语言技术

近30年来中文语言知识资源发展及应用*

詹卫东

（北京大学　中国语言文学系 / 中国语言学研究中心 / 计算语言学教育部重点实验室　北京　100871）

提　要　本文利用互联网搜索引擎，调研了中国大陆和港台地区，以及北美、欧洲等多地的中文语言知识资源，包括语料库、知识库及相应的检索系统的现状。得益于经验主义研究范式在自然语言信息处理以及其他语言应用研究领域近 30 年来的快速发展，中文世界的可用语言知识资源已经积累到了相当可观的规模。本文从 4 个方面讨论了中文语言知识资源在汉语研究及教学中的应用价值，并简要分析了资源建设面临的挑战及对汉语语言学未来发展可能造成的影响，指出汉语语言学研究的理想进路应是将基于理性内省的语言学研究范式与基于真实海量语言数据的实证分析相结合，而不是将二者对立起来。

关键词　语言知识资源；语料库；知识库；检索系统

中图分类号　H002　**文献标识码**　A　**文章编号**　2096-1014（2018）04-0058-12

DOI　10.19689/j.cnki.cn10-1361/h.20180405

An Overview of the Advances and Applications of Online Chinese Language Resources over Three Decades

Zhan Weidong

Abstract　In the past three decades, empiricism paradigm in research prevails in natural language processing and other language application fields, which leads to the boom of online language data resources, including corpora, knowledge bases, and the related search engines.. With regard to Chinese language online resources, numerous Chinese corpora, lexicon and dictionaries, large or small, have been established and open for search and research purposes, which has given great impetus for Chinese language studies. This paper examines the development and application of the online Chinese language resources, and discusses their possible impact on linguistics and the challenges for their further development. First, it gives a brief introduction of the background of corpus development. Second, it presents an overview of the Chinese language resources constructed since the 1990s to date. Third, it uses some concrete examples to demonstrate the application of online resources in linguistic research and language teaching. Fourth, it discusses the challenges for the construction of Chinese language online resources and the difficulties in their applications. In conclusion, it suggests a closer integration of introspection-based theoretical analysis and data-driven statistical analysis to benefit language studies.

Key words　Chinese language resource; corpus; knowledge base; search engine

---

作者简介：詹卫东，男，北京大学教授，青年长江学者，主要研究方向计算语言学、现代汉语语法和中文语言知识工程。电子邮箱：zwd@pku.edu.cn。

*　本文工作得到国家重点基础研究发展计划（2014CB340504）、教育部人文社科重点研究基地重大项目（13JJD740001，15JJD740002）经费支持。北京大学中国语言文学系研究生黄思思、田骏、苗宇晶、李安然、夏雪、赵贤为中文语言知识资源的调研收集了大量材料。本文初稿主要内容在暨南大学“汉语方言学大型辞书编纂数字化建设高端论坛”（2017 年 6 月 10–12 日）上报告过，得到与会专家的宝贵意见和建议。在此一并致谢。

一、引 言

伴随着网络和计算机技术日新月异的发展脚步，在语言学研究和语言教学方面，也越来越重视大规模语言资源的作用。特别是进入21世纪的“大数据”时代之后，移动互联网、社交媒体、机器学习理论模型的飞速发展，在人工智能、自然语言处理技术领域掀起了新的热潮，语言数据资源的重要性受到了学术界、工业界以及政府相关部门前所未有的普遍重视。而近30年来，中文世界的可用语言知识资源也已经积累了相当可观的规模。本文的目的，就是在这一大背景下，梳理当前互联网环境中的中文语言知识资源，包括语料库、词库及相应的检索系统的现状，并讨论对语言学未来发展可能造成的影响。

在展开论题之前，有必要简略说明有关语言学研究范式之争的一些背景情况，这对读者深入思考和探讨语言数据资源之于未来语言学发展的意义，应该会有很大的作用。

20世纪50年代末，随着乔姆斯基（Chomsky 1957）《句法结构》一书的问世，语言学领域掀起了所谓的“乔姆斯基革命”，将语言学的研究目标确定为解释人的语言生成机制或者说是人内在的语言能力，而不再是像传统的结构主义语言学者以及功能主义的语言学者那样，把关注重点放在语言表现和语言单位的具体使用方法方面。这一研究范式的转变，突出地表现在，生成学派语言学者的研究方法或者说主要的注意力，从以往结构主义语言学者倡导的观察、描写鲜活的真实的语言现象，变成了以“内省”为主要手段的研究。前者面对的主要是语言使用中的“正例”，即人们口中实际说出来的句子，而后者则反其道而行之，开始关注语言系统的所谓“反例”，即那些人们从来不说，或者很少会说的“句子”。乔姆斯基把句法和语义分成了语言生成机制中两个独立的部门，并由此逐步设计出一整套的普遍语法理论框架（Chomsky 1981，1993，1995，2000）。语言学研究的目标，被确定为去发现人脑中内在的语法机制如何工作，生成合语法且能得到语义解释的句子，同时避免生成出不合语法或无法得到语义解释的句子。在乔姆斯基看来，要实现这个目标，靠收集人们口中说出来的真实句子是行不通的，收集再多的真实语料也无助于去发现语法系统的本质规律。在2004年的一次访谈中，在被问到如何看待语料库语言学的迅猛发展态势以及对词汇语义学、构式语法研究的积极影响时，乔姆斯基一如既往地对基于语料库的经验主义语言学研究方法不留情面地批评：“语料库语言学毫无意义。这有点像是物理学和化学研究不基于实验，而是拿着录像机去记录世界上发生的事情，尽管可以收集非常多的录像带，也许能从中看出些什么一般规律。但是，如你所知，科学并不是这么开展研究工作的。”（Andor 2004）

