北京大学CCL语料库的研制*

北京大学 詹卫东 郭 锐 常宝宝 谌贻荣 陈 龙

提要：北京大学中国语言学研究中心CCL语料库是面向语言学本体研究和语言教学的大规模语料库，目前包括现代汉语、古代汉语和汉英句对齐平行语料，规模超过7亿汉字。CCL语料库检索系统以包括汉字、字母、标点等在内的字符为基本索引单位，提供普通查询、批量查询、模式查询等多种检索方式。同时该系统支持限定范围查询、基于复杂检索表达式的查询、统计模式频次、对查询结果进行排序、下载查询结果等功能。本文介绍CCL语料库的建设情况与主要功能，具体涉及语料分布概况、语料库查询功能和使用方式、语料库索引与检索技术架构等。

关键词：北大CCL语料库、汉英双语对齐语料、语料检索、语料索引

1. 引言

在语言本体研究及语言应用领域（如语言教学、教材编写、词典编纂等方面），语料库都发挥着非常重要的作用。另外，在自然语言处理领域，数据驱动的方式亦是当前主流。国际上的BNC语料库、COCA语料库、美国宾州大学LDC收集的多语种语言资源；国内的国家语委语料库、北京语言大学的BCC语料库（荀恩东等 2016），以及本文要介绍的北京大学CCL语料库等，均被广泛应用于语言学研究、教学领域以及自然语言处理中。可以说，经过半个多世纪的发展，语料库已经成为现代语言学相关领域必不可少的基础资源和研究工具[1]。

北京大学CCL语料库是由北京大学中国语言学研究中心（简称CCL）开发的大规模中文语料库。CCL成立于2000年1月。同年9月被教育部批准为全国普通高等学校人文社会科学重点研究基地。成立之初，设定的目标即为“努力把基地建设成为国际一流的汉语语言学研究中心、国际一流的汉语语言学研究信息资料库、国际一流的汉语语言学研究学术交流中心”。其中第二项目标的主要工作内容就是构建大规模语料库，免费向全球用户开放，服务汉语研究和教学[2]。根据这

* 本文写作得到2015年度教育部人文社科重点研究基地重大项目（15JJD740002）的支持，特此致谢。CCL语料库的建设工作得到了北京大学中国语言学研究中心历任领导陆俭明教授、蒋绍愚教授、王洪君教授、陈保亚教授和北京大学计算语言学研究所俞士汶教授的关心和支持。在北京大学CCL语料库构建工作过程中，先后有多位老师和同学给予大力支持，包括北京大学计算语言所段慧明、柏晓静、靳志辉等，北京大学中文系杨灵叶、沈薇、张洁、曾石铭等，中国科学院信息工程研究所王斌、李鹏等。此外，海内外广大用户也对CCL语料库提出过很多宝贵的改进意见和建议。在此一并表示诚挚的谢意。

个定位和具体的任务要求，CCL研究人员联合北京大学计算语言学研究所，研发了CCL语料库，于2004年底在CCL网站上发布了首个版本。此后分别在2006年、2009年、2014年历经多次语料扩容和检索系统功能升级，系统运行至今已有近十五年，而且仍在根据研究工作需要和用户反馈持续更新。CCL语料库在海内外汉语研究和教学领域得到了广泛应用，产生了较大的影响。本文将详细介绍CCL语料库的研制情况和使用方法。

2. CCL语料库的设计理念与语料分布

语料库语言学发展过程中，对于原始语料是否应加以标注，一直有两种对立的主张。一种观点认为语料库应该保持原样，不做标注。因为标注意味着预设的理论注入，可能带来谬误。真正的语言学知识，应该来自原始语料本身，不带任何预设的偏见。另一种观点认为语料标注有助于更好地研究语言，标注信息丰富的语料库可以在包括语言本体、语言认知等领域提供更好的工具支持。支持前一种观点的有一些著名的语料库语言学家，比如Sinclair（2004）、Teubert（2005）等。支持后一种观点的语言学者大概更多，可参见Leech（1993，1997，2005）、Gries（2012）等。这两种主张的背后，其实是对语料库在语言研究中所起作用的定位差异，即所谓语料库驱动的语言学（corpus-driven linguistics）与基于语料库的语言学（corpus-based linguistics）之分。前者把语料本身作为语言学理论的数据来源，追求在纯粹的原始语言数据基础上，构建全新的、区别于传统的、真正意义上的语料库语言学理论；后者把语料库作为工具看待，主张在标注语料基础上，检验并发展已有的语言学理论。

在北大CCL语料库系统设计之初，研究人员一方面受到上述语料库语言学“两种路线之争”的大背景影响；另一方面，也更主要的是，中文文本语料的自动分词、词性标注技术在当时的技术条件下还不够成熟，如果要进行词汇级的标注处理，需要较多的人工干预，成本较高，因此选择了基本保持自然文本状态，仅做文本的篇章分类和少量信息标注的路线。把语料库建设的工作集中在两个方面：一是完成基本的电子文本的文字校对；二是开发功能丰富的、支持语言学研究的例句检索系统。这样可以大大缩短语料库的开发周期。

