# 大数据背景下BCC语料库的研制

北京语言大学 荀恩东 饶高琦 肖晓悦 臧娇娇

**提要：**“北京语言大学语料库中心（BLCU Corpus Center，简称BCC）”是以汉语为主、兼有其他语种的在线语料库。BCC总规模达数百亿字，是服务语言本体研究和语言应用研究的在线大数据系统。BCC检索式由字、词和语法标记等单元组成，并且支持通配符和离合查询。本文将概述BCC的总体情况，包括语料库建设情况和检索引擎开发等，重点介绍BCC形式化检索语言和在线系统的使用方法。

**关键词：**BCC语料库、大数据、语言检索、检索式

## 一、引言

在大数据背景下，语言本体研究、语言教学和语言应用研究都离不开语料库的支持。在语言本体研究中，利用大规模语料，对语言现象进行穷尽式考察，可以归纳、完善、验证语言理论或观点，又可以通过实证方法，为语言理论的研究提供数据支撑和量化分析；在语言教学中，语料库可以提供真实的语言素材，用于教学内容制定和讲解，使语言教学内容选取和教学实施过程更加科学，并可以支撑辞书和教材的编纂；同时，语料库作为模型训练知识库，在语言信息处理各种应用中起着不可或缺的作用。

采用语料库进行实证研究历史悠久，国内外一系列语料库系统推动了语言研究的进步和发展。中文语料库方面，有“国家语委语料库”、“北京大学现代（古代）汉语语料库”、“中国台湾中央研究院语料库”、“兰卡斯特汉语语料库”等；在英语语料库方面，有“英国国家语料库（BNC）”、“美国当代英语语料库（COCA）”等。语料库发展到今天，出现了新的特点和需求：

1）语料库规模越来越大，逐渐进入大数据时代。随着信息社会的发展，个人微机的迅猛发展和存储数据的硬盘造价持续下降，使得能够记录语言生活的终端设备越来越普及，数据存储能力越来越强，网络传输速度越来越快，每天产生的语料数量大大超过以往。这些发展都为大规模语料库的采集提供了技术支持。

2）语料库成为语言技术进步的知识库。在语言大数据基础上，语言应用技术快速发展，人工智能在多个应用领域取得突破性进展。这些新技术进步，正在改

变社会语言生活，为语言研究不断提供新课题并提出新的挑战。

3）语料库形式多样。语料的领域越来越细化，语料加工越来越深入，网络社交语料异军突起。

4）语料库使用越来越便捷。在线语料库查询和统计功能更加人性化，除了面向个人在线使用外，语料库还利用云服务接口，通过云调用大大拓展了语料库的应用范围。

“北京语言大学语料库中心（BCC）”（http://bcc.blcu.edu.cn）是以汉语为主、兼有其他语种的语言大数据，目标是为语言本体研究提供一个使用简便的在线检索系统和构建大数据的语言应用基础平台。BCC支持云服务，通过API调用方式为开展知识抽取、模型构建等研究和应用工作提供便利。

本文首先概述BCC研制的总体情况，重点介绍BCC检索式，并在附录中给出了BCC检索式实例和中英文词性体系。

## 二、BCC语料库研制

一个语料库系统的建设，主要包括三方面工作：语料库资源建设、检索引擎开发和提供语料库检索服务。如图1所示，语料库的资源建设是构建语料库数据内容的基础。BCC主要包括三种类型语料：多语种单语语料库、双语对齐语料库和深加工的树库。语料库检索内核是实现语料库系统的技术基础，采用基于后缀串的全文检索算法，并且支持通配符和离合模式匹配。检索服务是指使用语料库系统的方式和方法。BCC提供两种服务方式：在线检索和云调用。