本文在概括近30年来中文语言知识资源发展的整体面貌时，试图始终保持一种反思的态度，即把上述乔姆斯基所代表的理论语言学界的认识作为一个思考背景，去探讨语料库等资源对于语言研究的意义，我们相信，这样的态度，对于未来的中文语言知识资源建设以及资源的利用，可能是更有利的。

本文第二部分概述我们调查中文语言知识资源的具体做法以及对当前中文语言知识资源的宏观认识；第三部分分类列举一些重要的有代表性的语言资源，并示例其应用价值；第四部分讨论中文语言知识资源未来发展面临的挑战及其对语言学研究和语言教学的可能影响；第五部分是结语，说明内省逻辑分析与语料统计分析相结合的中间道路应是未来发展较好的路径。

二、现状：中文语言知识资源调查概述

本文对中文语言知识资源的调查主要依靠互联网搜索引擎（如谷歌和百度），以及网上一些知名学者、汉语教师整理的在相关课程教学（如语料库语言学、汉语教学课程）中作为参考的语言资源列

表等[①]。考虑到汉语书面语有简体和繁体两个系统，另外在非汉语环境中的中文资源情况可能有所不同，我们的调查是分地区进行的，主要包括：（1）中国大陆；（2）港台地区；（3）亚洲其他地区（日本、韩国、新加坡等）；（4）北美；（5）英国；（6）欧洲大陆。搜索查询的关键字包括："语料库 / corpus""词典 / dictionary""词库 /lexicon""中文 / Chinese，Mandarin"，以及知识库的应用，如"自然语言处理 / natural language processing""机器翻译 / machine translation""中文语言教学 / Chinese as a second language education"等。网络搜索结果比较庞杂，资源类型多样。[②]其中语料库方面，按照规模、类型多样性程度、影响力程度的不同，大致可以分为3个层级：（1）提供丰富的多种类型语言知识资源管理和服务的门户型网站，如语言资源联盟（以下简称LDC）、中文语言资源联盟（以下简称Chinese LDC）等。（2）提供在线查询服务的大型独立语料库系统，比如北京大学的中国语言学研究中心语料库（以下简称CCL语料库）、北京语言大学的汉语语料库（以下简称BCC语料库）、HSK动态作文语料库、台湾地区"中研院"的系列中文语料库、Word Sketch Engine、WebCorp Search Engine等同时包含汉语和其他语言的语料库查询系统。（3）本领域代表性研究机构和学者研制的服务于中文研究和教学、中文信息处理的各类专项语言资源，一般规模相对小一些，比如加州大学洛杉矶分校汉语书面语料库（简称UCLA汉语书面语语料库）、香港中文大学的香港双语儿童语言资料库等。

知识库（lexicon和dictionary）资源方面，也大致可以分为3种情况：（1）面向学术界和信息产业界的以中文为内容主体的知识库，其中有代表性的如北京大学计算语言学研究所研制的"现代汉语语法信息词典""现代汉语语义词典""中文概念词典"等，台湾地区"中研院"语言学研究所研制的中文词汇网络（Chinese Wordnet）、中文双语本体知识，中国机器翻译学界知名学者董振东研制的知网（HowNet）知识库，等等。（2）互联网上众多的面向一般社会公众使用的词库、字典网站资源以及移动端应用程序，在互联网时代中文的传播、教学和语言文字应用中发挥着重要作用，比如：中国台湾国科会数位博物馆的"搜文解字"、台湾地区的免费字典门户网站线上字典导览网[③]、中国大陆的汉典网[④]等，其中汉典网集成了可以方便普通民众查询《康熙字典》《说文解字》等常用古籍字典类工具书。（3）为展示世界语言的丰富性，为语言学理论研究、语言类型学研究等基础性研究提供数据支持的专业网站和数据库，在收集世界范围内的语言数据中也包含了汉语（及其方言）的数据。如语言类型学协会网站[⑤]、世界语言结构地图在线数据库网站[⑥]（以下简称WALS），以及一些专门收集世界语言语音数据的数据库（UPSID[⑦]、P-Base[⑧]）等。

如果结合现代语料库和语言知识工程发展的早期历史来认识上述调查结果，可以更深切地认识到中文语料库与知识库资源在近30年来的飞速发展。冯志伟（2016）在为《应用语言学中的语料库》一书写的导读中介绍了从20世纪60年代以来英语语料库的发展简史，以及中国语料库从20世纪80年代以来的发展状况。早期是以手工收集为主的小规模样本语料，主要服务于字频、词频统计等任

① 比如美国加州长滩大学谢天蔚收集的中文教学网站资源有216个，分为21类，参见 http://web.csulb.edu/~txie/pcr.htm。

② 主要的中文语言知识资源网址参见"语言战略研究"微信公众号。

③ 参见 http://www.tradict.net/lang_guoyu.php。

④ 参见 http://www.zdic.net/。

⑤ 参见 http://www.linguistic-typology.org/resources.html。

⑥ WALS的全称：World Atlas of Language Structures，参见 http://wals.info/。

⑦ UPSID是美国加州大学洛杉矶分校的语音数据库（全称为UCLA Phonological Segment Inventory Database），可以通过世界语言语音数据库门户网站 http://phoible.org/ 访问查询。

⑧ P-Base是加拿大渥太华大学的Jeff Mielke博士构建的语音数据库，参见 http://aix1.uottawa.ca/~jmielke/pbase/。