根据当时的技术条件和已有的电子文本积累，CCL语料库确定入库的语料类型包括三大类：现代汉语语料、古代汉语语料和汉英句对齐语料。从规模上讲，以现代汉语和古代汉语语料为主。主要为汉语本体研究提供服务，汉英句对齐语料可以为汉英对比研究提供支持。下面分别介绍三类语料的小类构成以及所占比例。

（一）现代汉语语料

CCL语料库中现代汉语语料近12亿字节[3]，包含10645个不同字形的汉字。其中1949年之前的语料为“现代”语料，1949年之后的为“当代”语料。现代文

献约0.15亿字节，占全部现代汉语语料的1.28%；当代文献，涵盖了口语、文学、网络语料、应用文等10类，约11.8亿字节，占现代汉语语料的98.72%。现代汉语语料的分类及规模统计详见表1。

表1　现代汉语语料规模统计

现代文献	字节数	百分比	当代文献	字节数	百分比
文学	14,052,591	92.15%	报刊	839,973,730	71.45%
戏剧	1,197,572	7.85%	翻译作品	90,046,147	7.66%
			文学	85,241,162	7.25%
			网络语料	54,680,142	4.65%
			应用文	48,286,885	4.11%
			电视电影	21,359,547	1.82%
			学术文献	20,655,712	1.76%
			史传	8,799,888	0.75%
			相声小品	3,480,086	0.30%
			口语	3,081,723	0.26%
总计	15,250,163	100%	总计	1,175,605,022	100%

（二）古代汉语语料

CCL语料库中古代汉语语料近4亿字节[4]，包含18,898个不同字形的汉字。古代汉语语料根据语料所在的朝代分类。对于一些不方便按照朝代分类的语料，CCL语料库将它们归入其他杂类。各朝代语料收录了从周代到民国的1.64亿字节的语料（占比41.05%）；杂类语料2.36亿字节（占比58.95%）。古代汉语语料的分类及规模统计详见表2。

表2　古代汉语语料规模统计

按朝代排序	字节数	百分比	杂类语料	字节数	百分比
周	292,382	0.18%	大藏经	72,455,219	30.76%
春秋	942,293	0.58%	二十五史	70,496,299	29.92%
战国	2,521,690	1.54%	历代笔记	4,666,829	19.81%
西汉	1,013,037	0.62%	十三经注疏	16,828,127	7.14%

（待续）

（续表）

按朝代排序	字节数	百分比	杂类语料	字节数	百分比
东汉	2,585,081	1.58%	全唐诗	8,789,487	3.73%
六朝	5,796,356	3.53%	诸子百家	7,001,200	2.97%
隋唐	9,004,695	5.49%	全元曲	5,751,855	2.44%
五代	1,555,366	0.95%	全宋词	3,890,311	1.65%
北宋	31,982,012	19.50%	道藏	1,847,085	0.78%
南宋	2,843,677	1.74%	辞书	1,464,667	0.62%
元	961,884	0.59%	蒙学读物	394,627	0.17%
明	21,038,301	12.83%			
清	48,109,077	29.33%			
民国	35,371,339	21.57%			
总计	164,017,190	100%	总计	235,585,706	100%

（三）汉英对齐双语语料

CCL语料库中汉英句子对齐语料约0.716亿字节，其中包含747个汉译英文件和1627个英译汉文件，约23.36万个对齐的句子对（具有翻译关系），含600多万汉字和近400万英语单词。语料以书面语为主，也包含少量口语，分为应用文、文学和新闻三类文体，涉及政治、科技、体育等多个领域。表3列出了这些对齐语料在不同文体中的统计信息。

表3　汉英句子对齐语料规模统计

文体	句对数	中文句数	英文句数	中文字数	英文词数	中文平均句长（按字数计）	英文平均句长（按词数计）
应用文	20,802	21,454	25,399	912,658	550,228	42.54	21.66
文学	192,178	215,518	238,546	4,385,701	2,871,770	20.35	12.04
新闻	20,609	22,453	23,979	878,187	512,611	39.11	21.38

CCL语料库中的中文语料和英文语料均未做词汇、句法信息标注。仅做了篇章层面的少量分类信息标注。对于中文语料，以文件夹和文件名表示领域分类、文体信息和作者信息等；对于汉英对齐语料，在原始语料的XML文件中，标注了一篇文献的领域、文体、作者、译者、原文语种等信息。这些标注信息可以在检索时由用户指定为检索条件。参见下文3.4和3.5小节的说明。

3. CCL语料库的检索功能

CCL语料库中，单语语料库支持普通查询、批量查询和模式查询三种查询方式，双语语料库在普通查询外还提供了检索界面更为友好的高级查询页面。以下结合示例详细介绍这些查询功能的具体使用方式。

