文心语料库检索平台的研制
—DCC 2.0 动态流通语料库

朱君辉[1,2,3] 刘鑫[1,2,3] 杨麟儿[1,2,3] 师佳璐[1,2,3] 刘鹏远[1,2] 杨尔弘[1,3]

[1]国家语言资源监测与研究平面媒体中心
[2]北京语言大学 信息科学学院
[3]北京语言大学 语言资源高精尖创新中心

2022年11月19日

ICALL 智能语言学习

研究背景

- 贡献：
  - 语料规模大
  - 语料选材广
  - 多领域语料
  - 历时大数据
- 局限性：
  - 缺乏对深层的句法结构检索需求的支持
  - 检索功能的全面性与用户友好性难以兼顾

研究背景

北京大学CCL语料库

国家语委现代汉语平衡语料库

北京语言大学BCC语料库

DCC 1.0 动态流通语料库

⋮

- 贡献：
  - 语料规模大
  - 语料选材广
  - 多领域语料
  - 历时大数据

DCC 2.0 动态流通语料库

✓ 功能特色：

✓ 支持远距离、深层次的句法结构的检索需求

✓ 不同检索模式兼顾检索功能的全面性与用户友好性

DCC 1.0 动态流通语料库

词义演化 新闻历时 小说历时 论文历时 多领域历时

改革

报纸频率

报纸频度

分类频率

分类频度

中国青年报 今晚报 北京晚报 北京青年报 华西都市报 南方

广州日报 深圳特区报 羊城晚报 钱江晚报 全部报纸

DCC 2.0 动态流通语料库

- 全领域、多覆盖的语料资源
- 语料深加工与多级标注
- 强大的检索功能

DCC 2.0 动态流通语料库

（一）全领域、多覆盖的语料资源

- 目前已入库新闻报刊与二语教材语料两种领域的语料。
  - 新闻报刊：主要来自DCC 1.0 动态流通语料库，时间跨度为1946年至2020年，长达70余年。
  - 二语教材：收录500多册市面上广泛使用的国际汉语教材，包括《博雅汉语》《成功之路》等。

语料类型	字节数/字	句数/句
报刊语料	1,600,000,000	39,500,000
二语教材语料	5,376,330	243,071

- 丰富的元信息：所在文档ID、所在文档标题、所在文档来源、所在文档日期等。
- 文学、微博、口语、学术论文将陆续上线……

DCC 2.0 动态流通语料库

（二）语料深加工与多级标注

词性标注

+

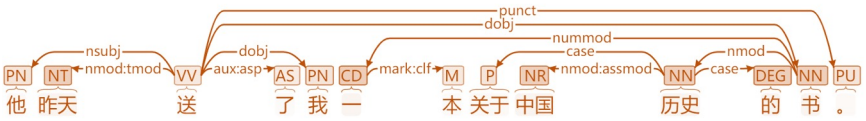
命名实体识别

+

依存句法标注

+

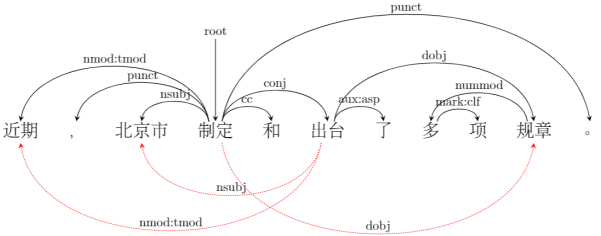
词汇等级标注

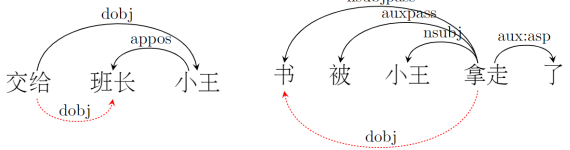


DCC 2.0 动态流通语料库

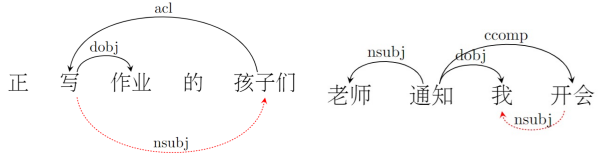
（二）语料深加工与多级标注

- 依存句法标注
  - 普通依存句法标注
  - 加强依存句法标注[1]





