义、话语研究的批判性与非批判性方法、话语研究的现状以及低碳经济等研究进行了综述；本章将概述语料库的发展历史，语料库在现代语言学中的应用，语料库方法在话语分析中的应用等，同时，我们将介绍用于本研究中所建的专题语料库的采样、语料处理以及用于语料库分析的软件 WordSmith Tools 6。

第一节 语料库概述¹

语料库在现代语言学中指“书面语和口语的采样文本集合，以机器可读形式，并且可以以各种语言信息形式赋码 (McEnery *et al.*, 2006: 4)。自上世纪 90 年代以来，基于语料库的语言研究方法不断创新，研究范式不断丰富，成果层出不穷，如：McEnery *et al.* (2006) 突破以往单以英语数据为对象的研究，深入探索了多语种语料库语言学研究，将以往语料库数据分析的多种方法有机结合，在分析 Sinclair (1996) 和 Leech (1992) 等学者对话料库的定义后，提出了语料库定义的四个要素，即：1) 机器可读性；2) 真实文本（包括口语记录的文字）；3) 采样；4) 某种语言和语言变体的代表性。不同的学者对此有不同的定义，各种定义的共同之处在于“机器可读的真实语言的收集”，然而，学界对于语料库的代表性却有争议。没有一个语料库能穷尽所有语言的收集来代表一种语言，这些收集的语言数据只能代表“语言变体的主要维度，比如：口语、书面语，正式、非正式，小说、非小说，英国英语、美国英语，或者代表不同年龄层

¹ 本节第二部分内容已发表于《外语教学与研究》2007 年第 3 期。

次、专家还是普通人所使用的语言 (Stubbs, 2001b: 305)。毋庸置疑, 过去的几十年里, 语料库语言学作为一种新的语言学研究方法已经渗透到语言研究的各个分支, 比如: 句法学、语义学、语用学等。

一、现代语言学中的语料库

自 20 世纪 60 年代计算机技术开始迅速发展以来, 计算机在对语言的理解和教学中发挥了越来越重要的作用。例如, 计算机出现之前, 是凭编者积累卡片来编词典和语法, 效率较低, 也几乎没有办法检查这些语言规则在现实世界中的真实使用情况。语料库的出现改变了人们思考语言的方式, 就像望远镜改变了天文学, X-光机器刺激了医学的进步, 录音机推动了社会语言学 and 口头话语研究的发展, 录像机促进了小组互动的研究, 摄谱仪 (和类似的设备) 深化了仪器语音学的发展 (Simpson & Swales, 2001)。

随着语料库的不断扩容, 用于处理语料库的软件也在不断更新。上世纪 90 年代出现了一场“语料库革命” (Chapelle, 2001)。语料库的使用越来越受到几乎所有和语言相关领域的关注, 经过几十年的发展, 已渗透到语言研究的各个分支, 如句法学、语义学、语用学、话语研究等; 实现了 McEnery & Wilson (1996, 2001) 的预言, 在语料库形态、语料库规模、语料选择、语料处理、专题语料库、语言理论、翻译研究、文体学、语法和词典开发、“数据驱动”学习和话语研究等方面取得卓著成就。如今语言研究中, 语料库相关的学问, 即便不是显学, 也称得上热门无疑。对此, 学界有激赏, 也有质疑 (许家金, 2017)。

尽管语料库语言学的概念在 20 世纪 80 年代初才出现, 但基于语料库的研究已有相当长的一段历史了。语料库的发展史可以追溯到 1898 年 Kaeding 编写的德语频率词典, 1921 年美国 Thorndike 编写的教师词汇手册, 1953 年 Micheal West 做的常用词义项频率统计, 以及 1959 年 Quirk 的英语用法调查等 (参见梁茂成、李文中、许家金, 2010)。在我国, 从 20 世纪 20 年代开始, 就有学者建立文本的语料库, 采用统计的方法来研究汉字的频率, 其目的在于制定基础汉字的字表, 如:《语体文应用字汇》 (陈鹤琴, 1925)。那时的语言学家也许用一个鞋盒子放入一些收集的书面语或者转写的文字纸片, 并无代表性, 只是些凭经验收集的小型观测数据 (McEnery *et al.*, 2006)。这些语料库并非计算机可读, 但它们是现代语

料库的雏形，在语料库的发展史上功不可没（冯志伟，2002）。1961年美国布朗大学开始100万词规模机器可读语料库的建设，上世纪70年代英国开始LOB语料库的建设，中国也开始进行机器可读语料库的建设，如：汉语现代文学作品语料库（1979年）；现代汉语语料库（1983年），2000万字；中学语文教材语料库（1983年），106万8千字；现代汉语词频统计语料库（1983年），182万字。发展至今，语料库无论是从规模到类型都发生了深刻的变化。20世纪60年代布朗家族语料库的诞生为语料库的发展做出了巨大的贡献，为后来的系列更大型的国家语料库以及更大型的平衡语料库建设提供了重要的参考。近年来，许多超大型的档案资料式语料库以及以网络为数据的语料库陆续建成，如：Google Books、LexisNexis、Hansard、NOW等，为多学科的研究开辟了新的路径和方法。

