立冬篇

北京大学 CCL 语言资源建设概况

——语言知识数据化和可视化

报告人

詹卫东，北京大学中文系教授。主要从事现代汉语形式语法、语言知识工程与中文信息处理、语言文字应用方面的研究。代表性成果有专著《面向中文信息处理的现代汉语短语结构规则研究》，国家语言文字标准《出版物上数字用法》；参编《计算语言学概论》《自然语言处理》《现代汉语》等多部教材；已在国内外学术刊物和会议上发表 90 多篇论文。

时　间

2023 年 11 月 8 日（立冬日）

地　点

北京大学数字化实验教学中心

主持人

郭琳，北京大学社会科学部副部长

交流嘉宾

马　皓，北京大学计算中心主任

范雪松，北京大学计算中心高级工程师

杨　加，北京大学计算中心人工智能应用工作组组长

苏　琪，北京大学外国语学院长聘副教授

高　山，北京大学外国语学院助理教授

成　沫，北京大学外国语学院助理教授

项目成果网址

http://ccl.pku.edu.cn:8080/ccl_corpus

主持人语

郭　琳
（北京大学社会科学部副部长）

各位老师大家好，欢迎来参加本期节气沙龙，这次的主题是关于语料库建设。自学校数字人文主题年以来，社会科学部致力于建设全校层面的文科数智化平台，就是为了打破学科壁垒，整合分散的学术资源，为文科研究注入数字化活力。高质量的语料库也是大模型等新一代 AI 技术的重要基础，希望文科的深厚积淀未来能够为科技研发提供核心价值。语料库作为平台的重要组成部分，也将为语言学、文学、历史学等多学科研究提供海量、精准的数据支撑，帮助老师们挖掘文本背后的深层规律，推动文科研究方法的创新与升级。希望今天的沙龙能为大家搭建交流的桥梁，碰撞出更多智慧火花。

主题报告

詹卫东
（北京大学中文系教授、北京大学中国语言学研究中心副主任）

大家好，非常感谢郭老师组织这次活动，也感谢计算中心马老师、范老师。我们之前就得到计算中心很多的支持，我们中国语言学研究中心（CCL）网站服务器都是放在计算中心托管的。

郭老师刚才介绍了一些背景情况，就是社科部在推动北大计算资源为文科的数据建设服务，我觉得这个想法非常好，我们也期待未来能得到计算中心更多的支持。我今天介绍的 CCL 语料库的工作，并不是我个人研究工作的主项，而是为了开展语言学的专题研究，首先需要建设一些基础资源，而建设基础资

源，又主要是工程技术性质的，所以平时不会把它当论文来写，来做比较系统的整理。为了准备今天的交流，我把一些零散的材料，包括已经发表文章中的数据与观点，收集到一起，重新进行了整理，希望能够描述出一个整体概况，这样也便于马老师、计算中心的老师们了解我们从语言学专业角度所做的一些工作，包括对于语言学资源建设的具体需求。这样在社科部的关心下，我们就更有希望从计算中心得到更多的技术支持。

先用图 7-1 来介绍北大的语言资源的总体情况，主要是北大中文系以及计算语言所相关的语言资源建设。

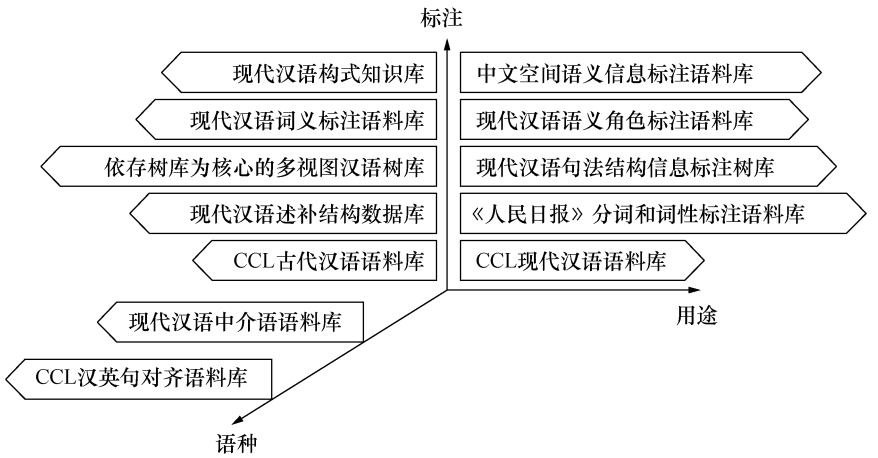


图 7-1　北京大学语言资源总体情况

图 7-1 从语言资源的用途、语料标注的深度以及涉及的语种三个维度给不同类型的语言学资源做了一个比较粗的定位。“标注”箭头方向代表标注信息增加，深度加深；“语种”箭头方向代表从单语向双语发展；“用途”箭头方向的含义相对含糊一些。这些语料库资源主要用途都是支持语言学研究、语言教学、自然语言处理应用。一般来说，标注信息越多的语料资源，用途范围也越广。

图中 CCL 语料库（现汉、古汉、汉英句对齐语料）、现代汉语句法结构信息标注树库、现代汉语述补结构数据库、现代汉语构式知识库，都是我主持建设的，中文空间语义信息标注语料库是我近两年在进行中的一个课题支持的资源构建工作，是为机器语言能力评测服务的。稍后我会展开介绍这些资源的情

况。另外几个语言资源是计算语言所的老师们主持的工作，包括俞士汶老师、段慧明老师主持的“《人民日报》分词和词性标注语料库”，王厚峰老师的“依存树库为核心的多视图汉语树库”，穗志方老师的“现代汉语语义角色标注语料库”，吴云芳老师的“现代汉语词义标注语料库”。除图上显示的外，还有刘扬老师的“中文概念词典（CCD）”，中文系做的“上古音查询系统”，袁毓林老师做的“名名组合释义系统，词类隶属度查询系统，同义名词辨析系统，动词词义蕴含系统”等多种类型的语言学知识库。限于时间，今天不展开介绍。

CCL 语料库检索系统

CCL 语料库大约于 2003 年上线试运行，2004 年推出第一个正式的稳定的版本。2014 年更新了一个版本，2024 年又做了一次比较大的更新，大概是 10 年更新一个版本。CCL 语料库在海内外有很多用户，在学术界影响比较大。这个语料库系统从问世到发展至今，都完全是“业余”作品，是我在教学和科研工作之余，跟学生一起建设的，开始并没有计划和规划。就我所知，CCL 语料库应该是最早在网上公开可用的，当时是最大规模（超过一亿字规模）的中文语料库。可能是因为占得先机，积累了一大批早期的忠实用户，所以在学术界很有影响，但是也因为年头久远，缺乏规划，现在的发展其实比它应该达到的高度要低。