通过增加、修改词语之间的依存关系，使得依存关系可以同时体现句子的句法结构和语义关系。

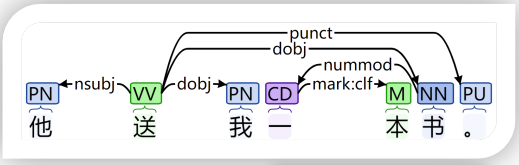


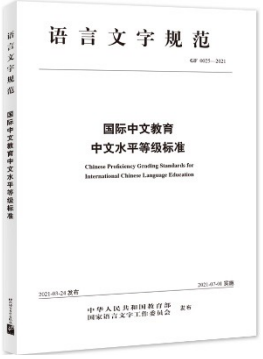
[1] 余婧思, 师佳璐, 杨麟儿, 肖丹, 杨尔弘. 汉语增强依存句法自动转换研究. 第二十一届中国计算语言学大会 (CCL 2022).

DCC 2.0 动态流通语料库

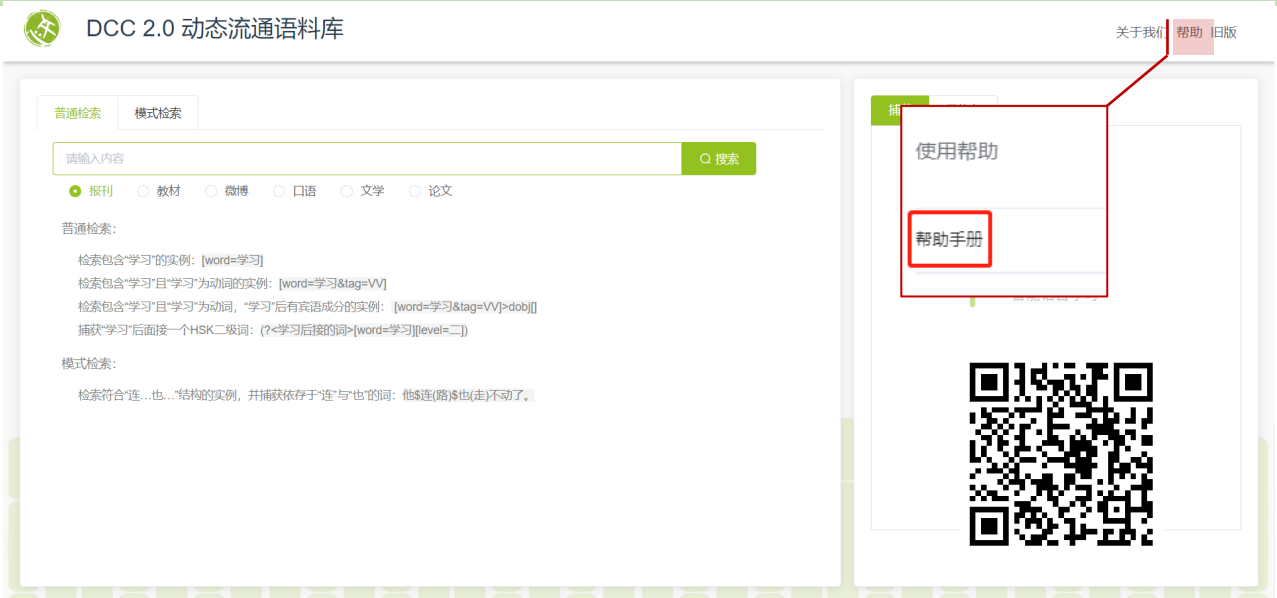
（三）强大的检索功能

- 支持正则表达式；
- 支持复杂的检索表达式查询（如依存句法、命名实体、词汇难度查询等）；
- 支持按出现频次对规定的捕获内容进行统计；
- 支持通过模式检索，查询与给定例句相同句法结构的句子；
- 支持利用元信息继续检索；
- 支持从网页上下载查询结果（txt 文件）。