（一）布朗家族语料库

1962年在美国布朗大学 Nelson Francis 和 Henry Kučera 的带领下开始创建世界上第一个计算机可读的英语通用语料库 Brown Corpus，1964年完工（Francis & Kučera，1964）。该语料库的建成意义重大，不仅因为它是历史上第一个机器可读通用语料库，而且是在当时美国乔姆斯基的转换生成语法理论盛行的背景下创建。乔姆斯基的理论引领了当时语言研究的方向，他明确反对使用语料库，认为语言研究应该围绕着语言能力而非语言表现。Brown Corpus 收集了1961年出版的美国英语的读物，共有1014312字，由500篇2000字的文本构成，包含15种文类，即：新闻报道、新闻社论、新闻评论、宗教、技能、行业和爱好、流行的传说、传记和散文、其他（报告、官方文件）、科学（学术散文）、普通小说、神秘与侦探小说、科幻小说、西方与冒险小说、浪漫小说、幽默。文本的选择标准注重代表性，其采样标准被许多语料库所采用，并先后建成了系列布朗家族语料库，如：Lancaster-Oslo/Bergen Corpus（Johansson *et al.*，1978）、Kolhapur Corpus of Indian English（Shastri *et al.*，1986）、Wellington Corpus of Written New Zealand English（Bauer，1993）、Macquarie Corpus of Written Australian English（Collins & Peters，1988）、Frown Corpus（Hundt *et al.*，1999）、Flob Corpus（Mair *et al.*，1999）、British English 2006 Corpus（Baker，2009）、Lancaster 1901（Leech & Smith，2005）、Lancaster 1931（Leech

& Smith, 2005)、Crown Corpus & CLOB Corpus (许家金、梁茂成, 2009), 以及汉语的 Lancaster Corpus of Mandarin Chinese (McEnergy & Xiao, 2004) 和 ToRCH 语料库 (许家金, 2009、2014)。

布朗家族语料库成为语言研究极好的资源, 成果层出不穷, 有历时比较研究、共时比较研究, 也有语法测量研究 (Mair *et al.*, 2002; Hudson, 1994; Rayson *et al.*, 1997; Granger & Rayson, 1998; Biber *et al.*, 1999; Rayson *et al.*, 2002), 如 Mair *et al.* (2002) 发现英语中名词的使用在 1961—1991 年间有显著的增加, 这和越来越多的名词化风格相关, 而动词却处于稳定状态。这些研究中还有对单个语言细节和语法特征变迁的观察, 如: McEnergy & Xiao (2005) 对英国英语和美国英语中 help 的用法做了历时和共时的比较研究, 发现历经 30 年语言的发展, help 后面的不定式 to 的脱落现象在英国英语中 (相比美国英语) 更明显。Leech (2004) 和 Leech & Smith (2006) 对英英和美英的语言发展趋势进行了研究, 发现了英国英语“口语化”和“美语化”的趋势。Rayson (2007)、Hundt (2004) 对进行时的被动语态进行了观察, Leech (2004) 和 Leech & Smith (2006) 通过对核心情态动词与半情态动词的历时变迁以及英音和美音的差异研究, 发现 30 年间情态动词使用的下降趋势和社会“民主化”以及“避免不平等”或“面子威胁”的趋势相关。还有其他的一些成果, 不胜枚举。如今, 这些语料库依然是研究语言变体和语言变迁的重要依据。

布朗家族语料库作为语料库的先行者为拓开语言研究的新思维功不可没, 但它存在许多缺陷, 比如, 语料库中不同单词形式的频率通常呈离散幂律概率分布 (Zipf, 1935), 其中第二个高频词出现的频率约为最高词频单词的一半, 第三个高频词出现的频率约为三分之一, 依此类推。如: Brown 语料库中出现次数最多的词“the”占语料库中所有词出现次数的 7%, 排名第二的是“of”占 3.5%。因此, 如果语料库要反映尽可能多的用法以及语言特征, 那就需要大型的、多种类语料 (Aston & Burnard, 1998)。针对 Brown 和 LOB 语料库, Sinclair (1991: 24) 也指出它们能提供的只是大量文本中出现的相对高频单词的可靠信息, 对于只出现在某些文本类型中的单词, 它们的可靠性要低得多。所以, 扩大各文类的量能在很大程度上解决这一问题。随着计算机技术的迅猛发展, 大型语料库建设成为可能, 国家语料库是其中重要的代表。