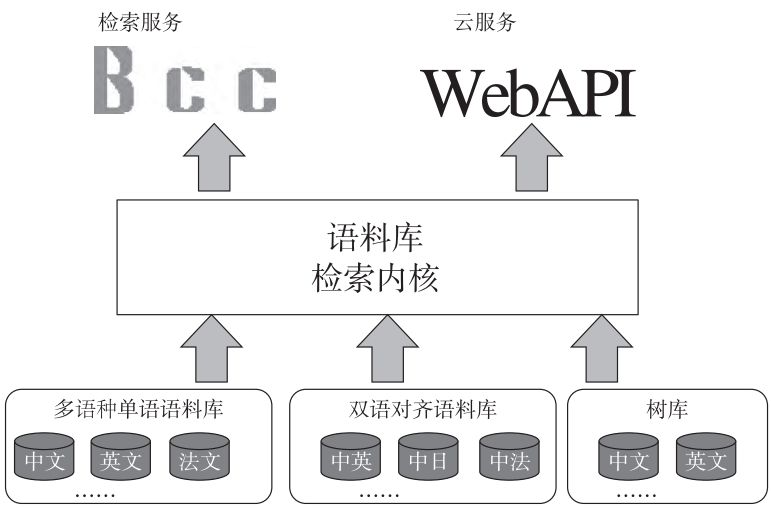


图1.　BCC语料库系统示意图

## 2.1　语料库资源建设

语料库建设是指在确定语料库内容、规模和形式后，对语料进行采集、加工和标注等，通过对自然语言文本的采集、存储、加工，可以凭借大规模语料库提供的客观语言事实为语言学研究提供支撑（黄昌宁、李涓子 2002）。BCC语料库具有以下特点：

语料库函盖多个语种

以汉语为主，兼顾其他语种的语料。目前BCC包含9种语言，如英语、西班牙语、法语、德语、土耳其语等。其中的英文语料主要采自《华尔街日报》，规模约为12亿单词。BCC语料以单语语料为主，也包括双语平行语料，如英汉、英德等双语对齐语料库。目前有9种语言互译，各类双语语料总规模约千万句。检索时，汉语最小的单位是汉字，其他语种最小的单位是单词，但单词不支持词形变化，保持原始语料中的形态，例如：英语The和the在语料库中是两个单词。

多层次语料加工

包括生语料、分词语料、词性标注语料和句法树。目前已对现代汉语、英语、法语的语料进行词性标注，除此以外的其他语料都是未加工的生语料；句法树包括中、英文树库，分别引自美国宾州大学的中文和英语树库。语料加工层次不同，支持检索的功能也不同，例如：生语料不支持带有词性信息的检索，树库支持短语类型标记的检索。

现代汉语语料和古代汉语语料兼具

对现代汉语语料进行了分词和词性标注，支持带有词性信息的检索；而古代汉语没有进行分词和词性处理，只能以字为单位进行检索。

汉语多语体

现代汉语语料涵盖新闻、口语（微博）、科技、文学、综合等多个语体。其中新闻、文学和综合语料标注时间、作者等组成信息，可以用BCC的“自定义”功能进行受限检索，即选择某一个子语料，限定在该语料中进行检索。

新闻语料：采自《厦门日报》、《厦门商报》、《厦门晚报》等；

口语（微博）语料：采自2013年新浪微博；

科技语料：采自国内学术期刊；

文学语料：采自国内外文学作品，对每个作品都标注了作品名称、作者、发表时间等信息。

综合语料：包括报刊、文学、微博、科技四个领域，语料内容独立，与其他语料不交叉，目标是建立一个“平衡”语料库。

共时语料和历时语料兼备

BCC对报刊语料和文学作品标注了时间信息，其中文学作品的时间信息体现在BCC的“自定义”功能应用上，用户可以选定某时间的文学作品进行限定检索；BCC“历时检索”主要是报刊语料，语料来自1945年至2015年的《人民日报》。历时检索是以图形可视化方式呈现的。

BCC语料库使用了语料采集、加工和语言分析处理等多种工具，例如对现代汉语进行分词和词性标注。为了完成语料采集、加工、标注等工作，开发了BCC语料库采集和加工平台，主要包括：