北大中国语言学研究中心成立于 2000 年。在确定中心的使命时，就有一条，要建设国际一流的语言学研究资料库。我当时在中心的工作任务就是在这个方向上努力。差不多从考虑语料库系统的样态到系统实现，用了一两年时间，都是利用的业余时间。当时有两个计算语言所的硕士研究生靳志辉和谌贻荣参与了这个工作，具体编程由谌贻荣完成。20 年前正是像 Google 这样的网页搜索引擎技术最火的时候。当时有一个非常好的开源的全文搜索程序 Lucene。我偶然在网上看到有一个叫车东的程序员，在网上发布了一个 WebLucene 系统，就是把 Lucene 的搜索内核包装了一下，从索引到检索，到输出查询结果，外部数据接口全部用 XML 格式的文件，统一用 XML 格式接口，非常方便调用。这样就有条件快速搭建网上语料库了。这在当年来说，是构建语

料库比较新的想法。中文世界除了我国台湾地区“中研院”有一个现代汉语平衡语料库，大约500万字规模，可以在网上查询（但是并不免费开放），其他还没有开放的、可以自由使用的在线语料库（Online corpus）。但在学术界，有学者个人收集了一定规模的语料文件，也有一些在个人电脑上可以检索的语料库检索软件。比如中文系郭锐老师就长期收集语料电子文件，积累了几千万字的现代汉语和古代汉语语料。另外清华大学孙茂松教授、北京语言大学宋柔教授当时都有可在个人电脑上运行的语料库检索工具。我参加了孙茂松老师的一个课题，其中可用的语料规模就超过了一亿字，有很多报刊语料的电子版文件。我们就把这些能找到的语料电子文件汇总到一起，做了简单的去重，尽量统一了篇章格式，形成了CCL语料库最初的基础语料。

图7-2是CCL语料库检索系统的结构流程图。分为创建索引（离线）和查询语料（在线）两个部分。索引和检索的核心都是Lucene。我们要做的就是将语料原始文件转为XML格式文件，提供给Lucene创建索引；然后再查询语料模块，定义用户查询表达式的规范，并将查询表达式解析为Lucene可以处理的最小检索表达式的组合形式，由Lucene返回文档检索结果，在浏览器网页上显示。

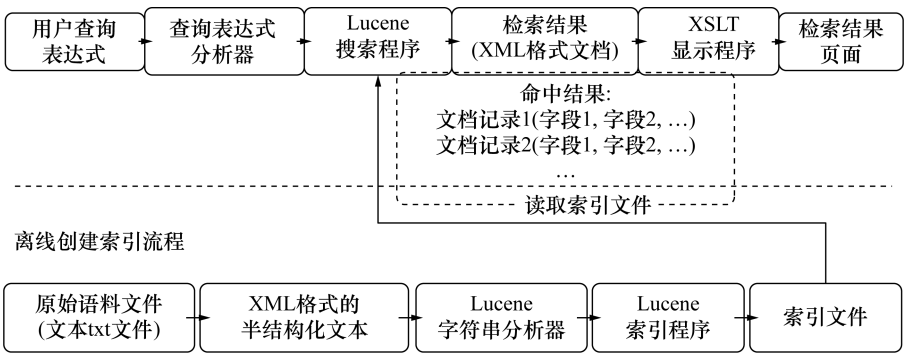


图7-2　CCL语料库检索系统结构流程图

CCL检索系统的架构一直没有变过。版本升级主要是扩充“查询表达式”

的定义范围[①]，满足用户更多样化的查询功能需求，以及增加语料文件的类型。表 7-1 展示了从 2004 年到 2024 年这 20 年间三个语料库版本的语料规模概貌。

表 7-1　三个语料库版本语料规模概貌

版本	汉语语料		汉英双语句对齐语料
	现代汉语	古代汉语	
2004 版	85,398,433	22,392,747	无
	107,791,180		
2014 版	581,794,456	201,668,719	中文字数：6,176,546 英文词数：3,934,609
	783,463,175		
2024 版	4,746,907,429	1,094,768,777	中文字数：192,057,581 英文词数：103,578,166
	5,841,676,206		

2005 年底我们统计当年的查询记录，全年服务器提供查询服务 351 天，查询次数共计 548,538 次，月均查询：548,538/12＝45,712，日均查询：548,538/351＝1,563。用户 IP 地址共计 18,020 个，在前 32 个访问次数最多的 IP 地址中，来自高校的 IP 地址占 26 个（81.25%）。6 月和 12 月是查询高峰，估计是为了完成课程论文和学位论文，学生在这些时段集中使用语料库的情况较为突出。值得一提的是，访问次数排在第 11 位的是韩国的 IP 地址（3,356 次查询）。当时 CCL 语料库没有任何宣传，影响很快就扩散到海外了。

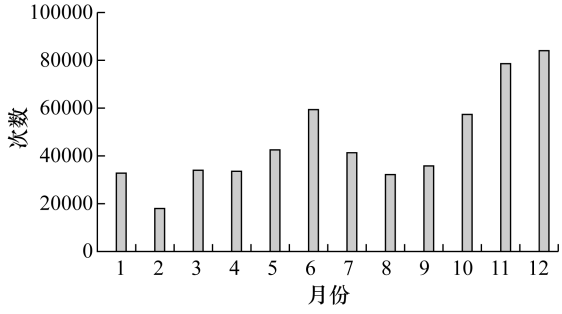


图 7-3　2005 年 CCL 语料库每月查询次数统计

① 查询表达式定义见：http://ccl.pku.edu.cn:8080/ccl_corpus/CCLCorpus_Readme.html.

表 7-2 是 2005 年查询表达式频次前 25 的记录。可以看出，用户对汉语语法功能词和虚词的查询需求非常突出。

表 7-2　2005 年查询表达式频次前 25 的记录

查询式	范围	查询次数	查询式	范围	查询次数
把	xiandai	1280	过	xiandai	555
很	xiandai	1088	起来	xiandai	555
有	xiandai	1028	在	xiandai	544
给	xiandai	931	了	gudai	516
被	xiandai	858	手	xiandai	502
下	xiandai	794	于	gudai	493
的	gudai	784	以	xiandai	400
上	xiandai	753	从	gudai	395
都	xiandai	728	每隔	xiandai	393
得	gudai	716	是	gudai	392
不	xiandai	707	就	xiandai	387
头	xiandai	582	首	xiandai	386
法	xiandai	571			

表 7-3 展示了查询离合词“澡……洗……”用法的示例。

表 7-3　查询离合词“澡……洗……”用法示例

序号	查询表达式	查询意图
1	澡 $4 洗	“澡”出现在“洗”的左侧（即句子中先说了“澡”，后面再说“洗”），目的是查找句子中“洗澡”这个词的离合用法。例如：他连澡都没洗，……
2	洗-5 澡	“洗”和“澡”之间间隔 0 到 5 个字。目的是把“洗……澡”这类离合用法，以及直接连用的“洗澡”非离合用法从查询表达式 1 的结果中剔除出去。
3	澡-0，	把“澡”后紧挨着逗号“，”的例句，再从查询表达式 2 的检索结果中剔除出去。
4	澡-0、	把“澡”后紧挨着顿号“、”的例句，再从查询表达式 3 的检索结果中剔除出去。