3.1　普通查询

普通查询功能通过查询表达式使用。查询表达式由关键字、数字、分隔符、操作符、基本项、过滤项、简单项、复杂项、子句等8项组成。这些项目的具体层级关系见表4。

表4　查询表达式的构成形式

项目	构成形式				
查询表达式	由一个或多个子句构成，多个子句之间以空格分隔。				
子句	由复杂项或过滤项构成。				
复杂项	由一个简单项直接构成，或由简单项和操作符及数字构成。				
简单项	由一个基本项直接构成，或由Operator1连接基本项构成。				
过滤项	由关键字、分隔符和简单项组成。				
基本项	不含操作符的任意字符串。				
操作符	Operator1	空格 |			
	Operator2	$　+　#　-　~			
	Operator3	!			
分隔符	:				
关键字	author	pattern	name	type	（单语和双语语料通用）
	ch	en	translator	enname	（仅限双语语料）
数字	1, 2, 3…				

表4中包含了3组特殊符号（共8个），第一组（Operator1）和第二组（Operator2）都是二元操作符，置于两个项目之间。第三组（Operator3）只有一个符号，是一元操作符，后接一个项目。这些符号的具体含义见表5。

表5　操作符和分隔符的含义与作用

符号	含义
空格	相当于逻辑关系符“且/AND”，用于连接基本项，表示两项同时出现。
|	相当于逻辑关系符“或/OR”，用于连接基本项，表示任意一项出现。
$	$后接一个非负整数n，表示它两边的项按左边的在前、右边的在后的顺序出现在同一句中，且两个项相隔不超过n个字符。
+	+后接一个非负整数n，表示它两边的项按左边的在前、右边的在后的顺序出现在同一句中，且两个项相隔n个字符。
#	#后接一个非负整数n，表示它两边的项不考虑先后顺序，出现在同一句中，且两个项相隔不超过n个字符。
-	-后接一个非负整数n，表示它左边的项出现在句中，且在其右边n个字符范围内，右边的项不出现。
~	~后接一个非负整数n，表示它左边的项出现在句中，且在其左边n个字符内，右边的项不出现。
!	表示它右边的简单项在显示查询结果时高亮且居于一行的中心位置。
:	将关键字和简单项区隔开，表示简单项是关键字所代表特征的取值。

需要补充说明的是：

（1）Operator2这一组操作符中，“$”和“+”这两个操作符可以在查询表达式中多次使用，且两个操作符还可以组合使用。其他三个操作符仅能使用一次，并且只能单用，不能与同组其他操作符同时使用。

（2）西文冒号“:”总是跟在关键字之后使用。关键字author代表“作者”，pattern代表“重叠模式”，name代表“中文语料文件名”，type代表“文章类型”，ch代表“中文句子”，en代表“英文句子”，translator代表“译者”，enname代表“英文语料文件名”。这些关键字相当于语料的特征，可以在全文检索的同时，进一步指定这些特征的值，从而达到更精准检索（或缩小检索范围）的目的。有点类似过滤操作，因此上面表4又把“:”称为分隔符，用于构成查询表达式中的“过滤项”。

（3）西文叹号“!”后接一个简单项，标示该简单项是查询表达式中的主要查询条件，即中心词。在显示查询结果时，该项匹配的字符串将置于一行的中心位置，并高亮显示（参见下文3.6）。如果查询表达式不包含“!”，则默认第一个简单项为中心词。一个查询表达式中有且仅有一个中心词（可以有0或1个西文叹号）。

为帮助理解上述符号的含义和用法，表6给出了一些查询表达式的示例。

表6　查询表达式示例

查询表达式	查询表达式的含义（检索目标）
了。|了！|了？	“了”后紧接句号或叹号或问号（大致相当于查询句尾“了”）。
了$0（。|！|？）	同上。
author:老舍 pattern:A来A去	“老舍”作品中符合“A来A去”模式的句子。
吃$5亏	“吃”在前，“亏”在后，二者共现且相隔5个字符以内。
因为#10所以	“因为”跟“所以”共现（顺序无关），且二者相隔10个字符以内。
（把|被）$10!给	同时含有“把”和“给”的句子，并且“把”在先，“给”在后出现，二者之间相隔10个字符以内；或者，同时含有“被”和“给”的句子，并且“被”在先，“给”在后出现，二者之间相隔10个字符以内。显示查询结果时，以句子中的“给”为中心词，居中定位，高亮显示。
被$10把$3!给$2了	“被、把、给、了”四个词在一个句子中共现，并且“被”在“把”前出现，二者之间相隔10个字符以内，“把”在“给”前，二者之间相隔3个字符以内。“给”在“了”前，二者之间相隔2个字符以内。显示查询结果时，以句子中的“给“为中心词，居中定位，高亮显示。
ch:以太网 en:Ethernet	汉语句子中包含“以太网”，英语句子中包含“Ethernet”的汉英对照句对。
和服-0（务|装）	含“和服”但不含“和服务”“和服装”的句子。