网上语料采集工具

BCC语料库中的语料主要源自互联网的页面文本，利用采集工具自动下载网页，把网页数据保存到本地。

语料加工整理工具

将网络作为语料库，是将以自然语言形式存在的整个网络电子文本当作一个庞大的语料库，可以通过征调主流搜索引擎的应用程序调用接口，获取搜索引擎的返回结果，再对其进行相应的语料库统计分析（熊文新 2015）。BCC语料加工整理的方式主要为：从网页中提取原数据信息，包括名称、出处等；网页数据清洗，从网页数据中剔除非内容数据，提取有效文本内容；对数据进行自动断句处理，为后续语言分析做准备；异常重复句子甄别和处理，剔除网页数据清洗阶段不能甄别的重复句子。

语言自动分析工具

原始语料完成断句后，在语言分析阶段对句子进行分词和词性标注处理。中文词性标注采用北京大学计算语言研究所提出的词性标注体系（俞士汶等 2000，2002），英文词性体系采用美国宾州大学词性体系。目前，BCC可以对现代汉语、英语、法语的语料进行自动分词和词性标注处理。

语料库标注平台

该平台的目标是通过人工标注来构建专门语料库。

## 2.2 BCC检索引擎

语料库建设是围绕内容进行的，用户通过检索使用语料库数据，而使用的检索功能是通过检索引擎实现的，因此检索引擎的性能直接影响语料库系统的使用体验。使用体验体现在多个方面，包括对数据规模的支持程度、语料类型的支持程度、响应检索的时空开销、检索式的支持功能、对服务器软硬件的适应性等。BCC检索引擎具有以下特点：

1）支持语言大数据。目前BCC检索内核支持建立超大规模语料库检索系统，单机可以索引的语料库规模最大可以支持64G（约320亿汉字），实际规模与机器内存相关。

2）支持多语种检索。BCC语料库检索内核技术支持中文、英文、日文等不同语种的语料库。

3）支持多种语料形式。BCC语料库包含原始语料、分词语料、词性标注语料，同时可以支持短语结构树的语料库检索。

4）支持功能强大的检索。BCC定义一种用户友好且功能强大的语料库检索语句，不仅具有模式查询和统计功能，支持带有词性的通配符和离合模式查询，还可以支持二次查询、自定义语料查询等，同时BCC还实现了在线统计以及在线反馈统计结果的功能。

## 2.3　语料库服务

BCC语料库服务包括两种形式：一种是在线检索，即在浏览器内使用BCC，输入检索式，以页面形式返回结果；另外一种是云服务，通过编程使用BCC的Web API接口形式来调用BCC。云服务一般用于BCC的二次开发，或者用于利用BCC进行语言的应用开发。

在BCC首页中可以选择不同语种的语料库，在输入框的上方，列出该语种的不同语料频道（如图2）如果想在某个频道中做更细化的查询，可以选择“自定义”搜索（如图3）通过点击语料库的组成窗口选择子语料库或者通过搜索定位子语料库。当用户选择一个子语料时，页面会给出该子语料库的语料规模，后续检索也会限定在该子语料库中进行。


图2.　BCC首页


图3. 通过“自定义”选择语料及查看语料组成和规模


图4. “帅气的n”检索结果页

输入检索式，点击“搜索”后得到检索页面，其中包括检索总条数信息、分页显示的检索实例等，如图4是“帅气的n”的检索结果。在搜索结果页面，BCC还提供在线统计、二次检索、下载结果、显示结果和查看原文等多个功能：

1）统计：BCC检索式中可以包含词性或者短语类型，也可以带有通配符。在结果页面，词性和通配符体现在具体的检索实例中。BCC通过统计实例，在线统计检索式在语料库中的分布情况，统计结果页如图5所示。