务，今天则是可以从互联网海量数据中获取既包括文本又包括语音数据在内的多媒体语料。而且，随着计算机存储能力和计算能力的提高，语料规模已达到GB级甚至TB级，比如Google Ngram统计了1500–2008年超过500万册的Google Books文本语料中的单词串到五元词串的历年频次，[①]受数据库规模限制，Ngram中只收录了出现在40本书以上的那些词串，即便这样，2009版的中文Ngram数据压缩后也达到共计7882MB的规模（计1510个压缩文件），[②]其中在LDC上发布的2009版5-gram中文数据规模约为30GB。[③]除资源规模巨大外，语料库类型丰富多样。从标注深度方面来看，标注层次由浅到深，形成了包括分词和词性标注语料库、句法树库、语义角色标注语料库、语义依存关系标注语料库、篇章关系标注语料库等为代表的多级加工语料库；从语料广度方面来看，大型语料库中既有单语语料，也有双语和多语平行语料；语体风格方面有书面语和口语，传统语言和网络语言，儿童语言和成人语言，现代汉语和古代汉语。此外，除母语者语料外，也有一定规模的面向第二语言教学研究的中介语语料库。知识库方面也呈现多样化发展的态势。在通用知识库（如“现代汉语语法信息词典”、知网等）之外，也有越来越多的更具针对性的中文知识库，如大连理工大学面向文本情感分析的“情感词汇本体”，北京大学针对名词（短语）的语义分析开发的名词物性结构知识库等（袁毓林，李强2014）。

中文资源取得如此迅速的发展，一方面是整个学术界对语料库和基于数据的经验主义研究方法日益重视的结果，另一方面，从实践层面来看，像LDC这样的语言资源收集和传播平台起到了相当重要的作用。LDC成立于1992年，其中文相关语言资源从1994年到2016年，共164种，呈现不断增长的趋势。按年份统计其语言资源种数得到的折线图如下所示：

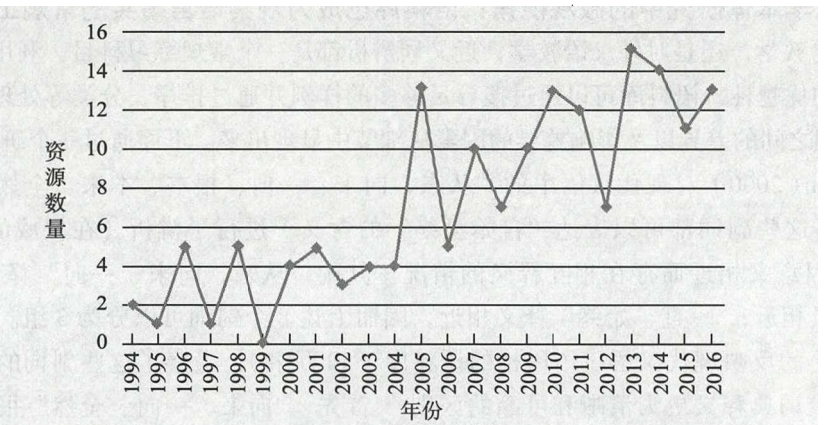

图1　LDC中文语言资源（1994–2016）种数按年统计图

在LDC中文资源中，文本形式的资源有109种，语音形式的资源54种，视频形式的资源1种。文本形式的资源中内容主要为词汇级信息的资源有5种，短语和句子级资源19种，语篇级资源7种，其余文本类资源78种，许多是服务于机器翻译的双语语料或机器翻译评测语料，以及用于信息检索、语言建模等应用目的的语料。文本语料的数据来源主要是报纸新闻、杂志新闻等。语音语料的数据来源主要是电话交谈、广播会话、广播新闻等。早期的语料以传统纸媒为主，后期则增加了不少来自网络（比如博客、论坛、新闻组等）的语料。

① 参见https://en.wikipedia.org/wiki/Google_Ngram_Viewer。在线检索可访问https://books.google.com/ngrams，下载数据可访问http://storage.googleapis.com/books/ngrams/books/datasetsv2.html。

② 参见https://archive.org/download/google_ngrams-chinese-simplified。

③ 参见https://catalog.ldc.upenn.edu/LDC2010T06。

尽管中文资源已经取得的成就值得充分肯定，但也还是有一些问题需要引起注意。在LDC的164种中文资源中，其中词库（知识库）类资源只有4种，其余160种均为语料库性质的资源。整个中文资源的情况也大致如此，即语料库资源相对较多，知识库资源较少。因为后者需要更多的语言学理论知识的注入，是专家精细知识的成果，建设成本很高。在语料库资源中，大规模的资源都是容易自动获得的、只需要较少的人工干预甚至不需要语言学专家知识干预的语料类型（如双语对齐语料、一般文本语料等），而深加工语料库，特别是涉及语义分析的语料库规模目前还不够大。此外，像中介语语料库、多媒体多模态语料库等支持对外汉语教学研究、篇章会话分析研究的专项语料库，也还有很大的发展空间。[①] 在语料规模扩张相对容易的今天，语料的质量问题一定程度上可能容易被忽视。以前述Google Ngram数据为例，有人调查了中文Ngram数据，发现1970-2008年间的数据比较可靠，而之前的数据存在明显的问题，究其原因，是因为对早期中文书籍通过OCR数字化得到的电子文本错误较多，严重影响了自动统计所得数据的质量。[②]

三、应用：中文语言资源在汉语研究和教学中所能扮演的角色

利用大规模语料库开展研究工作，成为越来越普遍的现象。尤其是在大数据和机器学习方法成为自然语言处理主流方法的今天，以语料库为代表的语言资源对自然语言处理的作用已是常态，本文不赘述。下面着重讨论语言资源在语言本体研究、汉语教学研究和语言社会价值方面的作用。