表6中例1和例2的查询结果相同。这也附带说明了，为达到一个查询目的，查询表达式可以有不止一种写法。例1和例2的查询结果在显示时会有细微差异。例1中“了。”是一个简单项，因此会作为一个检索单位，居中定位显示。例2中“了”是一个简单项，“。”是一个简单项，二者紧邻出现，会作为两个检索单位，其中“了”是居中定位的词语。

值得一提的是，CCL语料库检索系统是搭建在开源的全文搜索引擎工具包

Lucene之上的（详见第4节）。作为全文搜索引擎，一般会屏蔽标点符号这类很少被搜索的符号[5]。但是，考虑到语言学研究中，标点符号是一类重要的字符，CCL语料库检索系统也支持对标点符号的检索，将标点符号跟一般汉字等同看待。在上面的查询表达式示例中，例1和例2展示了标点符号的作用，可以在句号等标点的辅助下检索句尾包含“了”的句子。

这里再举一个例子说明标点符号检索的作用。比如在比较“高兴”和“快乐”的用法差异时，查询它们跟书名号《》共现的情况，查询表达式分别为“《$5高兴$5》”和“《$5快乐$5》”。前者在CCL语料库中仅检索到5条结果；后者则检索到214条结果。不难发现，“快乐”用于标题的概率远多于“高兴”。这无疑可以为分析二者的词义和用法差异提供一定的线索。

查询表达式中的“基本项”是不含操作符的任意字符串。系统关键字如果不紧跟西文冒号，也会被当作普通字符串看待，比如查询表达式“author”，将返回的结果是包含author的文本行。对于“基本项”的搜索规则，汉语是精确匹配。英语是兼容词形变体的精确匹配。例如在查询单词“take”时，会将took、taken、taking、takes等同时作为匹配目标，返回包含这些词形的文本行。

3.2 批量查询

在批量查询页面，用户可以把符合格式规范的多个查询表达式写在一个文本文件中，每个查询表达式占一行，然后将该文件上传到CCL语料库检索系统，进行批量查询。系统默认允许的最大查询个数为30个查询表达式。文件需采用GBK编码，不支持UTF-8编码。

批量查询可以把用户感兴趣的语料库检索任务集中在一起，一次性完成。比如，用户想对比“把”字结构跟“着、了、过”分别共现的情况，就可以使用批量查询来实现。表7为批量查询“把$4了”“把$4着”“把$4过”这三种格式的查询结果[6]。返回的结果网页中列出了每个查询表达式命中的结果的个数。每个查询表达式上都有一个超链接，点击后可进入该查询表达式对应的具体查询结果。对比显示，“把”字结构跟“了”共现的频率远远高于“着”和“过”。

表7 批量查询返回的结果页面示例

查询	结果数	状态	耗时
把$4了	31,229	成功	312.0 ms
把$4着	3,487	成功	140.0 ms
把$4过	6,769	成功	156.0 ms

3.3　模式查询

在模式查询页面，用户可以指定特定的模式检索跟该模式匹配的例句，例如“爱X不X”“X来Y去”等，模式中字母为变项，相同字母代表相同的文字，不同字母代表不同的文字。例如：“X来X去”将匹配包含“跑来跑去、说来说去”等的例句；“X来Y去”将匹配包含“思来想去、颠来倒去”等的例句。

为了与原文中的字母进行区分，在模式查询表达式中，要求匹配的变项字符用括号括起来。此外，变项的长度也可以由用户指定。表8是一些模式查询的示例。

表8　模式查询示例

模式查询表达式	含义	返回结果示例
爱（X）不（X）	“爱”与“不”被一个字符串X隔开，且X在“不”之后重复出现	爱理不理 爱搭理不搭理
（X，=2）不（X）的问题	字符串X长度为2，在“不”前跟“不”后重复出现	面子不面子的问题 原则不原则的问题
（X，=1）的不是（Y，<4），是（Z，2-4）	字符串X，Y，Z顺序出现在“…的不是…，是…”中的“…”位置，其中X为1个字符，Y为1到3个字符，Z为2到4个字符	冷的不是夜，是孤独 玩的不是魔术，是科学 望的不是天花板，是蓝天 卖的不是熊掌，是牛蹄筋

模式查询表达式中的变项可以有三种方式指定字符串长度。如表8中所示：（X，=2）表示变项X为2个字符长度；（Y<4）表示变项Y长度小于4个字符，即1到3个字符长度；（Z，2-4）表示变项Z的长度介于2到4个字符之间。相同的两个变项，长度也相同，如果给两个相同变项指定不同的长度，则系统会报错。相同变项的长度只需指定一次即可。此外，变项长度未指定时，系统默认变项长度为1-10个字符。也就是说，模式查询表达式“爱（V）不（V）”等价于“爱（V，1-10）不（V）”。此外，需要注意的是，模式表达式中的“V”不是代表动词，“V”跟“X”的作用是一样的，仅代表变项。模式表达式中相同字母代表相同的变项，不同字母代表不同变项。