2）筛选：筛选提供在线二次检索的功能，即在现有的返回结果中保留或者剔除符合检索式的语料实例，得到二次检索结果。筛选的检索式同一般的BCC检索式一致。

3）下载：下载检索或统计结果，登录用户可以下载更多的结果。

4）高级：可以设置返回结果的显示形式；可以随机生成实例，如设定上下文显示字数，设定是否以句形式显示结果等。

5）全文：点击该按钮可以查看检索实例更多的上下文。


图5.　统计结果页

BCC语料已经过自动分词和词类标注，并加工成为全文检索索引，制作成为“现代汉语词汇历时检索系统”。历时检索第一代系统（荀恩东等 2015）于2012年上线，使用的语料来源于1946年到2012年的《人民日报》语料，检索结果以年为单位，显示检索式的次数，并以可视化的方式呈现给用户，如下页图6所示。点击图中的每个柱形可以具体显示该年的实例结果。语料仅有分词标注，无词类标注。

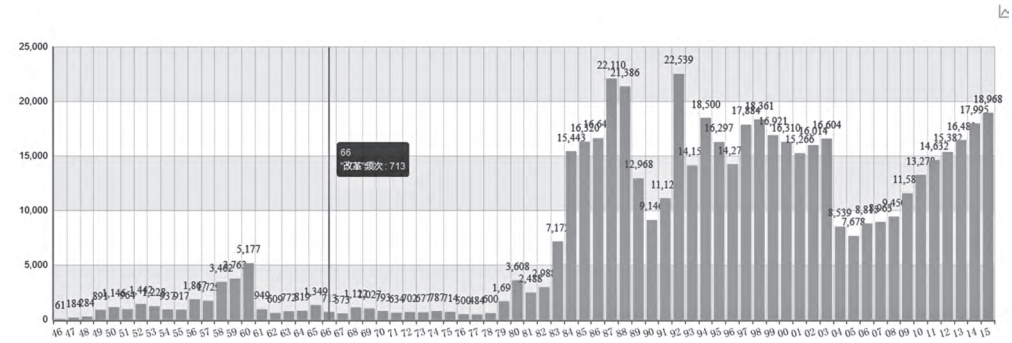


图6. 检索式“改革”频次历时结果

2015年底，历时检索第二代系统上线。历时语料库在分词的基础上增加了词类标注，在保留原有用户体验的同时开始提供多模态检索功能。在该功能的支持下，用户可以在对任意词串（不限于词）进行检索之外对词类串和字符词类混合串进行检索。如图7所示。第二代历时语料库在国内外引起强烈反响，为语言学和许多社会科学领域的相关研究提供了很大的便利（Rao & Xun 2015）。

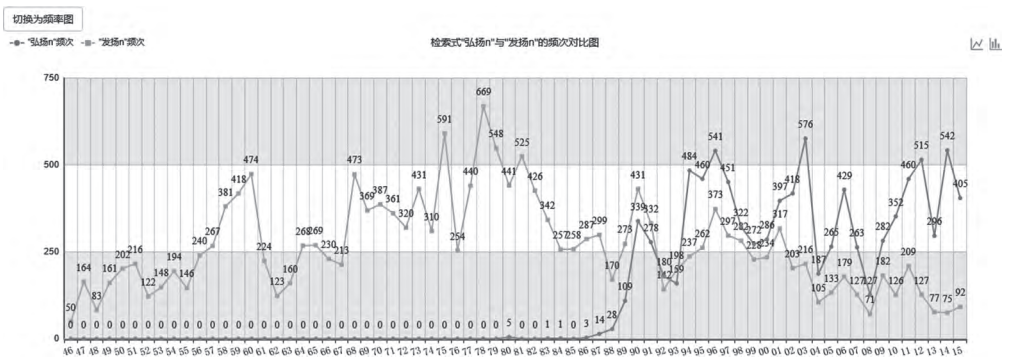


图7. 检索式“弘扬n”与“发扬n”频次历时对比

## 三、BCC检索式

一些语料库采用交互式生成检索式的设计，即以输入查询和界面控件设置相结合的方式进行查询。这种方式有利有弊：如果设置项少，往往限制检索功能的发挥；如果功能复杂，设置项过多，便会影响用户的使用体验。

