



# 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<sup>1,2,3</sup> 刘鑫<sup>1,2,3</sup> 杨麟儿<sup>1,2,3</sup> 师佳璐<sup>1,2,3</sup> 刘鹏远<sup>1,2</sup> 杨尔弘<sup>1,3</sup>

<sup>1</sup>国家语言资源监测与研究平面媒体中心

<sup>2</sup>北京语言大学 信息科学学院

<sup>3</sup>北京语言大学 语言资源高精尖创新中心

2022年11月19日

# 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

## □ 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

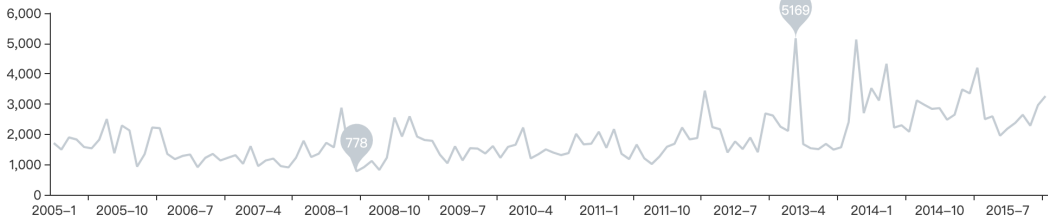
改革



○ 中国青年报 ○ 今晚报 ○ 北京晚报 ○ 北京青年报 ○ 华西都市报 ○ 南方都市报 ○ 广州日报 ○ 深圳特区报 ○ 羊城晚报 ○ 钱江晚报 ○ 全部报纸

**报纸频率关系**

来源：国家语言资源监测与研究平面媒体中心



报纸频率

报纸频度

分类频率

分类频度



# DCC 2.0 动态流通语料库

DCC 1.0 动态流通语料库



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

PN NT VV AS PN CD M P NR NN DEG NN PU  
他 昨天 送 了 我 一 本 关于 中国 历史 的 书 。

命名实体识别

XXXX-XX-XX      NUMBER      COUNTRY  
DATE                      1                      中国  
他 昨天 送了我 一 本关于 中国 历史的书。

依存句法标注

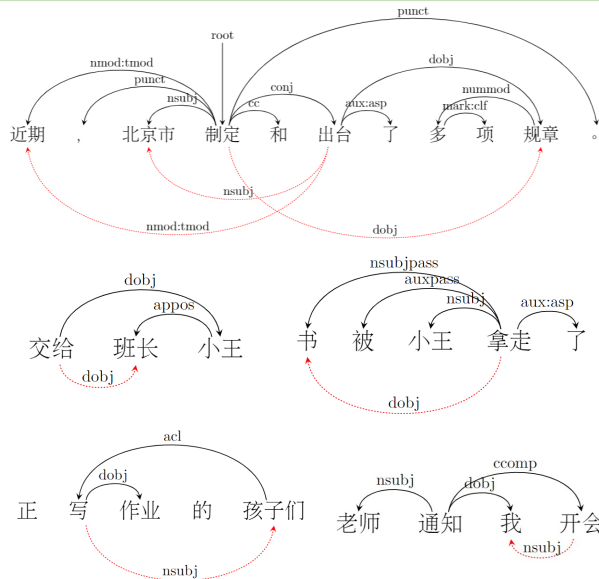
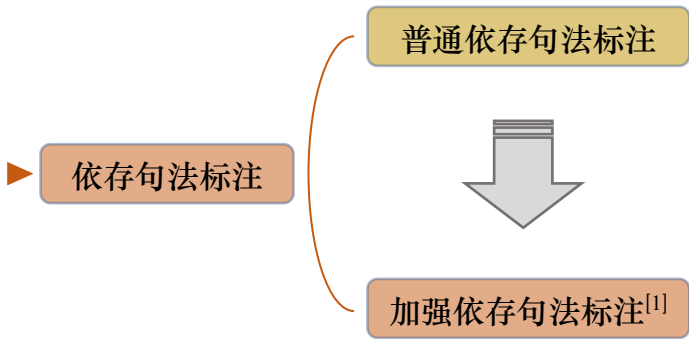
nsbj      dobj      nummod      punct      nmod  
nmod:tmod      aux:asp      mark:clf      case      case  
PN NT VV AS PN CD M P NR NN DEG NN PU  
他 昨天 送 了 我 一 本 关于 中国 历史 的 书 。