从上面这个简单示例可以看到，语料库查询的需求跟一般的信息检索（网页检索）很不一样，返回结果不是整篇文档，而是句子或一个自然段。用户查询表达式比较常见的查询对象是虚词（如“的、了”）甚至标点符号（如逗号、顿号等），这些在一般信息检索中通常是拒绝查询的停用词（stop words）。查询表达式除一般的关键字（连续字符串匹配）外，还要考虑多个关键字之间的间隔距离，位置关系上的顺序和逆序，以及常项关键字和变项（通配）关键字的查询模式等诸多查询需求。比如要查“爱吃不吃、爱理不理、爱看不看、爱去不去……”这样的语例，查询表达式中就要涉及“爱、不”两个常项（固定关键字），还要涉及“吃、理、看、去……”等变化的单音节动词。CCL 语料库提供的模式查询功能可以满足这方面的查询需求，更多的细节可以参看网页上的使用说明文档。

不过，CCL 语料库的句子语料是未经标注的生语料，能够支持线性字符串序列的模式匹配查询，但无法提供语法结构的层次查询。要更精准地查询不同句型不同结构的语例，就需要对原始语料进行更深层次的标注，即信息经过预加工后，才能提供更强的功能支持。下面就介绍这方面的代表性语料库：句法结构信息标注树库（Treebank）。

现代汉语句法结构信息标注树库

树库语料是把句子标注为它的句法结构树，就是把一个线性字符串形式的句子，变成括号标记的层次树结构。我们的树库大致是 100 多万字的规模，是参照国际上比较有名的宾州树库来建设的，美国宾州大学发布的中文树库语料规模（Chinese Treebank 6.0[①]）就是 100 万字级别的。图 7-4 是树库标注示例。

① https://catalog.ldc.upenn.edu/LDC2007T36.

原始句子形式：妇女经济地位的提高，是实现男女平等最重要的基础。

树库标注形式为：（zj（！dj（np（np（np（！n（妇女））！np（np（！n（经济））！np（！n（地位））））udel（的）！vp（！v（提高））））wco（，）！vp（！vp（！v（是））np（vp（！vp（！v（实现））dj（np（！n（男女））！ap（！a（平等））））！np（ap（dp（！d（最））！ap（！a（重要））））udel（的）！np（！n（基础））））））wfs（。））））

图 7-4　树库标注形式

层级树形图形式为：

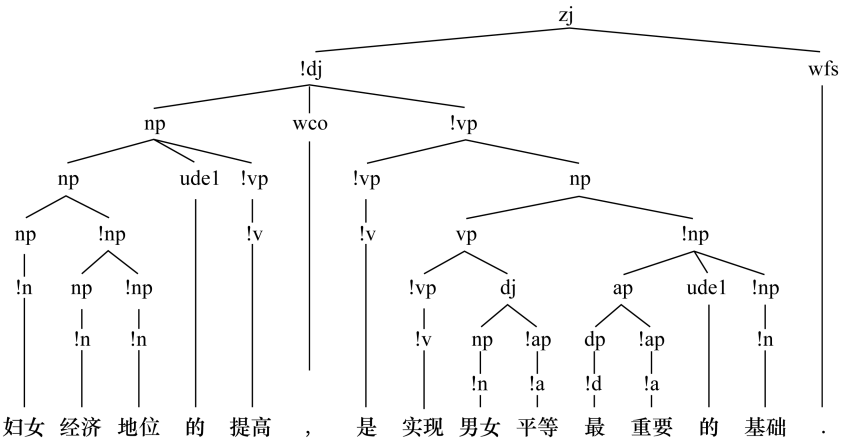


图 7-5　层级树形图形式

树库标注流程如图 7-6 所示：

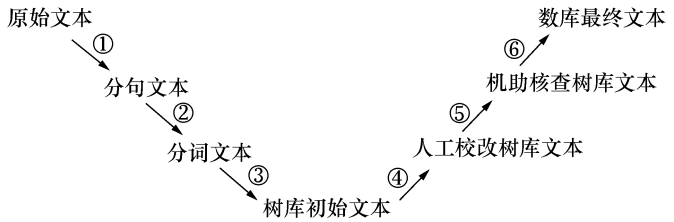


图 7-6　树库标注流程

步骤①：将原始语料文件处理为一句一行的文本文件，由断句程序自动完成，辅以少量人工校正。

步骤②：将每个文本行的句子切分为按词分隔的序列，每个词附加词性标

记，由中文分词和词性标注程序自动完成，辅以少量人工校正。

步骤③：用括号标记每个句子的结构层次，每个括号标记一个语法单位，该语法单位要标记短语类别和中心词，由一个基于规则的中文句法分析程序自动完成。

步骤④：在 TreeEditor 树图可视化编辑环境下，由人工逐句对自动产生的句法分析结果进行校改，使得每句的句法分析符合树库加工规范中对汉语句法结构标注的要求。

步骤⑤：由 TreeEditor 中的句法结构知识提取工具和树结构合法性检查工具对人工逐句校改树库文本进行处理，帮助发现各个短语结构范畴内部可能存在的标注不一致问题，提示出来供人工再次进行甄别校改。

步骤⑥：由 TreeCheck 工具进行一致性检查，确认没有违反标注规范的错误，得到最终的树库文本。北大中文树库到目前为止标注的语料规模为 55,742 句，总词数 899,365 词，总字数 1,318,488 字。语料类型包括语文课本（56.85%）、句型例句（13.29%）、新闻语料（13.00%）、科技语料（9.43%）、政府白皮书（7.43%）。各类型语料平均句长（以词数计）的分布如图 7-7 所示：

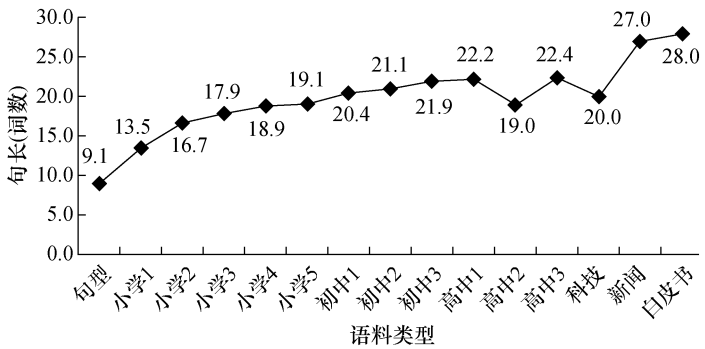


图 7-7　各类型语料平均字长分布

有了树库标注语料，对句子的构成情况就可以用结构化方式进行查询。比如“把”字结构中动词短语的结构类型，在树库中就可以直接查询到。表 7-4 展示的查询结果包括 vp 的内部构造规则，具体的句子或词组片段为：