BCC设计简洁，在界面中没有各种复杂的控件，选定语料后，输入符合语法的检索式可以直接搜索查询。检索式的设计也需要平衡考虑：一般来讲，检索式的语法直接影响语料库功能和用户友好性。复杂的检索式设计可以支持强大的检索功能，但是会对用户学习和使用造成负担。例如，检索系统采用正则表达式的方式，虽然语句标准、功能强，但是不易理解，需要付出更多的学习代价。

为平衡用户友好性和检索功能，BCC设计实现了一种简单的查询语言，即BCC检索式。选取语料，输入检索式，即可查询符合检索式的语言片段。在BCC中，这些语言片段通常是一个（检索式中无离合符号）或者两个（有离合符号）连续出现的字串或者词串。如果检索式中不包含标点或表示标点的词性，检索结果将会限定在一个标点句中。

BCC检索式主要由汉字串（或者词串）、属性符号、通配符、集合符号、离合符号、属性约束符号、空格或“+”组成。

**汉字串（或者词串）：**

汉字串不需要给出分词信息，如果在相邻汉字串之间加入空格或者“+”，BCC也会自动过滤掉这两个符号，检索时作为一个整体字符串匹配。这与通常搜索引擎的使用体验不同。例如：

检索式：“与其说是”，检索包含“与其说是”的实例；

检索式：make a promise，检索包含make a promise的实例；

检索式：“提高 水平”或“提高+水平”，等同于“提高水平”；

检索英语语料库时，按照词的原始形态进行匹配处理，BCC不对单词的形态做处理，也不对单词的大小写进行处理。

**属性符号：**

属性符号是指在标注语料库中，字、词或短语所具有的类型标记。它可以是词性符号、短语类型符号、语义符号等，具体与标注语料库的内容和标注体系相关。目前，BCC中的属性符号主要是指词性符号和短语类型符号，而短语类型符号仅用于具有短语标注的树库中。

BCC中汉语语料库采用北京大学的词性体系，英语语料库采用美国宾州大学的词性体系，具体见附录2。限于语料性质或加工工具，并不是所有BCC语料都有词性信息，即除现代汉语、英语、法语外的其他语种的语料都没有进行词性标注。

为了便于使用，在汉语（除古汉语）和英语（除标准词性标注集）语料库中，同一属性可以用多个等价符号表示，例如：汉语动词可以用v、V、verb、Verb、VERB等不同符号替代。例如：

检索式：“不得不说v”，检索“不得不说”后接属性符号v（动词）的实例。以下形式都是等价的：“不得不说 v”（加空格）、“不得不说 Verb”、“不得不说V”。

检索式：“w 吃 n w”，检索属性符号w（汉语标点符号）后接“吃”再接属性符号n（名词）和属性符号w（汉语标点符号）。这里n和w之间要用空格分隔，表示是两个独立的属性符号。

检索式：v dt problem，在英文语料库中检索属性符号v（动词）后接属性符号dt（定冠词）后面再接单词problem的实例。

**通配符：**

与属性符号相比，通配符代表更宽泛的语言单元，说明某个位置是任意一个汉字或单词。BCC支持三种通配符“.”、“~”和“@”。

1）“.”在汉语语料库中，表示任意一个汉字，检索其他语种的语料库时，表示一个单词。例如：

检索式：“洗...澡”表示检索“洗澡”中间插入三个字的实例。

检索式：make..promise表示检索make promise中间插入两个单词的实例。

2）“~”是汉语语料库专用符号，表示任意一个词。其他语种的语料库则是使用通配符“.”表示任意一个词，而“~”只作为普通符号。BCC限制该符号只能在检索式中出现一次，不支持多个连用的情况。例如：

检索式：“洗~澡”表示检索“洗”后接任意一个词再接“澡”的实例。该通配符通常用来统计上下文词语的搭配情况。该检索式的统计结果见图8示意。

检索式：“w 吃 ~ w”，表示属性符号w（标点），后接“吃”再接任何一个词，再接属性符号w（标点），即检索“吃”后接一个词并单独成小句的实例。该检索式的统计结果见下页图9示意。