模式查询功能返回结果可以像普通查询一样，返回原文例句，也可以点击模式查询页面上的“统计”按钮，对模式中的变项进行频次计数，并按频次大小降序输出。比如查询模式“爱（X，<3）不（X）”中变项X的统计信息为：X共有51种，频次最高的前5个是：爱理不理：98；（X，理）、爱得不得：18；（X，得）、爱信不信：18；（X，信）、爱听不听：8；（X，听）、爱吃不吃：8；（X，吃）。跟模式匹配的实例后面是出现频次值，括号中是变项X的值。

3.4 高级查询

在CCL汉英句对齐语料库中，提供了高级查询页面。该页面的查询功能跟3.1节介绍的普通查询是一样的。但通过区分不同的查询关键字段，使得用户表达查询意图更为直观方便。用户可以根据自己的需要，指定语料的作者、译者、类型、模式（特指汉语的各种重叠形式）、中文篇名、英文篇名等作为查询条件。各条件之间是逻辑“并”的关系。比如在“作者”字段指定查询“谌容”，在模式字段指定查询“AA”（表示单音节重叠），则相当于普通查询模式下输入查询表达式：author:谌容 pattern:AA。

3.5 指定查询范围

CCL语料库默认情况下是针对整个语料库进行检索。同时也提供了用户选择语料查询范围的功能。选择的范围可以具体到若干篇特定的文本。上文3.1节已经提到，CCL语料库在文件目录结构和文件名中标记了语料的一些篇章信息，比如“口语”目录下，存放的是口语体的相关文件，在“西汉”目录下，存放的是西汉时期的文献语料，文件“张承志北方的河.txt”的文件名中包含了作者“张承志”的信息。在查询范围选择页面上，CCL语料库的全部文件目录和文件名信息以树状方式呈现，用户可展开每个目录及其子目录直到列出该目录下包含的所有文件，通过勾选树节点前的方框，来指定将某个特定的目录或文件列入查询范围。

图1显示了选择语料范围的界面，左图是现代汉语语料文件目录的示意，右图是古代汉语语料文件目录的示意。左图中勾选的查询范围是：“口语”目录下的“1982北京话调查资料.txt”文件和“电视访谈”目录下的所有文件。右图中勾选的查询范围是：“春秋”目录下的“左传.txt”以及“西汉”目录下的“刘向战国策.txt”这两个文件。

图1 查询范围选择界面

通过上面的界面指定查询范围，跟在普通查询表达式中用过滤项来表示检索条件可以达到相同效果。例如：“现代”目录下有子目录“小说”，其下有文件“老舍四世同堂.txt”“老舍短篇.txt”“老舍长篇1.txt”“老舍长篇2.txt”等四个文件是作家老舍的语料。在查找范围页面上勾选这4个文件，相当于在普通查询表达式中指定“author:老舍”作为查询项。

3.6　查询结果的显示与下载

CCL语料库的查询结果以原始语料文件（纯文本格式）中的一个自然文本行为单位输出显示，用户可以指定查询结果的显示长度（左右n个字范围），默认为一行60个字符。如果想显示查询关键字所在的整行，可以通过指定足够大的显示长度（比如1,000）来实现。当用户指定的显示长度超过原文本行的字符长度时，以原文本行长度为限显示查询结果。查询结果中，被查询项会被标成红色，称为标亮词。中心词是特殊的标亮词，在显示查询结果的每行文本时，以中心词位于页面水平中心位置对齐。用户可用操作符“！”指定中心词。若不指定，则默认第一个标亮词为中心词（参见3.1节表5和表6的说明）。下面是查询“被$10把$3!给$2了”的结果页面示例，“被、把、给、了”四个词为高亮词，均以红色加下划线形式显示，“给”是居于中心位置的高亮词。

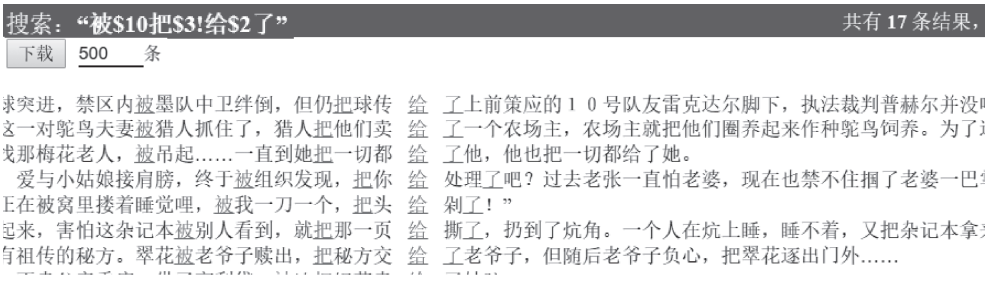

图2　检索结果页面示例

如图2所示，检索结果页面左上角位置有“下载”按钮，用户可指定下载的查询结果条数（默认为500条），点击“下载”按钮，可将查询结果以本文文件（*.txt）格式保存至本地电脑。每句之后注明该句所在文件名、文件作者等信息。