DCC 2.0 动态流通语料库



© BLCU-ICALL 2022 版权所有

普通检索

- 对单独字词的检索
  - 中文　[word=中文]
  - 学习 中文　[word=学习][word=中文]
- 词性/命名实体约束
  - [word=只&tag=/M|AD/]
  - [entity=ORGANIZATION]
- 通过依存关系约束
  - [word=刻画&tag=VV]>dobj[]
  - [word=袭击] >/nsubj|nsubjpass/ []
- 通过词汇等级约束
  - [word=只有] [level=三] [word=才]
  - [word=及时&tag=/VA|JJ/] <amod [level=三]

普通检索-检索结果


DCC 2.0 动态流通语料库
普通检索
模式检索
[word=刻画&tag=VV]>dobj[]
搜索
报刊
教材
微博
口语
文学
论文
结果下载
客栈外的开阔地上,有一座1米高的石碣,上面竖立着一尊堂吉诃德手执矛盾的铁塑像,传神地刻画出了出征前踌躇满志的神态。
随着大众传媒的迅速发展,女性以瘦为美的概念和图像被迅速传播,逐渐在世界范围被接受,拥有纤瘦身材的女性被刻画为"富有吸引力、自信、快乐和成功"的形象。
人民日报任仲平文章《标注共产党人的精神坐标》,从涤荡"四风"到制度构建,从问题清扫到人心回归,于层层递进中标注出当代共产党人的精神坐标,以严、实、廉、清、敢、党6个字刻画了政治生态的改变。
我们有专门的人物版和多个栏目,既生动描绘个体,也注重刻画新时代奋斗者群像。
其次,要坚持马克思主义哲学精神的指引,坚持刻画好党的形象的理想性格的"质的规定性、表现的丰富性、情致的始终如一性"的辩证统一。

普通检索

- 对单独字词的检索
  - 中文　[word=中文]
  - 学习 中文　[word=学习][word=中文]
- 词性/命名实体约束
  - [word=只&tag=/M|AD/]
  - [entity=ORGANIZATION]
- 通过依存关系约束
  - [word=刻画&tag=VV]>dobj[]
  - [word=袭击] >/nsubj|nsubjpass/ []
- 通过词汇等级约束
  - [word=只有] [level=三] [word=才]
  - [word=及时&tag=/VA|JJ/] <amod [level=三]

捕获

通过捕获可以直观地展示观察内容在语料中的频次。使用检索式(?<name> ...)来捕获和命名。

普通检索-捕获

- 对单独字词的检索
  - 中文　(?<中文>[word=中文])
  - 学习 中文　[word=学习][word=中文]
- 词性/命名实体约束
  - (?<只> [word=只& tag=/M|AD/])
  - (?<机构名> [entity=ORGANIZATION])
- 通过依存关系约束
  - [word=刻画&tag=VV]>dobj(?<dobj>[])
  - [word=袭击] >/nsubj|nsubjpass/ (?<主语>[])
- 通过词汇等级约束
  - [word=只有] (?<三级词> [level=三]) [word=才]
  - [word=及时&tag=/VA|JJ/] <amod(?<三级词>[level=三])