图8. 检索式“洗~澡”的统计结果


图9.　检索式“w 吃 ~ w”的统计结果

3）“@”在各种语料库中，都表示任意词，该符号往往用于统计的功能，即统计该位置对应不同词性出现的频次。BCC限制该符号只能在检索式中出现一次，不支持多个连用的情况。例如：

检索式：“w 吃 @ W”，在检索实例时，结果同“w 吃 ~ W”，不同的是检索式的统计结果，如图10所示。


图10.　检索式“w 吃 @ W”的统计结果

**集合符号“[]”：**

在符号“[]”内，可以写多个汉字字符串、单词或者词性，之间用空格分隔，表示可以对应括号内任意一项。例如：

检索式：“[美丽 靓丽]”表示检索包含“美丽”或者“靓丽”的实例。

检索式：“v[上来 下去]”表示检索动词后面接着“上来”或者“下去”的实例。

检索式：“打击[n vn]”表示检索动词“打击”后面接着名词n或者动名词vn的实例。

**离合符号“*”：**

通常，BCC检索式对应连续的字符串。引入该符号的目的是描述语言中的各种离合现象。使用该符号的一般形式为：“检索式1*检索式2”，表示在句子内（对于汉语是小句内），检索符合“检索式1”后接其他成分再接“检索式2”的实例。要注意离合表达的顺序和检索所表达的范围。BCC中限制该符号最多只能出现一次，即不支持多个语言片段连续出现的检索功能。例如：

检索式：“洗*澡”是检索“洗澡”离合出现的情况。

检索式：“见*面”是检索“见面”离合出现的情况。

**属性约束符号“/”：**

该符号作用于一个检索式，后面给出属性符号，约束检索式对应的实例所具有的特定属性，比如“检索式/属性符”的格式。例如：

“./Vg人”表示单音节动词后面接“人”的实例。

“打./v”表示以“打”字开头的双音节动词。

**空格或者“+”：**

除了“/”外，一般情况下，不同表达内容之间需要用“+”或者空格分隔，如果在没有歧义的情况下，也可以连接在一起。例如：

检索式：“我想吃n”，检索“我想吃”后面紧接着一个名词的语言实例。与通常搜索引擎含义不同，在BCC检索式中，有歧义表达时，需要加空格，起到分隔的作用。如“我想吃n”在汉字串后接一个半角的属性符号“n”没有歧义，所以在检索时可以省略空格。

例如：“我们 大家”等同“我们大家”，“打击 n”等同“打击n”。在检索式中，如果连续出现两个或多个词性标记，或在外文语料库检索时，单词之间要用空格分隔。例如：“一q n”，表示检索“一”后面连着一个量词，量词后面是一个名词的实例。多个词性相连时，用‘ ’（空格）分隔。另外，空格在集合符号“[]”中使用，用来分隔多项内容；配合“/”使用，可以用来表示词边界。

## 四、结语

BCC语料库为语言本体研究提供数据和技术支持，在大数据背景下，可以证实、证伪或者发现语言现象；BCC作为语言应用开发的基础平台，为信息抽取、构建知识图谱、语言自动分析等提供便利；同时，也为语言教学研究提供统计数据和实例支撑等。

BCC是动态发展的，本文没有提供现有BCC在线服务语料的细节信息，最新的语料和规模可以通过BCC的“自定义”功能或在线说明文档获得。今后BCC将会纳入更多的语种、更大规模的数据、更多形式的语料，从文本语料向多模态语料拓展，从语法属性为主的检索向语义信息检索方面发展。BCC的建设目标是打造一个大型知识库。从一个语料库发展成为一个知识库，这不仅能支持语言本体研究，也能为语言相关应用的研发提供知识支撑。

### 参考文献

Rao, G. & E. Xun. 2015. Words and characters in official newspapers since the founding of the PRC: *Guizhou Daily and People's Daily* as examples [J]. *International Journal of Knowledge and Language Processing* (2): 23-33.