CCL语料库检索结果的计数规则是每一个命中查询表达式的实例计1次。在显示检索结果页面，如果一个文本行（相当于自然段）有多个实例命中，则每次命中均占一行显示。在该行的左侧用两级编号来标识。比如“1.1”表示第一个文本行的第一条命中记录，“3.2”表示第三个文本行的第二条命中记录。

关于查询结果的显示，CCL语料库还提供排序和扩展功能：排序功能可对查询结果按照中心词左边或右边的字符进行排序，排序依据为字符的计算机内码升序或降序。扩展功能包括扩展显示命中记录的“上下文”，“在结果中检索”等。限于篇幅，这里就不展开说明了。

4. CCL语料库检索系统的开发

CCL语料库的设计理念是在原始未标注文本基础上提供尽可能丰富的检索功能，方便语言研究和教学工作者查找例句。因此，检索系统的程序实现，采取了在开源的全文检索工具包基础上，根据语言研究需要再做二次开发的路线。这样相当于站在巨人的肩膀上借力登高望远，可以大大缩短开发周期，同时实现比较好的开发效果。

CCL语料库检索系统的核心引擎基于Lucene开源工具包。Lucene是一套用于全文索引和检索的开源工具包，由Apache软件基金会支持，其Java程序语言的版本被广泛应用于需要进行全文检索的各类应用系统中。Lucene因其索引结构具有可增量维护的特色，检索效率高，问世后很快在计算机全文检索系统开发领域受到关注。为方便将Lucene用于网页环境下的检索系统开发，全文检索系统开发人员在Lucene基础上，又进一步搭建了以XML文件格式来包装Lucene核心检索引擎的WebLucene，其系统框架如图3所示。

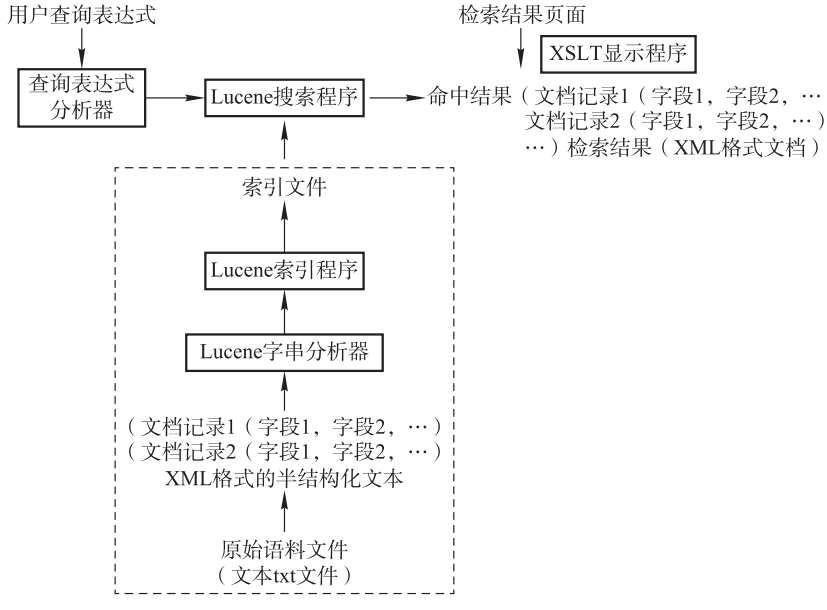

图3　Lucene检索系统的工作框架

图3中程序模块用实线方框表示，比如“查询表达式分析器”就是一个程序模块。其余文字内容表示的是各种数据（字符串，文件等），比如“用户查询表达式”就是一个字符串；“检索结果页面”就是一个基于html（超文本标记语言）的网页文件。跟一般的检索系统架构类似，Lucene也分为“索引”和“检索”两个部分。图3虚线框内是索引模块。在索引环节，WebLucene的作用是将原始txt文本文件改为XML格式的文件，相当于把无结构的文本，转为半结构化的类似数据库格式的文件：一个文件由若干个文档记录（doc）组成，一个文档又由若干个字段（field）组成（比如语料的“作者、类型、重叠模式”等等，都可以表示为单独的字段），这样就可以针对不同字段，设置更具针对性的索引，从而在检索阶段，提供更为丰富的检索条件组合功能。图中Lucene字串分析器包含了一个针对汉字的CJK字符分析器，专门处理汉字文本。Lucene索引程序以这样的XML文件为输入，生成索引文件，索引文件包含了原文内容，因此实际检索系统只需要访问索引文件获取检索结果，不再需要从原始文件中抽取检索结果。在增加语料文件时，Lucene索引可以不改变现有的索引文件，实现增量式索引构建，便于语料库的扩容或订制专用的语料库。在检索环节，WebLucene的作用是将检索结果输出为网页，支持用户通过浏览器访问检索结果。