（一）语言学本体研究中的微观视角：语料库已成为观察语言事实的常规工具

无论是母语教学，还是对外汉语教学，近义词辨析都是一个常规练习科目。利用语料库对近义词进行辨析有显然的优越性。语料库可以通过搜寻足够多的样例并通过排序、分类等处理，弥补个人语感的不足，将近义词之间的差异以及影响差异的因素从细节中显现出来。下面通过一个研究案例来具体说明。

Tao Hongyin（2000）对现代汉语中的“从来、向来、一向、根本、本来、全然、一直、始终”等8个近义副词（这些副词都可以表达“自始至终”的含义）进行了辨析。在权威的《现代汉语词典》中，这些副词因意义相近而存在相互释义的情况。其中“从来、向来、一向”释义相近；“根本、本来、全然”释义相近；“一直、始终”释义相近，因而上述8个副词可以分为3组。作者利用了一个现在看来规模较小的反映现代汉语口语风格的语料库（50万字），考察了这些副词的用法。结果得到了这些词语之间较词典释义更为清晰和可靠的区别。首先，“向来、一向、全然”很少用于口语中。其次，“根本”跟“本来”相比，前者全部用于否定语境中，后者则是中性的。“一直”跟“始终”相比，前者一般跟在表达时间范围的小句之后，后者无此限制，可以独立使用。最后，“从来”跟“根本”都用于否定语境，但从语义模式上讲，“从来”后接的否定表达本身是说话人肯定的“积极”行为或事实，因而从说话人立场而言，“从来”小句表达的并非消极而是积极的语义。“根本”后接的否定表达则不是说话人肯定的“积极”行为或事实，而是说话人的直接否定（消极）语义。例如：

1）真的，我当时**根本**没想过会有今天这种事儿。（原文第5例）

2）银行有律师，靠我可打不赢官司，我从**来**不打官司。（原文第26例）

3）一九四二年回国后，我一**直**在东北。（原文第17例）

4）“她这是胡说，她**始终**不明白，‘按能承包’也得分地。……”她丈夫说。（原文第21例）

① 截至2016年，LDC资源种数为787种，其中英语资源434种，是汉语资源（包括一些方言）的两倍多，无论从数量还是从类型丰富性角度，中文语言资源离英语资源都还有不少距离。

② 参见http://digitalsinology.org/when-n-grams-go-bad/。

尽管以上通过语料库实例获得的观察结论并不能说已经完美地解决了这8个近义副词的全部辨析问题，但无疑比传统的基于语言学家、词典编纂者有限语感的辨析有了明显的进步。传统的一些权威词典也观察到“从来”和“根本”对否定语境的明显的选择倾向性，但是，并没有进一步去区分表面“否定”形式背后的实际语义差异，即“从来”的否定实际上是说话人积极的正面评价（如“不打官司”是好事，是应该的），“根本”的否定则可能关联消极的负面评价（如“没想过会有今天这种事儿”是不对的，不应该的）。此外，基于语料库的观察无疑可以拓宽人们在辨析近义词用法时的视野，正如作者所指出的，传统的研究往往把副词的观察范围限于单句，但语料中的实例则可以在篇章范围内展示更为真实的差异。

5）我没上过大学，也不想上大学，我根本没必要去念那些本来就是错误的理论。（原文第7例）

6）我一直在看中文书。（原文第30例）

例5显示，“根本”用于一个连续的语气不断加强的小句序列中，“根本”所在的小句居后，承接前面的小句并起到在后加强语气的作用，“本来”则没有这种特点。和“根本”相比，“本来”没有在连续小句序列中居后的限制，可以作为独立小句[①]中的副词使用。例5中的“根本”和“本来”不能互换位置。例6作为独立句子并无语法错误，但在实际使用中，“一直”往往要求前面有表达时间范围的成分，从篇章（而不是单独的句法）角度来看，母语者很少会像例6这样来使用“一直”。

Tao（2000）是早期语料库在汉语研究中的典型研究案例。随着语料规模的扩大以及分词和带有词性标注信息语料的出现，人们可以利用像互信息（Huang *et al.* 2005）等度量随机事件相关性的统计指标[②]来测量词语之间的共现（搭配）模式和强度，从而更加量化地呈现近义词之间的区别，如Hong和Huang（2006）利用中文Gigaword语料库（规模为11.1亿汉字）开展的关于汉语中“吃”跟“喝”的比较研究。此外，中介语语料库中包含了学习者词语使用偏误的例子，利用中介语语料库的数据来辅助近义词辨析在近期也受到汉语研究者的重视，如Hong（2014）在中国台湾师大一个3亿字的中介语语料库上对“便利”跟“方便”的偏误调查。

（二）语言学本体研究中的宏观视角：标注语料库作为对语法结构进行量化研究的工具

上文是针对具体的语言现象，利用大规模语料库进行调查。语料库还可以在宏观层面为认识语言系统的整体性质提供帮助。下面以汉语句法结构歧义的定量分析为例来说明。

传统的关于自然语言歧义的分析往往是针对具体的歧义实例，比如汉语中经典的歧义结构实例“咬死了猎人的狗”，从词性序列来讲，可以抽象为“v + n + 的 n”，符合这个词性排列的词语串，都有可能分析为vp或np，例如“发现 敌人 的 哨兵”“怀疑 张三 的 老师”等。除了这类现实中的对人来说确实具有两种语义理解的歧义外，对计算机句法分析来说，还存在更多的可能分析出两种句法结构，但具体实例不一定有两种语义解释的格式。比如：“pp vp vp”这样的短语序列，可能分析为下面两种结构。