黄昌宁、李涓子，2002，《语料库语言学》[M]。北京：商务印书馆。

熊文新，2015，《语言资源视角下的语料库建设与应用研究》[M]。北京：外语教学与研究出版社。

荀恩东、饶高琦、谢佳莉、黄志娥，2015，现代汉语词汇历时检索系统的建设与应用[J]，《中文信息学报》(3)：169-176。

俞士汶、段慧明、朱学锋、孙斌，2002，北京大学现代汉语语料库基本加工规范[J]，《中文信息学报》(6)：49-64。

俞士汶、朱学锋、段慧明，2000，大规模现代汉语标注语料库的加工规范[J]，《中文信息学报》(6)：58-64。

**附录1. 检索式示例**

构词

| | |
|---|---|
| ../v | 双音节动词 |
| 打./v | 以“打”为首的双音节动词 |
| .性/n | 以“性”为结尾的双音节名词 |
| ./v 货/n | 单音节动词，后接名词“货” |
| ./v ./n | 单音节动词，后接单音节名词 |

搭配

| | |
|---|---|
| 讨论n | “讨论”后邻“名词” |
| ~讨论 | 任意词后邻“讨论” |
| 提高*n | “提高”后面离合接名词 |
| @的提高 | 任意后接“提高” |
| 提高../[vn n] w | 提高句尾后接双音节名词或者动名词 |

离合

| | |
|---|---|
| 洗*澡 | “洗”后接“澡” |
| 洗.澡 | “洗澡”中间有一个字 |
| 洗..澡 | “洗澡”中间有两个字 |
| 澡*洗 | “澡”后接“洗” |

句型

| | |
|---|---|
| 是*的[。？ !] | “是”后接“的”，“的”后面是“。”或“？”或“！” |
| 是*的w | “是”后接“的”，“的”是句尾 |
| 把*v[上下起]. | “把”后接动词，动词后邻“上”或“下”或“起”，后面再接一个字 |
| 被*v[上下起]来 | “被”后接动词，动词后邻“上来”或“下来”或“起来” |
| 被n v一下 | “被”后邻名词、动词和“一下” |
| 被n v一下w | “被”后邻名词、动词和“一下”，“一下”是句尾 |

定界

| | |
|---|---|
| w吃 | “吃”做句首 |
| w吃.W | “吃”做句首的二字短句 |
| [，。]吃W | “吃”是单字短句，句首前标点“，”或“。”，句尾符号不限 |
| 吃W | “吃”做句尾 |
| [，。]吃[，。] | “吃”是单字短句，句首前标点“，”或“。”，句尾符号是标点“，”或“。” |

构式

| | |
|---|---|
| a不到哪里去 | 形容词后邻“不到哪里去” |
| 还n尼 | “还”后邻名词，再接“尼” |
| v就v | 动词后邻“就”，再接动词 |
| v不着 | 动词后邻“不着” |
| v不到 | 动词后邻“不到” |
| n连n都 | 名词后邻“连”、名词、“都” |
| n连n也 | 名词后邻“连”、名词、“也” |
| 有一种n叫n | “有一种”后邻名词、“叫”、名词 |
| 非[a v n]不可 | “非”后邻形容词或动词或名词，再接“不可” |
| 活活[a v n]死人 | “活活”后邻形容词或动词或名词，再接“死人” |
| 放着n不v | “放着”后邻名词、“不”、动词 |
| v不过n | 动词后邻“不过”、名词 |
| n说起来v | 名词后邻“说起来”、动词 |