图3所示的索引和检索框架是对一般的全文检索系统而言的，并不能完全满足语言学研究的需要。以语言研究为目的检索例句，往往是以句子为搜索范围的。而且，语言研究时查询关键字除了普通的词语，往往需要检索不连续共现情况（比如离合词用法），或者符合某些形式特征的字符串模式（比如动词重叠等）而不是确定的词语字符串。针对上述检索需求，就有必要对原始文档做分割，设计特定的查询表达式，并增加索引字段。上文3.1和3.3节详细介绍了CCL语料库检索系统设计的操作符和多种组合的查询表达式，以及模式查询表达式，具体实现的策略包括：

（1）在原始语料预处理阶段，将原始纯文本格式文件转为XML格式文件时，一个原始文本按照自然段落切分为若干个文档记录，即相当于把一个句子（或一个自然段）看作一篇文章，从而在Lucene全文检索系统（针对一篇文章进行检索）的框架中达到为语言研究服务的目的；

（2）在原始语料预处理阶段，将查询表达式中定义的关键字，均设置为XML文件中的字段，从而为这些字段的文本内容建立独立的索引，为后续的查询做准备；

（3）在“查询表达式分析器”程序中，将CCL检索系统定义的复杂查询表达式分解为若干项简单的Lucene查询表达式的组合，然后交由Lucene检索程序去访问索引文件查询每一个简单的查询表达式，最后将各个简单查询表达式的检索结果进行合并，返回最终结果；

（4）在模式查询中，先利用查询表达式中确定的字符串（常项）部分，按照

普通查询表达式的检索办法，得到命中结果的文档集，然后根据模式查询表达式的模板槽（变项）部分，生成字符串方程，通过求解方程，过滤出文档集中满足条件的特定文档（例句）。方程需要满足两方面的条件：一是方程的字符串解拼接后能还原得到原文字符串；二是字符串解要满足长度限制。例如，查询“（X，=2）不（X）的问题”这一模式，检索系统先按照普通查询表达式“不$2的问题”进行查询，先把该表达式拆解为“不”和“的问题”两个基本项，在索引文件中查找包含这两个字符串的文档，求交集，并核查“不”和“的问题”在原文中的位置偏移量之差是否不超过2个字符，将这样得到的结果文档集编号作为第一个阶段的结果输出。假设该集合中包含两个字符串：A. 不是需不需要的问题 B. 不是文化不文化的问题。A对应的字符串方程为：X=“是需”，Y=“需要”，Length（X）=2；B对应的字符串方程为：X=“文化”，Y=“文化”，Length（X）=2。显然，A文档，X≠Y，不符合查询要求。B文档符合查询要求，作为结果，将B文档“文化不文化的问题”作为最终的实例结果输出，同时输出模式中的变项X的字符串解为“文化”。

简而言之，CCL语料库检索系统在Lucene和WebLucene全文检索引擎的架构基础上，针对语言学研究的需要，在查询表达式解析，检索结果后处理等环节，做了许多针对性的改进，丰富了检索功能。

此外，为提高在网络环境下响应检索请求的效率，CCL语料库检索系统还引入了Memcached服务器。Memcached是一套分布式的高速缓存系统，常用来提高网站的访问速度。因为模式查询中求解字符串方程是实时进行的，并无事先索引，因而耗时较长。针对这一问题，CCL语料库检索系统利用Memcached服务对模式查询的结果进行了缓存处理。当用户的模式查询发送到服务器，服务器会首先查看Memcached服务器中是否有对应的结果，若已有结果则直接返回，若无再执行具体的模式查询，并将查询结果以json格式存储到Memcached中。

5. 结语

语料库的建设总是包含着两个重要的方面，一是选什么语料，二是语料如何使用。CCL语料库是2003年开始设计，历时一年多完成了第一版的系统开发。当时确定的选材原则是语料要规模尽可能大、覆盖领域尽可能多。开始选取的主要是相对传统规范的文本，后期在发展过程中又逐渐吸收了一定比例的网络语料，以反映汉语在新时期的发展情况。在语料使用方式方面，则是确定了通过互联网提供免费查询服务的原则。当时谷歌、百度等互联网搜索网站问世不久，正逐渐成为新一代的主流检索工具，而在面向语言研究的检索方面，中国还没有类似的在线语料检索系统。北大CCL语料库是比较早做此尝试的系统。在这样的大背景

下，CCL语料库上网发布不久就引起了海内外汉语学界的关注[7]，逐渐成为汉语研究和教学领域非常常用的语料库之一。

近十年来随着互联网的飞速发展，世界逐渐进入到“大数据”时代。语料库的规模也已从亿字级跨入百亿字级甚至万亿字级。越来越多的大规模在线语料库开始提供多语言的查询服务，还有的语料库系统把互联网搜索引擎作为语料来源，以整个网络的文本资源作为检索对象。语料库检索系统也不再只是提供例句作为检索结果，而是在检索的基础上进一步提供丰富的数据分析以及数据可视化。可以说，语料库的类型越来越丰富，检索手段和结果呈现形式也越来越多样。