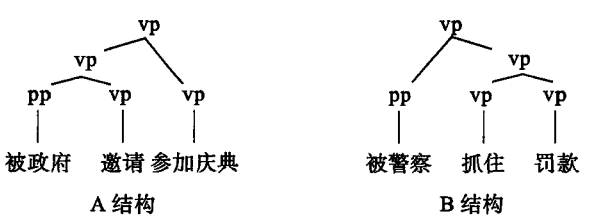

图2 “pp vp vp”的两种句法结构

① “本来”所在的小句在例5中作为内嵌的定语从句使用。

② 参见 https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/。

这两种结构对应的实例“被政府邀请参加庆典”和“被警察抓住罚款”各自都只有一种语义，没有歧义，只是在计算机对这两个实例进行分析的时候，由于缺乏有关词例（如“罚款、邀请、抓住”等）的知识，既可能分析为左图的树结构，也可能分析为右图的树结构，而且两种结构下，根节点均为vp，这类歧义称为“内含型歧义”，不同于上面的“v + n + 的 n”格式歧义，后者对应的两种结构根节点不一样，分别为vp和np，属于“外显型歧义”。对于计算机自动句法分析来说，要得到像上面这样的树图分析结果，所依赖的语言知识主要是“vp → pp vp”“vp→ vp vp”这样的上下文无关文法规则。对于中文的句法分析而言，有多少这样的规则？在给定规则情况下，如果分析任意的多项短语排列格式（如“np vp np”“pp vp vp”等），可能碰到多少有歧义的格式？歧义的程度如何？这些无论对于计算机的自动句法分析，还是对于认识汉语的语法结构系统，都是很值得思考的问题。我们曾在手工规则的基础上利用计算机程序对汉语的3项短语序列进行过考察，得出了在给定规则集下文法系统歧义程度的量化分析结果（詹卫东等1999）。在标注了100多万字现代汉语句子的树结构后，我们从树库中抽取出更能够代表现代汉语真实语料组合模式的规则，重新进行了歧义度量化分析，得到的结果跟之前基于手工规则所做的歧义度结果对比如下表所示：

表1 现代汉语3项排列短语序列歧义格式及歧义度统计表

短语序列 \ 文法规则集			手工文法（246条规则）		树库文法（657条规则）	
			数量	比例	数量	比例
短语序列（全排列）			9^3=729	100%	11^3=1331	100%
a. 不合法的短语序列			360	49.38%	661	49.66%
合法的短语序列	b. 无歧义的序列		84	11.52%	150	11.27%
	c. 有歧义的序列	c1. 外显型歧义	194	26.61%	386	29.00%
			平均歧义数：6.55		平均歧义数：6.61	
		c2. 内含型歧义	91	12.48%	134	10.07%
			平均歧义数：2.37		平均歧义数：2.29	

跟手工规则相比，树库规则增加了两个短语范畴标记（从原来的9类增加到11类）。从这11个语言范畴（如ap、dp、np、pp、vp等）中任取3项组合为一个短语序列（如“np vp np”），可得到共计1331个序列，将全部短语序列交由计算机句法分析程序（parser）进行分拣处理，得到上表中a、b、c1、c2四类序列的数量。其中c1、c2分别是外显型歧义序列和内含型歧义序列。从对比结果来看，基于树库文法和手工文法两个数据源，歧义序列的数量占比基本相当，没有明显的差异。因为树库标注的语法体系是在原有的手工文法200多条规则基础上扩展而来，树库文法跟手工文法相比，规则数量增加的主要贡献是提高了对现实语料句法结构的覆盖程度，与此同时并没有带来歧义程度的明显增加。歧义度统计结果说明两个文法的基本范畴和整个语法体系决定了文法内在的歧义程度。就上表中b和c类的数量对比而言，有歧义的序列数量上是绝对多数，说明目前树库短语范畴的划分，对于计算机分析的应用目的而言，有很高的歧义性，仍有很大的优化空间。

基于树库语料还可以抽取多种语法结构知识，以多种方式来提取真实语料中的歧义序列（包括以短语范畴标记的歧义序列和以词类范畴标记的歧义序列），计算歧义序列的熵值，等等（詹卫东2013）。此外，语言学家基于依存树库（dependency treebank），可以对语言结构的复杂度进行定量研究（比如Liu *et al.* 2009；Jing & Liu 2015；等）。这些考察都超越了传统上以具体的语法点或某个具体的歧义格式作为语法研究对象的研究方式，能够真正从宏观层面，对汉语语法现象、歧义现象、语义

结构复杂度做全局式的俯瞰。这类研究工作，都得益于现代深度加工语料库和计算机自然语言处理技术的进步。

（三）语言类型学及跨语言对比视角

上面谈的主要是基于语料库的语法本体研究工作。下面再看一项基于语言知识库的跨语言的研究工作。WALS是语言类型学数据库。目前其“语言库”（Languages）有2679条记录，“特征库”（Features）有192条记录，每个特征取值在2-28个值之间。数据库中也包含了有关汉语的丰富的特征信息。冯胜利（2015）对声调、语调和语气词之间的关系提出了新的观点：（1）有声调语言都有句末语气词。（2）无声调语言没有句末语气词（特殊情况除外）。（3）声调越多越复杂，句末语气词也越多越复杂。（4）句末语气词越多越复杂，语调就越简单越贫乏。（5）某一语言从非声调语言变成声调语言，会带来“从无句末语气词到有句末语气词”的平行发展。

叶述冕（2016）利用WALS数据，基于206种语言声调系统的表现和是非疑问句表达手段的数据，分析了声调、语调和语气词之间的类型学意义上的相关性。对冯胜利（2015）的看法进行了分析，跨语言的数据调查既有支持冯的看法的部分，也有不支持冯的观点的事实。叶文的主要结论包括：（1）声调的有无跟语气词的有无没有显著的类型学相关性。声调的有无可能跟句末语气词的有无存在类型学相关性：无声调的语言倾向于无句末语气词，但有声调的语言则不一定。欧亚大陆的有声调语言倾向于使用句末语气词，美洲的有声调语言则倾向于不使用句末语气词，非洲和巴布亚-南岛区域的语言则无明显倾向。（2）语调和语气词之间可能存在关联，但“语气词可分析为语调的一种变体”仍需要寻求语言事实（数据）的支持。