捕获

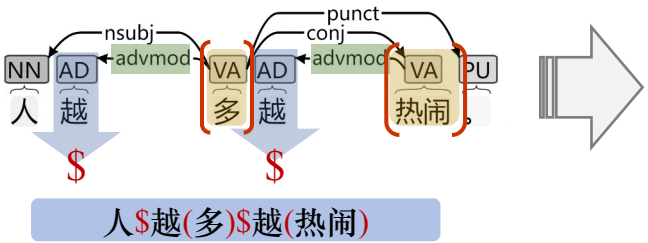
通过捕获可以直观地展示观察内容在语料中的频次。使用检索式(?<name> …)来捕获和命名。

普通检索-捕获结果

- [word=及时&tag=/VA|JJ/] <amod(?<三级词>[level=三])

模式检索

- 锚点：用于精准匹配固定的成分，用“$”表示
- 捕获：待观察的成分捕获为变量，用“( )”将内容括起来
- 未被标记为捕获或锚点的词，则被认为是帮助句子符合语法要求，使句子完整的成分。
- 在输入一个带标记符号的句子后，系统会返回例句的依存句法图呈现给用户，并依据该依存句法图匹配结果。



模式检索

- 连……也……
  - 他$连(路)$也(走)不动了。

模式检索

- 介词“对”后可以跟的谓词性成分
  - 这$对我(有利)

检索结果的过滤与下载

- 点击语料前的 ⓘ 标志，可以查看语料详情，包括语料id、文章类型、题目、日期、来源、文章段落等。
- 在界面右侧的“元信息”下，可对检索结果的元信息进行选择。
- 检索结果页面右上角位置有“结果下载”按钮，用户可指定下载的检索结果条数（默认为 50 条）与文件名，点击“结果下载”按钮，可将查询结果以本文文件（*.txt）格式保存至本地电脑。

DCC 2.0 动态流通语料库的应用

■ 国际中文教育语法点自动抽取[2]

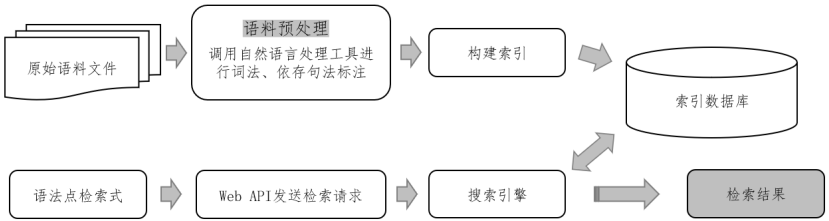
- 对象：《国际中文教育中文水平等级标准》(GF0025 - 2021) 中《附录 （规范性）语法等级大纲》中的语法点。
- 语法点检索式实例

id	级别	语法项目	类别	细目	语法内容	
93	二级	固定格式			还是……吧	[word=还是&tag=AD][]*[word=吧&tag=SP]
94	二级	固定格式			又……又……	[word=又&tag=AD]<advmod[tag=/V.*/][word=又&tag=AD]<advmod[tag=/V.*/]
95	二级	固定格式			（在）……以前/以后/前/后	[word=在]?[word=/以前|以后/&tag=NT] <nmod:tmod [tag=/V.*/]
168	三级	固定格式			除了……（以外），……还/也/都……	[word=除了&tag=P]<case[]>case[word=外|word=以外&tag=LC][word=','}?[word!=','}*[word=/也|还|都/&tag=AD]
169	三级	固定格式			从……起	[word=从&tag=P]<case[] </nmod.*/?[]? >ccomp?[]? >/advmod.*/?[]? >case[word=起&tag=LC]
170	三级	固定格式			对……来说	[word=对&tag=P]<case[]?</nmod.*/?[]?>/advmod.*/?[]?>case[word=来说&tag=LC]
171	三级	固定格式			一……也/都+不/没……	[word=/一|一点|一点儿/]<nummod[]<nsubj[][]>advmod[word=也|word=都]
172	三级	固定格式			越……越……	[word=越&tag=AD]<advmod[]>>{1,4}[]>advmod[word=越&tag=AD]
240	四级	固定格式			一+量词+比+一+量词	
241	四级	固定格式			（自）……以来	[word=自&tag=P]? <case [] >case [word=以来&tag=LC]
242	四级	固定格式			由……组成	[word=由&tag=P]<case[]<nmod:prep[word=组成&tag=VV]
243	四级	固定格式			在……方面	[word=在&tag=P]<case[word=方面&tag=NN]
244	四级	固定格式			在……上/下/中	[word=在&tag=P]<case[]*>case[word=/上|中|下/&tag=LC]
313	五级	固定格式			从……来看	[word=从&tag=P]<case[]>case[word=来看&tag=LC]
314	五级	固定格式			到……为止	[word=到&tag=P]<case[]>case[word=为止&tag=LC] [word=到此为止]
315	五级	固定格式			够……的	[word!=不][word=够&tag=VV][tag!=PU]*[word=的][tag=PU]
316	五级	固定格式			拿……来说	[word=拿&tag=P]<case[][word=来&tag=MSP]<aux:prtmod[word=说&tag=VV]