表 7-4 “把”字结构内部构造规则

构造类型	频次	比例	结构规则	示例
述宾式 vp	999	39.74%	vp→！ vp np vp→！ vp sp vp→！ vp qp vp→！ vp mp ……	（把 x）交给 新干部 放在 桌子上 放在 第一位 砍去 一半 …
述补式 vp	907	36.08%	vp→！ v v vp→！ v a vp→！ v ude3 ap ….	（把 x）扔 掉 清理 干净 布置 得 非常漂亮 …
状中式 vp	354	14.08%	vp→dp ！ vp vp→pp ！ vp vp→ap ！ vp …	（把 x）也抛出来 在几个工作人员中分配一下 直接 倒到喉咙里去 …
附加式 vp	187	7.44%	vp→！ vp ule vp→！ v uzhe …	（把 x）摔坏 了 珍藏着
连谓式 vp	58	2.31%	vp→！ vp vp vp→！ vp wco vp …	（把 x）带回家放好 变成电信号，再加以放大 …
其他	9	0.36%	vp→！ v vp→c ！ vp …	（把 x）公开 剥一 剥 …
合计	2514	100%	48 种	

一般语法教科书在描述“把”字句语法特点时，往往强调“把”字句的语义是处置，但具体在语法形式上通过哪些动词性结构类型来表达处置义，通常描写不太充分。如果没有像树库这样的深度标注语料，只是在生语料库（像 CCL 语料库）中就很难直接查到。而从树库语料中查到如表 7-4 所示的结果，就能更细致、更清楚地呈现“把”字句的面貌：“把”后动词性短语（vp）的核心往往是述补结构，如“扔掉、清理干净、放好、摔坏、变成、抛出来”等等，但实际用例中述补结构又可能再跟其他成分组合成更大的动词性词组，比如频次排在第一位的述宾结构，如“交给＋新干部、放在＋桌子上”，还有如状中结构、连谓结构等等，可以看到“把”后动词性短语实际用法的类型丰富

性。此外，语法书上一般都提示说“把”后的动词性成分要求是复杂形式，排斥简单形式。但实际语料中也有少量的单个动词用在“把”后的情况，如“把＊公开”，类似的还有“分解、平分、消灭、相加、还原、神化、抽象化、形式化、置之度外”等等，这对一般关于“把”字句特点的认识是一个有益的补充。

除提供更精细化的查询功能外，树库语料还可以帮助从宏观视角来审视一个语言的结构特点，特别是句法结构歧义的面貌。通常语言学家在观察和分析结构歧义现象时，一般是从具体语例出发，像朱德熙先生举过的经典例句“咬死了猎人的狗”，就是有代表性的结构层次歧义，既可以理解为“猎人的狗被咬死了”（结构组合是：咬死了＋猎人的狗），又可以理解为“狗把猎人咬死了”（结构组合是：咬死了猎人＋的＋狗）。还可以类推出更多这样的例子，比如“发现敌人的哨兵、喜欢跳舞的女孩”等等。那么，问题来了：这样的歧义结构到底有多少？歧义实例在语言中的实际分布情况如何？歧义程度有无高低之分？要宏观地回答这些问题，就得借助像树库这样的深加工语料库资源。我们可以把树库中的同形词性序列（比如“n＋v”“n＋v＋n”“v＋n＋的＋n”……）全部列举出来，然后看这些词性序列构成的树结构之间是什么关系。如果存在层次或者节点标记的差异，就有潜在歧义，如果有语句实例在树库中标注为不同的树结构，就有真实发生的歧义。而歧义程度的大小，可以借助信息熵的公式来计算。表 7-5 展示了根据北大树库资源计算得到的熵值最高的前 6 个有潜在歧义的词性标记序列。

表 7-5　熵值最高的前 6 个有潜在歧义的词性标记序列

词类组合	根数	根节点	频次	合计	熵值	示例
n v	3	dj	1838	2777	1.26	前人 开路
		vp	514			全线 崩溃
		np	425			燃料 供应
n v n	3	np	685	1176	1.04	电子 发射装置
		dj	482			麦子 需要 春雨
		vp	9			重金 奖励发明人

（续表）

词类组合	根数	根节点	频次	合计	熵值	示例
v n undel n	2	vp	590	1161	1.00	解释 工厂的困难
		np	571			划分 句型的标准
v n n	2	vp	1291	1622	0.74	解决 技术 问题
		np	331			有 问题 农药
v n	2	vp	14005	16941	0.67	符合 国情
		np	2936			炼钢 工人
a v	2	vp	1173	1353	0.67	努力学习
		ap	149			胖 起来

表中对应“咬死猎人的狗”这样的歧义序列（模式）排在第 3 位，非常能说明问题，语言学家的语感直觉跟语料库的计算结果非常吻合，这种歧义是汉语中系统性存在的非常突出的歧义结构模式。表中第 4 位的“v＋n＋n”序列对应的实例“有问题农药”，就是一个真实歧义的例子（既可以理解为“有＋问题农药”，也可以理解为“有问题＋农药”）。

树库是对汉语语法结构的全面标注。下面要介绍的三种资源侧重有所不同，更突出专题性，分别是关于两种语法对象汉语“述补结构”和“构式”的知识库资源，以及针对文本中空间信息的语义理解进行专门标注的专题语料库资源。因为时间关系，此处仅做概要介绍，不详细展开。

现代汉语述补结构数据库

述补结构数据库是我们与早稻田大学砂岡和子（Sunaoka Kazuko）老师合作开发的。目的是用于帮助把汉语作为第二语言的学习者。数据库的建设目标是尽量多地收集述补结构实例，对述语和补语的基本信息（如词性、读音、义项等）进行描写，同时提供述补结构的释义、语义角色分析、例句（有中、日、英三种语言对译）等。图 7-8 展示了“摆”跟“满”组合为“摆满”述结式这个条目的信息。

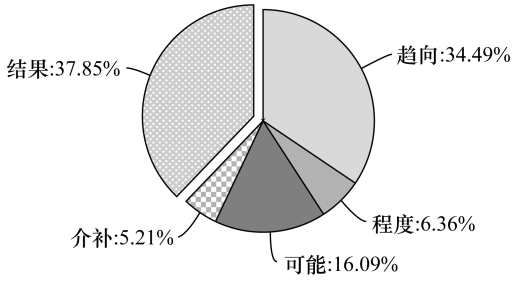
摆1 【bai3】	词性：动词	HSK：甲	该述语有 4 个义项，点击下方链接查看各义项 摆1 摆2 摆3 摆4	
摆满	bai3 man3	补语类型：结果补语	CCL频次：687	来源：砂冈

	中文	日文	英文
释义	安放的对象布满某个空间或范围。		
例句	1. 架子上摆满了各种各样的容器。 2. 书架摆满了书。 3. 商店里摆满了包装精美的商品。	1. 棚はさまざまな容器でいっぱいだ。 2. 本棚に本がびっしりと並んでいる。 3. お店には綺麗にラッピングされた商品がいっぱい置いてある。	1. The shelves are packed full of different kinds of containers. 2. The bookshelves are packed full of books. 3. a)The shop is packed full of products with exquisite packaging. 4. b)Products with exquisite packaging fill the shop.