叶文的研究是一个很好的示例。以往语言学理论研究依据少量语言调查的材料做出假设和理论分析，在世界语言数据不断积累、描写日益精细化、数据检索日益便利的今天，基于大规模跨语言的数据调查，可以为语言类型学、为人们了解人类语言的更多奥秘提供更好的支持。

此外值得一提的是，跨语言语料库目前主要还是以双语平行语料库为主。大多数双语平行语料库的构建目的，都是为了训练机器翻译系统。不过，除了用于支持机器翻译的模型训练外，在语言对比研究中，双语语料库也可以发挥重要作用。比如柏晓静、詹卫东（2006）对汉语“被”字句表被动与英语被动句差异的研究，[①] 马千（2011）从英汉对比来分析汉语“这、那”的定指功能，等等，都是基于汉英句子对齐语料库开展的研究工作。

（四）社会语言学视角

从2004年到2008年，由教育部语言文字信息管理司牵头，先后成立了国家语言资源监测与研究中心[②]的6个分中心：平面媒体语言中心、有声媒体语言中心、网络媒体语言中心、教育教材语言中心、少数民族语言中心、海外华语研究中心。随着这6个分中心的启动与工作的展开，语言信息作为一种公共资源的意识受到学术界和社会各界越来越多的关注。而这些中心所建设的大型动态流通语料库，不仅在语言研究与教学领域发挥作用，而且为政府和社会公众了解当代中国语言实际运用方方面面的状况提供了坚实的数据支撑。从2006年开始，教育部语信司每年都组织编写《中国语言生

① 汉语“被”字句跟英语被动句的对比研究显然对机器翻译也有重要的意义。柏晓静、詹卫东（2006）曾测试过一些机译系统对英语被动形式的汉译情况。结果显示被动句的机器翻译问题比较多。即便是在机器翻译技术已经引入神经网络方法的今天，测试句中的英语句子“They love to read and be read to.”机器译成中文仍然是“他们喜欢阅读和阅读”。

② 参见 http://www.moe.gov.cn/s78/A19/A19_xglj/201309/t20130929_158028.html。

活状况报告》[①]，全面反映当年的中国社会语言生活实态，为社会公众提供语言信息服务，为政府制定相关的语言文字政策与语言规划提供依据，在引领社会语言生活走向更健康更和谐方面起到了良好作用。比如每年的报告都记录了当年的年度热字、热词、热语，能够反映社会百态的活跃的语言现象，像“互联网+、阅兵蓝、重要的事情说三遍”等众多的新词新语都被忠实地记录下来。有媒体统计过，《中国语言生活状况报告》在2006到2015年发布的报告中，记录了像“微博、中国梦、微信、正能量”等体现时代脉搏的新词语共计5514个。[②]此外，每年举办的“十大新词语、十大流行语、十大网络用语”评选，吸引了全社会的关注和参与，为推动社会各界参与语言资源建设、支持中国语言资源发展起到了积极作用。

除政府联合学术界力量积极推动中国语言资源建设外，海内外学术界也有不少学者关心不同地域的汉语（或称华语）之间的联系和区别。邹嘉彦主持建设的香港城大泛华语共时同题语料库（简称LIVAC语料库）从1995年开始一直收集泛华语区8个城市的报纸语料，迄今已达到25亿字的规模。这样的语料资源为比较不同地区汉语书面语的变异情况提供了丰富的资料。

互联网为语料收集提供了极大的便利，[③]詹卫东、陶红印（2016）利用网络爬虫在短时间内就构建了一个近亿字的北美汉语网络语料库，通过对该语料库中多个词语变项以及语法项目的考察，可以看到，尽管海外汉语中仍有受到英语影响的特有表达形式以及带有南方方言色彩的说法[④]，但从总体趋势而言，海外华语跟中国的汉语越来越趋向同步发展。互联网的发展使身处“地球村”中的各地汉语的语法变异更多的是在趋同，而不是趋异。新兴语言现象因新媒体的传播可以迅速波及全球，这是社会语言学研究汉语面貌（特别是社会方言变异）时应该重视的一个重要外部因素。

四、挑战：语言资源的未来发展及其对语言研究和教学的影响

无论是理论层面的关于汉语本体知识的研究、跨语言的语言类型学研究、针对儿童语言的习得研究，还是应用层面的词典编纂、汉语作为第二语言的教学、语言政策的制定和规划等方面的研究，各种类型的中文语料库和知识库资源都在发挥着积极的重要的作用。一方面，与资源配套的辅助的检索工具功能越来越强大（如BCC语料库、CCL语料库都支持复杂的模式匹配、跨词检索等），界面也越来越友好，便于使用；另一方面，正如上文已经提到的，随着语料规模的扩大、加工成本的提高、维护的难度加大，资源的质量问题也需要引起足够的重视。特别是目前有的语料库系统大量的原始材料来自互联网文本，原文错误较多又难以进行人工校对，同时存在大量重复，用这样的资料来支持研究，就要特别小心。以一个利用搜索引擎（如微软Bing搜索）接口实时从互联网文本中查询网页，再从网页中提取包含用户关键字的句子返回结果的语料库系统[⑤]为例，查询汉语词串“虽然短暂”，