**附录2. 词性标注集**

汉语词性列表

| 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 |
|---|---|---|---|---|---|---|---|
| Ag | 形语素 | I | 成语 | o | 拟声词 | vn | 名动词 |
| a | 形容词 | J | 简称略语 | p | 介词 | w | 标点符号 |
| ad | 副形词 | K | 后接成分 | q | 量词 | x | 非语素字 |
| an | 名形词 | l | 习用语 | r | 代词 | y | 语气词 |
| b | 区别词 | m | 数词 | s | 处所词 | z | 状态词 |
| c | 连词 | Ng | 名语素 | Tg | 时语素 | un | 未知词 |
| Dg | 副语素 | n | 名词 | t | 时间词 | h | 前接成分 |
| d | 副词 | nr | 人名 | U | 助词 | g | 语素 |
| e | 叹词 | ns | 地名 | Vg | 动语素 | nz | 其他专名 |

（待续）

（续表）

| 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 |
|---|---|---|---|---|---|---|---|
| f | 方位词 | nt | 机构团体 | V | 动词 | vd | 副动词 |

英语词性列表（缩减版）

| 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 |
|---|---|---|---|---|---|
| WRB | Wh-副词 | PRP | 人称代词 | DT | 冠词 |
| WP | Wh-代词 | POS | 所有格 | CD | 数词 |
| WDT | Wh-限定词 | PDT | 后接成分 | CC | 连词 |
| W | 标点 | NN | 名词 | | |
| VB | 动词 | MD | 情态词 | | |
| UH | 语气词 | LS | 名语素 | | |
| TO | to | JJ | 形容词 | | |
| SYM | 符号（#$） | IN | 介词 | | |
| RP | 叹词 | FW | 外来语 | | |
| RB | 副词 | EX | there | | |

英语词性列表（完整版）

| 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 |
|---|---|---|---|---|---|---|---|
| CC | 并列连接词 | MD | 情态动词 | RBR | 副词，比较级 | VBP | 动词，非第三人称单数现在式 |
| CD | 基数 | NN | 名词，可数或不可数 | RBS | 副词，最高级 | VBZ | 动词，第三人称单数现在式 |
| DT | 限定词 | NNS | 名词，复数 | RP | 小品词 | WDT | wh-限定词 |
| EX | 存在型there | NNP | 专有名词，单数 | SYM | 符号（数学或科学） | WP | wh-代词 |
| FW | 外文单词 | NNPS | 专有名词，复数 | TO | to | WP$ | 所有格wh-代词 |

（待续）

（续表）

| 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 | 词性编码 | 词性名称 |
|---|---|---|---|---|---|---|---|
| IN | 介词 | PDT | 前位限定词 | UH | 感叹词 | WRB | wh-副词 |
| JJ | 形容词 | POS | 所有格结束词 | VB | 动词，基本形态 | # | # 符号 |
| JJR | 形容词，比较级 | PRP | 人称代名词 | VBD | 动词，过去式 | $ | 美元符号 |
| JJS | 形容词，最高级 | PP$ | 物主代词，所有格代名词 | VBG | 动词，动名词/现在分词 | . | 句点 |
| LS | 列表项标记 | RB | 副词 | VBN | 动词，过去分词 | , | 逗号 |
| : | 冒号，分号 | ) | 右括号 | ‘ | 左单引号 | ’ | 右单引号 |
| ( | 左括号 | “ | 双引号 | “ | 左双引号 | ” | 右双引号 |

**通讯地址：**100083 北京市北京语言大学大数据与语言技术研究所

# The construction of the BCC Corpus in the age of Big Data

.....*XUN Endong, RAO Gaoqi, XIAO Xiaoyue & ZANG Jiaojiao* (93)

Beijing Language and Culture University Corpus Center (BLCU Corpus Center, BCC) Corpus is a large full-text retrieval corpus with multiple languages, including Chinese and other languages as well. BCC is an online data system with a size of about ten billion words, ideal as a data source for studies in linguistics as well as applied linguistics. BCC search queries support wildcards, splittable words, as well as character-based, word-based and POS-tag based expressions. This paper introduces the BCC Corpus in detail, including the construction of the corpus and the design of the search engine, with a particular focus on the query language and tips for corpus search.