编辑词条　删除词条　新增词条

图 7-8　“摆满”述结式条目信息

这个数据库项目完成时收集到 21,082 个述补结构实例，包括结果补语、趋向补语、可能补语、程度补语、介词补语等 5 种类型，图 7-9 展示了不同类型的述补结构在数据库中的记录条数和占比情况。



补语类型	数目	比例
结果补语	7980	37.85%
趋向补语	7272	34.49%
可能补语	3392	16.09%
程度补语	1340	6.36%
介词补语	1098	5.21%
总计	21082	100%

图 7-9　不同类型的补语数量统计

对于每个述补结构记录，我们查询 CCL 语料库，获得其全部用例，并在数据库中记录了频次信息，根据频次大小，在图 7-10 所示的卫星图上，以述语（动词）为中心，将补语（形容词或动词）放置在周围位置上，离述语近的表示频次高，远的表示频次低。比如“放—走、放—大、放—好、放—平……”等就是相对高频的述补结构，“放—黄、放—明白、放—光……”等就

是相对低频的述补结构。点击图上每个补语节点，会弹出文本框，展示该述补结构的基本信息、例句、语义角色等内容。

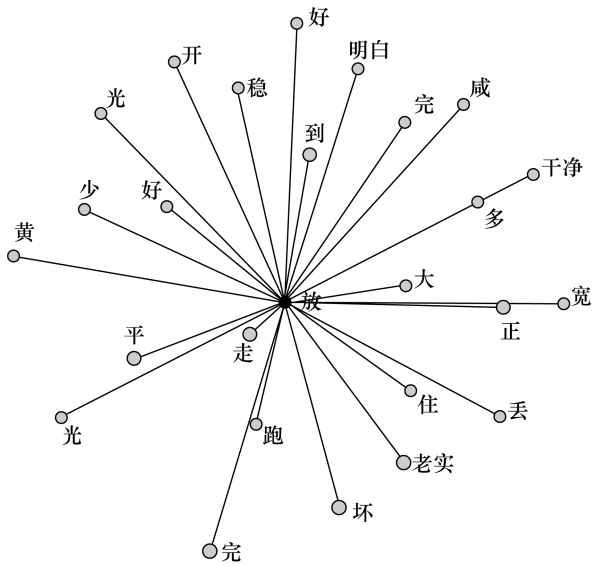


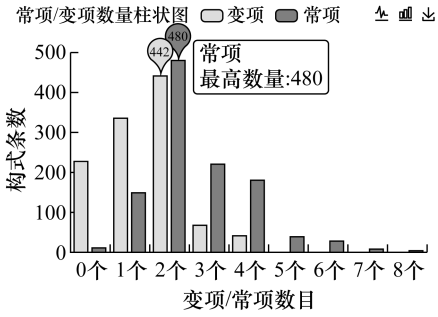
图 7-10　频次信息卫星图

述补结构是汉语中相对较为特殊的句法结构类型，在对外汉语教学中是一个难点。通过构建述补结构数据库这样的资源，并提供可视化展示信息的网页方式，可以为对外汉语教学提供一个方便实用的途径，提升教学效率。

现代汉语构式知识库

构式（Construction）指的是语言中形式和意义的对应关系比较特殊的语法单位，它的整体语义不能由构成成分的意义简单组合得到。比如“你罚你的款，他违他的章”不仅仅是表达“他违章，你罚款”的字面意义，还有“他不理会罚款，继续违章（我行我素）”这样的字面外的意思。而像这样的表达形式，就对应着一个构式，可以记作：r＋v＋r＋的＋n，其中 r 代表人称代词（如“你、他”），v 代表动词，n 代表名词。我们在数据库里用这个词性标记的序列来代表一个构式，对这个构式的语义和用法特点进行描述。我们收集了 1000 多条类似这样的构式，形成了 CCL 构式知识库，可以在 CCL 网站上

检索[①]。每条构式都由常项（具体词语）和/或变项（语法标记）组成。图 7-11 展示了常项和变项数量分布的情况。



常项统计

常项数目	0	1	2	3	4	5	6	7	8	总计
构式条数	10	149	480	220	180	38	27	7	3	1114
构式占比	0.9%	13.4%	43.1%	19.7%	16.2%	3.4%	2.4%	0.6%	0.3%	100%

变项统计

变项数目	0	1	2	3	4	5	6	总计
构式条数	228	335	442	67	40	1	1	1114
构式占比	20.5%	30.1%	39.7%	6.0%	3.6%	0.1%	0.1%	100%

图 7-11　构式中常项和变项数量分布情况

图 7-12 是“差点儿＋没＋vp”构式的信息描述，包括该构式的变体形式、构式类型、构式特征、构式释义、构式举例等。这个构式的特征标记了“冗余｜含否定成分”，比如“差点儿没割破手”是这个构式的一个实例，释义等价于“差点儿割破手”，表示这里面的“没”是个冗余成分。整个表达形式虽然包含了两个否定成分“差点儿”和“没”，但语义跟一般的双重否定等于肯定的释义方式不同，并不表达肯定义，也就是说，不是“割破手”的意思，而是“没割破手”的意思。这个构式对其中变项 vp 有两个语义特征上的要求：通常来说，vp 所代表的动作、行为，首先在客观上比较容易发生；其次在主观上，也就是说话人的认识或者社会常识意义上大多数的认识当中，是不期望发生的。比如“割破手”，在切菜或者类似的生活工作场景中，一般人就感觉比较容易出现这种情况，这是一个特点；同时这又是当事人不期望发生的事情，这是第二个特点。在这两个条件作用下，用“差点儿没＋vp”表达的意思，就跟“差点儿＋vp”表达的意思相同，都是庆幸一个不想发生又容易发生的事情没有发生，表达了“还好、还好没……”的庆幸的心态。

构式表达往往跟主观态度、情感语义有联系。在构式知识库基础上，我们选取了 100 条构式，从 CCL 语料库中查找这些构式的实际用例（总数约 1000

① https：//ccl. pku. edu. cn/ccgd.

图 7-12　“差点儿＋没＋vp”构式信息描述

条），对它们所表达的情感义进行了标注。比如像前边提到的“爱吃不吃、爱理不理、爱看不看”这类“爱＋v＋不＋v”构式，往往就是表达消极情感，持不满或不置可否的偏负面态度，情感强度介于极高和一般之间（如图 7-13 所示）。关于构式语料的标注工作，这里就不进一步展开了。

图 7-13　“爱＋v＋不＋v”构式情感强度

中文空间语义信息标注语料库

从 2020 年开始，我参与了计算语言所穗志方老师的科技部课题[①]，研究任务是设计能有效评测机器语言能力的数据集。当时就提出了一个思路，以中文文本中的空间语义信息理解为主题，从表层语义到深层认知语义，逐级加大理解难度，来测试机器的理解能力。按照这个思路，我们先是从 CCL 语料库以及互联网上收集了大量富含空间信息表达的文本，形成一个语料池（共计 89,740,377 字），内容分为两大类：一类是一般性的语料，一类是空间相关的专题性语料。前者具体包括报刊语料（36%），文学作品语料（25%），中小学语文课本语料（20%），跟空间表达研究相关的语言学论文语料（不到 1%），语义角色标注课题中涉及空间角色的语料（不到 1%），少量自拟语料（不到 1%）等；后者包括交通事故判决书语料（9%），体育动作语料（6%），地理百科语料（2%）。然后基于一个空间词表，对语料中的空间方位词进行大量的替换，形成对照句对，标注替换后的文本空间语义信息是正常还是异常（二分类任务）。比如图 7-14 展示的就是去年标注的一个例子。原文的“放进嘴里”的“进”被替换成“出”，形成的“放出嘴里”就是一个“搭配不当”的错误表达形式。这条语料最终会进入测试集中，由计算机程序来判断这段文本是否包含错误的空间信息，并定位到具体哪个语言片段是错误的。