① 又称“语言生活绿皮书”，可参见李宇明（2007）的详细介绍。也可浏览教育部语信司官方网站了解每年中国语言生活状况报告发布的情况，http://www.moe.edu.cn/s78/A19/yxs_left/moe_813/s237/。

② 参见 http://www.jyb.cn/china/gnxw/201510/t20151017_640001.html。

③ 谷歌公司利用其强大的网页搜索和数据处理能力，支持了一个名为全球事件、语言与语调数据库（简称GDELT）的项目，收集全球超过100种语言（包括汉语）的新闻报道，实时获取全球各地新闻内容，并对文档进行情感/褒贬指数（sentiment index）计算，详见 http://www.gdeltproject.org/。这样庞大的实时语言资料库可以为社会语言学研究提供支持。

④ 例如北美汉语中有“说回那道栗子汤”这样的“V-回”动趋式带宾语的用法，这是普通话中没有的语法结构（搭配）。普通话中非位移动词带趋向补语（如“V-回来”）跟宾语共现的例子很少，比如像已经固化的表达形式“话又说回来”。

⑤ 参见 http://www.webcorp.org.uk/live/index.jsp。

就返回了大量重复的句子：

表2 在WebCorp语料库检索系统中查询“虽然短暂”返回检索结果示意

……
10) http://www.le.com/tv/89028.html
Text, Wordlist, text/html, UTF8 (Content-type), 2017-01-01 (Copyright footer)
29:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:25 第2集
30:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:24 第3集
31:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:25 第4集
32:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:24 第5集
33:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:25 第6集
……

汉语的虚词是语法研究也是语言教学的难点。对于近义实词的辨析，像Word Sketch这样的大规模语料库可以提供一些自动分析的参考结果，比如上文提到的基于Word Sketch对动词“吃”和“喝”、形容词“方便”和“便利”的对比研究等。但对于虚词的辨析，就要困难很多。比如用Word Sketch的Sketch Difference功能来检索“一直”和“始终”、“根本”和“从来”，系统就无法像处理“吃”跟“喝”那样自动给出比较结果，而是返回“数据量不足以支持得到具有统计显著性的分析结果”。如果想利用Word Sketch查询单个词的搭配模式，对于汉语虚词而言也存在较多问题。比如搜索“的”，尽管“的”的实例无疑是最多的，[①] 但Word Sketch返回结果仍然是“Insufficient data for 的”（无法给出“的”的搭配模式）。此外，该语料库检索系统基于的数据源是经过计算机中文自动分词的语料，因自动分词带来的错误也会影响这些系统对要求较高的检索需求的支持。比如以Search功能搜索单词“了”，返回结果是19 471 367个（每百万字9242.80次）。再以Word Sketch功能搜索“了”，可以返回“了”的搭配模式，但其实是错误的结果，模式中列出的实例都是作为补语动词的“了（liao）”，而不是作为助词的“了（le）”。原因显然是语料库分词处理中并没有区分这两个字形相同但实际语法性质完全不同的语言单位。

除近义词辨析外，利用语料库，特别是新兴网络语言语料库对新词或词语新义进行研究，也能很好地展现语言资源的价值。比如现代汉语动词“刷”在权威的《现代汉语词典》中有两个动词义项[②]：

②动用刷子清除或涂抹：～牙 |～鞋 |～锅 | 用石灰浆～墙。③动除名；淘汰：他不守劳动纪律，让厂里给～了 | 他报名参军，因视力不合格被～了下来。

考察北京大学CCL语料库和北京语言大学BCC网络语料可以发现“刷”的许多新用法，都是传统的词典义项覆盖不了的。比如：刷微博、刷电影、刷屏、刷粉丝、刷好感、刷系统、刷人品。要进一步分析这些新用法中“刷”的义项，以及这些新义项的生成机制，仅靠统计计量分析就无法做到，仍然需要利用传统的认知语言学中的概念整合理论来分析“刷”自身的语义要素跟典型语境中的名词性成分的语义要素整合出“刷”的新义项：（1）长时间、大量高效地做某事（刷微博）。（2）按照顺序从头到尾经过某空间（刷马路）。（3）通过重复（快速）的行为获得商品（刷票）。（4）更新（刷系统）。（5）通过特定行为显示人的品质或特性（刷优越感）。这种经过深入分析和归纳得到的语言知识，是语言知识库的内容，目前还没有可靠的方法从语料库中自动获得这样的知识。这是未来语言资源系统努力的方向之一。

① 用Word Sketch系统的Search功能搜索单词“的”，可以返回99 838 775个结果（每百万字47 392.00次）。

② 摘自《现代汉语词典》（第7版）。“刷”还有名词义项（如“刷子”），这里省略了。

五、结 语

中文语言资源在近30年的发展可以说是成绩斐然，有众多的在线语料库、知识库和检索系统对理论研究和教学应用提供着切实的支持。不过，未来需要提高的地方也很多。比如在移动设备上提供更多可用的语言资源及应用程序，为语言数据提供更好的可视化展示系统，等等，都还需要大力发展。另外，语言资源的建设应考虑层次性，面向不同人群（如语言小学研究者、语言教师、语言学习者、普通爱好者等）的语言资源及相应的应用系统需要更加丰富。

随着语言数据积累的规模日益扩大，机器学习算法的不断改进，利用计算机工具，基于大规模自然语言资源数据为人类提供各种语言服务的能力也越来越强了，比如中国台湾地区“清华大学”张俊盛教授主持开发的Linggle系统，利用英语文本大数据来支持二语者的英语写作，可以抽取大量的英语搭配实例供用户选择，为非母语者写出更地道的英语论文提供帮助（Boisson *et al.* 2013）。像这样的基于语言资源的应用今后会越来越多，而且会越来越好用。另外，如果语言资源的应用跟语言学理论结合起来（比如上面举的有关“刷”的新义项发生机制的分析），可以让语言资源发挥更大的作用。正如Levin *et al.*（1997）指出的那样，在利用语料库进行英语词典编纂的时候，如果结合词汇语义学的理论指导，可以让语料库中动词看似杂乱的用法系统化，从而在词典中做出更科学的描写。