[2] 朱君辉,刘鑫,杨麟儿,王鸿滨,杨尔弘.汉语语法点特征及其在二语文本难度自动分级研究中的应用[J].语言文字应用,2022(03):87-99.

检索式的应用

■ 语法点抽取流程



■ 语法点在文本中的抽取结果实例

文本

北京有座美丽的中山公园，公园里有个用五色土砌成的社稷坛。

（《成功之路．成功篇》第2册《古迹今日》首句）

语法点

一级语法点及频数：“结构助词”的 (2 次)，方位名词“里”(1 次)，量词“个”(2 次)，“有字句”(2 次)，陈述句(1 次)；

二级语法点及频数：偏正短语 (3 次)，名词性短语 (3 次)，方位短语 (1 次)。

[2] 朱君辉, 刘鑫, 杨麟儿, 王鸿滨, 杨尔弘. 汉语语法点特征及其在二语文本难度自动分级研究中的应用[J]. 语言文字应用, 2022(03):87-99.

参考文献

[1] 郭慧志, 王强军, 刘华, 张普. 大规模动态流通语料库的构建. 第二届全国学生计算语言学研讨会论文集. 2004

[2] 靳光瑾, 肖航, 富丽, 章云帆. 现代汉语语料库建设及深加工. 语言文字应用, 2005,(2).

[3] 李正华. 汉语依存句法分析关键技术研究. 哈尔滨工业大学, 2013.

[4] 谭晓平. 现代汉语文本语料库建设及应用现状研究. 对外汉语研究, 2018,(2).

[5] 荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 大数据背景下BCC语料库的研制. 语料库语言学, 2016,(1).

[6] 詹卫东. 近30年来中文语言知识资源发展及应用. 语言战略研究, 2018,(4).

[7] 詹卫东. 北京大学CCL语料库的研制[J]. 语料库语言学, 2019.

[8] 张永伟, 刘婷, 刘畅, 吴冰欣, 俞敬松. 融合句法信息的文本语料库检索方法研究. 数据分析与知识发现,2022

[9] De Marnee M C, Manning C D. Stanford typed dependencies manual, 2010.

[10] Kilgarriff A, Baisa V, Bušta J. The Sketch Engine: ten years on, 2014,(1).

[11] 余婧思, 师佳璐, 杨麟儿, 肖丹, 杨尔弘. 汉语增强依存句法自动转换研究. 第二十一届中国计算语言学大会 (CCL 2022).

[12] 朱君辉,刘鑫,杨麟儿,王鸿滨,杨尔弘.汉语语法点特征及其在二语文本难度自动分级研究中的应用. 语言文字应用, 2022(03):87-99.

欢迎大家使用和反馈问题！

欢迎大家使用和反馈问题！


北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY


国家语言资源监测与研究中心
语言资源
北京语言大学
全面媒体语言分中心

请大家批评指正！