除判断正误任务外，我们还设计了深层理解类的任务，对文本中的空间语义信息进行全面的角色标注。图 7-15 展示了一个空间语义角色标注的例子。这是一个辅助标注的网站，标注员可以通过鼠标操作，选中文本中的片段，标注其空间信息，比如“戒指”这个实体的空间信息以“处所”角色标注，取值为“美国人手上”，相关的事件动词是“戴”。而从动词的视角来看，就是标注动词的论元角色，比如“戴”，就标注了 3 个论元，arg0 是“美国人”，arg1 是“戒指”，argS（表示处所）是“手上”。

---

① “国家科技创新 2030‘新一代人工智能’重大项目——以自然语言为核心的语义理解理论、模型与方法”（项目号：2020AAA0106701）。

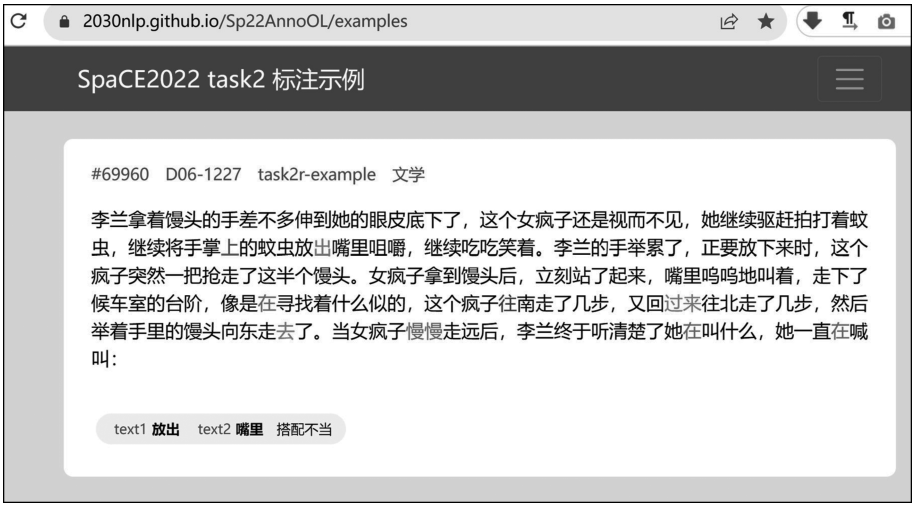


图 7-14　空间信息表达标注示例

图 7-15　空间语义角色标注示例

通过这样的标注，就把自然文本中的空间表达转化为结构化信息，可以再进一步制作为测试题。以上标注的语料经过一定的筛选条件，进入最终的数据集（划分为训练、验证、测试 3 个集合），用于测试机器对中文文本中空间信息的理解水平。我们按照这个工作模式，已经连续组织了 3 届空间语言理解评

测大赛[①]（SpaCE2021—2023）。将来计划把历届大赛的数据集进行整理，形成一个可供语言学研究和自然语言处理研究的空间语义信息标注专题语料库。2022 年 11 月底，ChatGPT 问世后，大模型一下子引起了全世界的关注，大模型的语言能力进步速度非常快，原来评测任务设计出来可能有半年一年的生命力，但大模型出来后，评测数据集就越来越“短命”了。这反过来促使我们要去考虑怎么来标注、构建更有新意的语言资源，来有效支持评测数据集的研制。

结语

最后做个简短的总结。我从博士毕业留校参与中文系和中国语言学研究中心的语言研究和教学工作，到现在二十多年，一直在做语言资源建设的工作。这对学术研究来说，是一个基础工作，但可能又不太会直接体现在专题的论文中，可见度不高。因为具体工作是大量的工程技术性质的一些比较琐碎的事情。但实际上我和学生们“业余”做的事情，要求并不低，就是要有（1）理论研究支撑；（2）应用研究驱动；（3）工程技术保障。所以特别希望今后能得到计算中心的更多的专业支持。具体的需求分为四部分：核心检索功能，网页前端技术，服务器架构与安全，数据分析与可视化呈现（见图 7-16）。

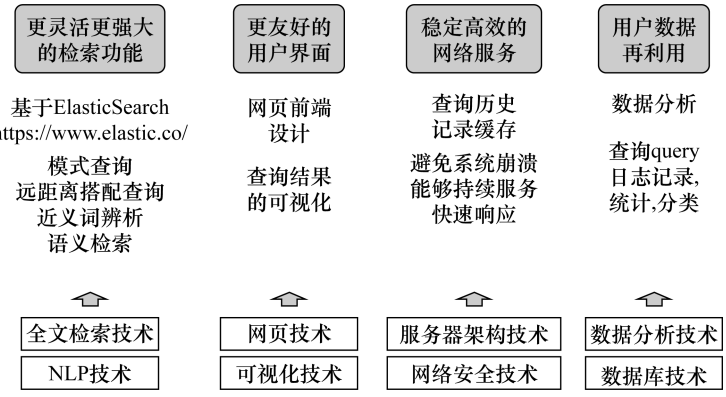


图 7-16　语言资源建设具体需求

① https://2030nlp.github.io/SpaCE2023/.

最后，要感谢许许多多为 CCL 语言资源的建设作出贡献、提供过支持和帮助的老师和同学，以及广大的 CCL 语言资源用户。谢谢大家！

◆ **郭琳**（主持人，北京大学社会科学部副部长）

特别感谢詹老师，起初我以为他只是介绍一下 CCL 语料库，没想到这是一个打包的整体的语言资源的报告。他讲得特别好，理论层面语言学本身的研究，作为资源建设的支撑，特别重要；然后，应用研究就是真正我们要干什么，为谁服务；工程技术保障也正是我们今天特别要讨论的，未来可以合作的，计算中心怎么为北大社科人文的研究提供支持，特别值得讨论。最后，我觉得还可以增加一个扩展性比较强的内容，就是在资源建设方面，大模型能不能发挥作用？另外我觉得 CCL 语料库未来对于我们研究大模型也会有特别好的帮助。

CCL 语料库未来是不是可以开发出一些产品。比如说无论是教外国人学中文，还是帮助中小学生提高中文水平，同时包括新闻记者和政府公文等规范表达，语料库能不能有帮助？

北大外院几位老师在外语语料库的建设方面也做了很多工作，可能跟 CCL 语料库有相通的地方，所以下面请各位老师自由发表意见。可以是对詹老师工作的一些评论或者意见建议，还有包括您自己做的语料库方面的相关研究，包括对计算中心有什么需求等，无论哪个方面，都可以自由交流。