回到本文引言中提出的问题，语料库之类的资源对于语言学理论研究到底意味着什么？学术界对语料库之于语言学理论的价值有两种看法（Gries 2012）：（1）语料库驱动的语言学（Corpus-driven linguistics），这种语言学构建的理论完全建立在语料数据基础上，拒绝标注语料库（即预设某种语言学理论）。（2）基于语料库的语言学（Corpus-based linguistics），这种语言学使用标注语料库，对现有语言学理论进行检验和改进。显然，前者是一个有追求的语言学家的理想，但现实中，后者是目前语料库语言学的现状。或许在将来的某一天，随着计算机自然语言处理技术的不断发展，通用人工智能（AI）的日益强大，语言资源会成为语言学的“阿拉丁神灯”，它会帮助有语言需求（比如学外语、机器翻译等）的人们，告诉人们“这是什么（What），怎么去做（How）”。但正如乔姆斯基所指出的，科学的主旨是回答“Why”的问题[①]。目前的语言资源以及基于资源的各种计算机程序，还无法具有像人一样的洞察力，来回答“Why”的问题。从这个意义上讲，人的内省洞察和逻辑能力跟语言资源的海量数据相结合，应是对未来语言学科发展具有建设性的积极态度，把基于内省的语言学理论分析跟基于真实语言数据的分析对立起来，互相贬低对方的价值，并不可取。

参考文献

柏晓静，詹卫东 2006《汉语“被”字句的约束条件与机器翻译中英语被动句的处理》，载邢福义《汉语被动表述问题研究新拓展》，武汉：华中师范大学出版社。

冯胜利 2015《声调、语调与汉语的句末语气词》，《语言学论丛》51辑。

冯志伟 2006《〈应用语言学中的语料库〉导读》，载霍斯顿《应用语言学中的语料库》，北京：世界图书出版公司。

① 2011年5月3–5日，麻省理工学院举办了“大脑、心智与机器”（Brians，Minds and Machines）专题研讨会（麻省理工学院建校150周年系列活动之一）。乔姆斯基在发言中表达了对统计方法在语言学研究中作用的质疑。之后谷歌科学家Peter Norvig撰文对乔姆斯基的质疑发表详细评论（见http://norvig.com/chomsky.html）。Norvig的文章中谈到了对科学（包括语言学）的性质以及目标的认识，对乔姆斯基的看法提出了批评。

李宇明 2007《关于〈中国语言生活绿皮书〉》,《语言文字应用》第1期。

马 千 2011《从英汉对比看汉语“这”“那”的定指表达》,北京大学硕士学位论文。

叶述冕 2016《声调、语调、语气词之类型学相关性》,《语言学论丛》53辑。

袁毓林,李 强 2014《怎样用物性结构知识解决“网球问题”?》,《中文信息学报》第5期。

詹卫东,常宝宝,俞士汶 1999《汉语短语结构定界歧义类型分析及分布统计》,《中文信息学报》第3期。

詹卫东 2013《基于大规模中文树库的汉语句法知识获取研究》,载郑秋豫《语言资讯和语言类型》,台北:“中研院”。

詹卫东,陶红印 2016《北美书面汉语语法特点探析——基于互联网中文文本的考察》,《全球华语》(*Global Chinese*)第1期。

Andor, J. 2004. The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1(1), 93–111.

Boisson, J., T. Kao, J. Wu, *et al.* 2013. Linggle: A web-scale linguistic search engine for words in context. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 139–144.

Chomsky, N. 1957. *Syntactic Structures*. Hague: Mouton Publishers.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. 1993. A minimalist program for linguistic theory. In K. Hale and S. Keyser (eds.), *The View from the Building 20: Essays in Linguistics in Honour of Sylvain Bromberger*. Cambridge: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge: MIT Press.

Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Gries, S. 2012. Corpus linguistics, theoretical linguistics, and cognitive psycholinguistics: Towards more and more fruitful exchanges. In Joybrato Mukherjee and Magnus Huber (eds.), *Corpus Linguistics and Variation in English: Theory and Description*. Amsterdam: Rodopi.

Hong, J. F. and C. R. Huang. 2006. Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, November 1–3, 2006, Huazhong Normal University, Wuhan, China.

Hong, J. F. 2014. Chinese near-synonym study based on the Chinese Gigaword Corpus and the Chinese Learner Corpus. In *Chinese Lexical Semantics,15th Workshop, CLSW 2014*, Macao, China, Volume 8922 of the series Lecture Notes in Computer Science, Springer International Publishing, 329–340.

Huang, C. R., A. Kilgarriff, Y. Wu, *et al.* 2005. Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 48–55.

Jing, Y. and Liu Haitao. 2015. Mean hierarchical distance augmenting mean dependency distance. in *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), Uppsala, Sweden, August 24–26, 2015, 161–170.

Levin, B., G. Song, and B. T. S. Atkins. 1997. Making sense of corpus data: A case study of verbs of sound. *International Journal of Corpus Linguistics* 2, 23–64.

Liu, Haitao, R. Hudson, and Z. Feng. 2009. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory* 5(2), 161–175.

Tao, Hongyin. 2000. Adverbs of absolute time and assertiveness in vernacular Chinese: A corpus-based study. *Journal of the Chinese Language Teachers Association* 35(2), 53–74.

责任编辑:丁海燕