在新的形势下，CCL语料库也在谋求新的发展。不过，正如十多年前创建时，CCL语料库基本上是自底向上（bottom-up）技术路线的产物，缺少一个自顶向下（top-down）的顶层设计。现在CCL语料库的升级之路，也仍然将延续这一方式：主要是利用现有的技术手段，逐步将更多类型的语料融入到CCL语料库检索系统中。近年来，北京大学中国语言学研究中心在一些项目的支持下，出于课题研究的需要，陆续收集和构建了一些专项语料库，包括：（1）早期北京话材料（如近代西人北京话教科书汇编、日本北京话教科书汇编、清末民初京味小说书系等）；（2）留学生汉语作文语料；（3）汉语构式语料库；（4）中文学术文献语料库；（5）海外华文网络语料等等。这些语料将以专题语料库的形式，融入现有的CCL语料库中。同时，在语料预处理方面，将尝试做一定的中文分词和词性标注；在检索功能方面，将提供更多样的统计数据信息，并增加对一些检索结果的可视化支持。此外，目前的CCL语料库仅针对用户的检索请求提供单向的查询结果反馈。未来将考虑增加用户与CCL语料库之间的交互功能，为用户提供更多的定制服务。总而言之，CCL语料库的发展愿景仍然是以建设“国际一流的汉语语言学研究信息资料库”为目标，希望在广大用户良性反馈的基础上，通过合理扩容和功能升级，为学界提供更优质的服务。

注　释

1. 可参见冯志伟（2002，2006）对世界上语料库发展的历史做的详细介绍；詹卫东（2018）对全球范围内近三十年来中文语言资源的建设和应用情况所做的述评。
2. CCL语料库网址http://ccl.pku.edu.cn:8080/ccl_corpus或http://ccl.pku.edu.cn/corpus.asp。
3. 对语料库规模的定量表示一般以字符（如汉字、英文字母、标点符号等）或词语数量为单位。CCL语料库中文文本未经分词，无法计量词语个数。CCL语料的文本均为GBK编码，即一个中文字符在计算机中以两个字节表示。这样字节数与字符数的对应

关系大致为2：1。12亿字节相当于6亿字符。除标点、非汉字的字母、阿拉伯数字等字符外，汉字字符数约为5亿。有关CCL语料库文本类别及字数统计的更多信息可访问网页查询：http://ccl.pku.edu.cn:8080/ccl_corpus/corpus_statistics.html。

4. CCL语料库古代汉语文本约2亿字符，1.637亿汉字。
5. 全文搜索引擎一般会把不大可能有人搜索的符号，比如标点符号、虚词（如“的”）等做屏蔽处理，即把这类符号加入搜索系统的停用词表（stopword list）中。
6. 在普通查询页面，指定查询表达式“把$4（了|着|过）”也可检索出所有包含“把”跟“了、着、过”分别共现的例句，但检索结果是混在一起计数，并返回例句的。批量查询模式是对3个查询表达式的检索结果分别计数，分别返回每个表达式的检索例句。
7. 2005年CCL语料库检索系统的日志文件显示当年的日均查询量达到5000次。很有意思的一个现象是，按月统计2005年CCL语料库的检索量中，5-6月份，12-1月份为全年检索量的两个高峰值，绝大多数检索的来源IP地址都显示检索请求来自中国教育科研网（CERNET）的网段范围，即集中在高校科研机构。很可能是在这两个时间段内，由于课程期末论文或学位论文的需要，产生了大量的语料检索需求。

参考文献

Gries, S. 2012. Corpus linguistics, theoretical linguistics, and cognitive psycholinguistics: Towards more and more fruitful exchanges [A]. In J. Mukherjee & M. Huber (eds.). *Corpus Linguistics and Variation in English: Theory and Description* [C]. Amsterdam: Rodopi. 41-63.

Leech, G. 1993. Corpus annotation schemes [J]. *Literary and Linguistic Computing* 8(4): 275-281.

Leech, G. 1997. Introducing corpus annotation [A]. In R. Garside, G. Leech & A. McEnery (eds.). *Corpus Annotation: Linguistic Information from Computer Text Corpora* [C]. London: Longman. 1-18.

Leech, G. 2005. Adding linguistic annotation [A]. In M. Wynne (ed.). *Developing Linguistic Corpora: A Guide to Good Practice* [C]. Oxford: Oxbrow Books. 17-29.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse* [M]. London: Routledge.

Teubert, W. 2005. My version of corpus linguistics [J]. *International Journal of Corpus Linguistics* 10(1): 1-13.

冯志伟，2002，中国语料库研究的历史与现状 [J], *Journal of Chinese Language and Computing*（1）：43-62。

冯志伟，2006，《应用语言学中的语料库》导读 [A]。北京：世界图书出版公司。

荀恩东、饶高琦、肖晓悦、臧娇娇，2016，大数据背景下BCC语料库的研制 [J]，《语料库语言学》（1）：93-109。

詹卫东，2018，近30年来中文语言知识资源发展及应用 [J]，《语言战略研究》（4）：58-69。

通信地址： 100871 北京市北京大学中国语言文学系（詹卫东、郭锐、陈龙）/ 100871 北京市北京大学计算语言学研究所（常宝宝、谌贻荣）