◆ **马皓**（北京大学计算中心主任）

CCL 语料库的查询语句是通用的吗？比如刚才查询“洗”和“澡”的离合词的例子，一步一步地去逼近目标，里边用了很多符号。类似这样的查询，别的语料库是什么样的形式？比如您提到的我国台湾地区的那个语料库，跟

CCL 语料库的查询形式一样吗？

◆ **詹卫东**

CCL 语料库的查询语句是自己专门设计定义的。目前语料库的查询表达式一般都是开发者自己定义，没有特别通用的大家都接受的表达形式。不太像数据库查询，大家都统一用 SQL 的语法。也有的语料库检索系统把自己的查询表达式规范称为语料库查询语言（Corpus Query Language，CQL）。不过 CQL 并不是大家共同遵循的规范。因为语料库的使用者多数都不具有很多技术背景。我们的语料库在提供由专门的符号组合拼接成的查询表达式的同时，也提供用勾选、表单等方式来表达查询需求的高级查询页面。像我国台湾地区“中研院”的现代汉语平衡语料库的查询页面也有这样的设计。应该说，不同的语料库系统提供检索服务的“思路”都是相通的，但语法形式目前还是各有特点，不完全一样。因此，查询功能也有差异。CCL 语料库的查询功能覆盖面比较广，除一般的关键字查询外，还允许多个关键字间隔的查询，两个关键字既可以顺序也可以逆序，还提供模式匹配查询，可以对关键字前后的变项进行约束，比如要求是名词或者动词等特定的词类，变项的长度比如是双音节词还是三音节词等等。

◆ **郭琳**（主持人，北京大学社会科学部副部长）

CCL 用户如何标注引用？研究者如果用了这个语料库，在论文标注和引用方面，有什么要求吗？北大社会科学调查中心掌握的数据非常多，服务的用户也非常多。您刚才也说到用户数据的再利用，不光是利用他们的数据，也是可以有一个持续跟踪的服务，甚至在这个基础上可以做成一个学术共同体，都是可以探索的，其实范围很大。另外调查中心还会有用户的培训，扩大用户群体。

◆ **马皓**（北京大学计算中心主任）

对，计算中心有学校高性能计算平台，支持学校的科研和教学。老师们通过这个平台计算得出结果，发表科研论文的时候，我们希望老师在致谢部分能够有一个表示，说明本工作的成果来自北京大学高性能计算平台。用户如果标注了引用，我们相应地可以奖励机时。比如发表在像 *Nature*、*Science* 这样的顶刊，可能奖励更多。因为这个平台也是要有一定付费使用的，它有一个计算设备的维护成本。这样也就形成了一个正向的循环。老师们利用这个平台做了科研，我们的平台也扩大了影响，可以吸引更多的老师使用，做成一个网上的用户交流的社区。甚至可以建一个论坛，大家分享一些代码和数据资源。

◆ **詹卫东**

谢谢郭老师和马老师的意见。我们对于论文标注和引用方面的意识其实是比较差的，没有做过主动的宣传和引导。早期是用户在论文中注明我们 CCL 语料库的网址并致谢，都是用户自发的行为。一般来说，学者们还是比较注意学术规范，用过我们的语料库都会主动致谢。所以在知网 CNKI 上用“CCL 语料库”作为关键字查询，可以看到很多引用。我 2017 年的时候查过一次，有 3700 多篇文献引用。2019 年，北京外国语大学《语料库语言学》杂志主编许家金老师联系我，说我应该写一篇介绍 CCL 语料库的文章，方便学界引用。我就写了一篇。现在 CCL 网站上也把这篇文章的信息放在了语料库使用说明文档的开头，明确提出请大家在发表论文时引用。

原来 CCL 网站上还有一个汉语语言学论坛 BBS，有一个语言学百科 CCLWiki。不过没过多长时间，就在上面发现了各种广告帖和散布非法信息的帖子。我们没有人手去做管理工作，就只好把这些培养用户社区的功能都下架了。

也曾经考虑过在像北大未名 BBS 上开设一个 CCL 语料库的版面，或者在像 GitHub、Stack Exchange 这样的平台上建设 CCL 语料库的用户社区，但都没有精力去做。CCL 语料库是没有用户系统的。互联网用户人人可用，不需要注册。我们在后台也只是记录一下查询时用户的 IP 地址、查询时间、查询

表达式。

◆ 苏琪（北京大学外国语学院长聘副教授）

很多语言学家的研究是从比较微观的点去切入的。詹老师考虑问题的特点是更全面地掌握材料，进行宏观分析。所以他从语言学研究的需求角度出发，考虑要建一个像 CCL 语料库这样的基础资源。正是因为有他的语言学专业的很多理念作背景，所以 CCL 的查询表达式、查询模式的设计，就很全面，很有特色。这是詹老师与其他许多语言学研究者不同的地方。我的课上用到 CCL 语料库还比较多，比北语的 BCC 语料库更多一些。不过 CCL 语料库的稳定性确实也是一个问题，有时候就会突然访问不了。后来有经验了，就养成了及时把查询结果截图备份的习惯。

詹老师刚才介绍的树库语料都是靠人工标注的。这个我觉得特别难。现在基本已经没有条件做这样的工程了。像俞士汶老师那一辈学者才能静下来做这么难的专业的语料标注工作。这样标注出来的珍贵的资源，利用率没有 CCL 语料库那么高，因为查询界面以及检索技术用起来还是有一些问题。

詹老师展示的这些资源还是分散的状态，最好能整合起来，使用统一的界面、统一的查询语法。这真的是非常需要技术支持。光靠詹老师的团队很难实现。中国语言学研究中心（CCL）的这些资源内部需要整合，另外现在社会科学的研究也都会用到语料，可以考虑在更大范围，跟北大外语的语料资源进行整合，跟社会科学其他院系的数据资源进行整合。学生的开发能力还是比较有限，如果能得到计算机团队的专业开发者的支持，就可能更好一些。我有一个项目本来也是学生开发的，后来王选计算机研究所的团队帮助我们重构了整个系统，效果就好很多。

◆ 马皓（北京大学计算中心主任）

国内类似 CCL 语料库的资源还有哪些？如果相互比较的话，CCL 语料库有什么特别的地方？计算中心如果要重构 CCL 语料库，大概工作量怎么估计？

◆ 詹卫东

CCL 语料库是国内最早的大规模在线语料库检索系统。因为历史长，积累了海内外的大批忠实用户。后来跟 CCL 类似的语料库主要就是北京语言大学的 BCC 语料库。BCC 语料库的规模据说是百亿字。不过 CCL 语料库的语料类型多样性、不同时代的语料的平衡性、书面语语料的规范性都是比较有特色的。苏老师说 CCL 的语言资源类型比较多，这些资源如果能更好地整合，可以发挥出更大效用。这本身也是一个特色。比如树库语料就很有特点，包含了中小学语文课本，还有专家精选的汉语句型例句。如果能够把这样的深加工语料库与 CCL 生语料库做更好的链接，除了支持做学术研究，还可以用于语言教学场景。2012 年的时候我在夏威夷大学报告过用树库语料来辅助汉语句型教学的设想[①]。当时参会的很多一线的海外汉语教师都表示很有兴趣。

其实 CCL 语料库的影响不仅仅在语言教学和学术的圈子。我这里还可以讲一个比较有意思的 CCL 语料库出圈的故事。去年我偶然看到一则社会新闻[②]，讲的是南京大牌档在一场侵权官司中胜诉。南京大牌档是原告。它举证安徽的两家公司也就是被告侵权的时候，就用了 CCL 语料库的检索结果，对比了语料库中“大排档”和“大牌档”的用法情况，说明前者才是通用词汇，而“大牌档”不是通用词汇，是受保护的商标。法院判决时采信了这条证据，判南京大牌档胜诉。这个小例子也能从一个侧面反映 CCL 语料库的权威性和影响力。

重构 CCL 语料库检索系统的工作量不太好估计。我们是从 2021 年 11 月份开始修改旧版系统的 bug，并增加一些新的查询功能，到现在差不多两年，基本完成了。没有集中时间来开发，是几个学生分工，断断续续地编程。之前的两个版本，集中开发时间每次是 6 个月左右。

---

① 詹卫东，2012，树库在汉语语法辅助教学中的应用初探，第七届国际电脑与汉语教学研讨会（TCLT7）特约工作坊报告，2012.5.25—27，University of Hawai'i at Manoa. 修改稿发表于 *Journal of Technology and Chinese Language Teaching*，2012，Vol.3，No.2，pp.16-29.

② 澎湃新闻报道：https://m.thepaper.cn/newsDetail_forward_19948903.

◆ **成沫**（北京大学外国语学院助理教授）

我来自外国语学院意大利语专业。我自己在语料库方面的经验，就是设计了一个非常小的、针对但丁诗歌韵脚的搜索小软件。我在做这个研究的过程中，感觉是类似业余半业余的工作，尤其是对于人文学者，可能的情况都是，研究引发了某种需求，然后市面上又没有现成的可以满足需求的软件，而又有一些技术手段，然后就自己动手去做。可能您的初衷也是这样的。

刚才您说到，CCL 语料库系统使用的很多场景是教学，还有学生写论文。这个是不是说明，像这样的一些系统，其目的更多的是面向一种大众的教学的，或者是一种较少创新的？

在意大利语领域的一些数据平台，真正的原创性的场景其实是比较少的。以词韵为例，更多的是已有前人的一种理解，然后通过更多的数据，比如说过去是通过 10 篇诗歌里面的一些词韵，根据它们之间的互文关系来建立一种理论。现在可以通过数字语料库的方式，用 50 个词，用更多的材料来进行验证。所以我也想请教，您自己在使用的时候，有没有一些开拓性的规则的创建？就是说在功能设计方面会不会有所偏重，能够让语料库成为一种开创性的平台？因为很多时候人们会问，我们为什么要有语料库这么一个平台？因为在进行科研的时候，可能教学是一部分或者展示是一部分，但更多的可能是会希望这些数字手段能够有一些新的贡献。这是我的一个疑惑，所以向您请教。

◆ **詹卫东**

语料库是一个辅助工具，主要是用于验证（证实或证伪）一个假设。语料库查询系统自己不会产生新的知识。我在设计 CCL 语料库的查询表达式的时候，主要是基于以往的研究经验，考虑会有哪些查询需求。而这些查询需求，就是字符串匹配，去真实语料中匹配有没有某种语法模式。比如我在上课的时候讲到汉语中有一条规则，数量词不能直接跟并列结构的名词词组组合。比如：可以说“三条鲤鱼”，也可以说“三条裤子”，但是“三条鲤鱼和裤子”组合到一起就不能表达“三条鲤鱼和三条裤子”的意思。如果在语料中有这样的表达形式，也只能理解为“三条鲤鱼”和“裤子”，“裤子”的数量是不清楚

的。我们可以现编一个这样的句子，比如“三条鲤鱼和裤子统统不见了”。这个句子的意思应该是好理解的，其中的“裤子”不可能理解为是刚好三条。那么，我提出的这条规则对不对呢？真实语料中有没有类似“三条鲤鱼和裤子”这样的用例？如果有，具体意思又是怎么理解的？这就要借助语料库查询工具，通过设计好的查询语言，来达到查询目的。所以，语料库所能起的作用，首先是验证一条规则是不是站得住，然后再进一步，通过统计和定量分析，说明一条规则在多大程度上站得住。

再有一个就是在查询结果的基础上做知识抽取，通过一些统计手段和可视化手段，比如对近义词的搭配差异进行分析。比如“涂”和“抹”、“刷”和“擦”这样的相近动词，学习者想知道它们有哪些用法上的共性和差异，语料库工具箱里就可以设计一个功能，在查询到的包含这些词语的例句基础上，对常用搭配进行自动提取，然后用可视化的方式，把对比情况展示出来。我们现在的技术还没有做到这些。未来希望计算中心的工程师团队可以帮忙，全部实现这些功能。这个可以算是语料库辅助发现知识，而不仅仅是验证一个已有的规则。

◆ 马皓（北京大学计算中心主任）

今天詹老师介绍得很全面。我们对 CCL 语料库的功能需求有了很多了解。我觉得重构 CCL 语料库的检索系统很有意义，很值得做。现在的检索技术比 20 年前有了很多进步。用 Elasticsearch 这样的框架来重构 CCL 语料库检索系统，肯定可以让检索服务响应效率更高，而且系统也会更稳定。我们还可以尝试用大模型来支持自然语言查询方式。这些都值得去试验。

◆ 詹卫东

谢谢马老师！通过今天的介绍，计算中心对我们 CCL 语料库的情况有了更具体的了解，特别是对我们的需求、哪些方面可以帮助我们，比之前更清楚了。非常感谢社科部组织这次交流活动，期待计算中心的技术力量可以帮助 CCL 语料库升级换代，为全球的用户提供更高质量的服务。

后续进展

在计算中心杨加老师团队的大力支持下，2024 年 1 月 22 日，CCL 语料库检索系统第三版正式上线运行。2025 年 10 月 11 日，我们统计了服务器记录的这段期间（628 天）的查询日志。总查询次数 12,319,326，日均查询次数 19,616，用户 IP 地址覆盖 132 个国家和地区。

2025 年，杨加老师团队基于 Elasticsearch 重构了 CCL 语料库检索引擎。CCL 语料库系统第四版已经进入内部测试阶段。

2025 年，我们整合了 SpaCE2021—2025 的评测数据集，形成了中文空间语义信息标注数据库（SpaCECorpus），已在北大中国语言学研究中心网站上线[①]。目前包含 56,197 段文本，语料规模为 6,121,850 字。

2026 年，CCL 语料库新版检索系统扩容升级后，以新的域名（https：//corpus. pku. edu. cn）上线运行。

---

① http：//ccl. pku. edu. cn：8084/SpaCECorpus 得到 2022 年度教育部人文社会科学重点研究基地重大项目“面向机器语言能力评测的综合型语言知识库研究”（项目号：22JJD740004）资助。