词汇等级标注

— — — — — — — 四 — 四 — —  
他 昨天 送 了 我 一 本 关于 中国 历史 的 书

# DCC 2.0 动态流通语料库

## (二) 语料深加工与多级标注



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。

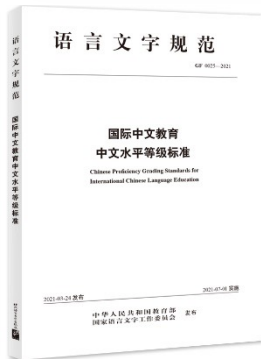
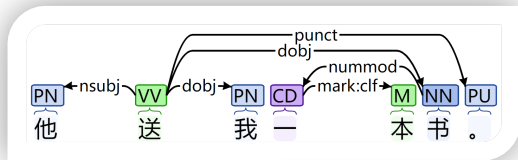
[1] 余婧思, 师佳璐, 杨麟儿, 肖丹, 杨尔弘. 汉语增强依存句法自动转换研究. 第二十一届中国计算语言学大会 (CCL 2022).



# 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）。



The screenshot shows the search results interface for the query '学习后接的词' (Words following '学习'). The interface includes a search bar with the query, a dropdown menu for '学习后接的词', and a list of results with their respective frequencies:

- (94) 学习教育
- (12) 学习生活
- (9) 学习讨论
- (6) 学习国际
- (5) 学习成绩
- (5) 学习活动

# DCC 2.0 动态流通语料库



普通检索

模式检索

请输入内容

搜索

报刊  教材  微博  口语  文学  论文

普通检索:

检索包含“学习”的实例: [word=学习]

检索包含“学习”且“学习”为动词的实例: [word=学习&tag=VV]

检索包含“学习”且“学习”为动词,“学习”后有宾语成分的实例: [word=学习&tag=VV]>dobj]

捕获“学习”后面接一个HSK二级词: (?<学习后接的词>[word=学习][level=二])

模式检索:

检索符合“连...也...”结构的实例,并捕获依存于“连”与“也”的词: 他\$连(路)\$也(走)不动了。

捕

使用帮助

帮助手册



# 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

- ① 客栈外的开阔地上,有一座1米高的石墩,上面竖立着一尊堂吉珂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**。
- ① 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**。
- ① 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**。
- ① 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**。
- ① 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])



DCC 2.0 动态流通语料库

关于我们 帮助 旧版

普通检索 模式检索

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

Q 搜索

● 报刊 ○ 教材 ○ 微博 ○ 口语 ○ 文学 ○ 论文

结果下载

① 司空见惯的新闻客户端的弹窗技术也是如此,只不过人们熟悉的是新闻事件的**及时传播**,甚至有时略有些信息过载,但当这种技术用到了刀刃上,就立刻发挥出传统模式所不具备的作用。**三级词**

① 乌干达总理鲁贡达表示,乌干达政府感谢中国政府**及时**的人道主义**支持**,称这是非中友谊的象征。**三级词**

① 乌干达总理鲁贡达表示,乌干达政府感谢中国政府**及时**的人道主义**支持**,称这是非中友谊的象征。**三级词**

① 现如今,气象人则追求更准确、更精细,只为**及时预报**,贴心服务。**三级词**

① 根据该方案确定的目标,到2022年,首都将健全完善方法科学、实施有效、**更新及时**的**标准**制定修订工作机制,基本建成体现北京特色、指标水平先进、系统构成完善的节能低碳和循环经济标准体系,努力打造全国节能低碳和循环经济标准创新中心、示范基地和辐射之源。**三级词**

① 袁慧琴建议,加大戏曲优秀领军人物、各行当优秀演员的新剧目创作力度,实施“当代戏曲名家优秀代表作推广工程”,让优秀演员演出的传统戏、新创戏得到**及时记录**和广泛推广。**三级词**

共有160条结果

捕获 元信息

三级词 --None--

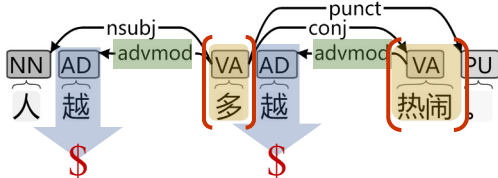
**三级词 (63 samples)**

- (17) 处理
- (14) 支持
- (12) 支付
- (7) 情况
- (7) 解决
- (7) 调整
- (6) 指导
- (5) 应用
- (4) 保护
- (4) 推广
- (4) 能力
- (4) 人员



# 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。



人\$越(多)\$越(热闹)

- 
- ① 邻居们越听越糊涂。  
多 热闹
  - ① 我问问题,你回答,回答得越快越好。  
多 热闹
  - ① 爷爷一口气背到了底,我静静地听,越听越佩服,不禁喊道:“爷爷真棒!”  
多 热闹
  - ① 据说是越叫得随意,越不易丢失。  
多 热闹
  - ① 元宵是一点一点滚出来的,越滚越大,里面是馅,外面是江米粉。  
多 热闹



# 模式检索

## ■ 连……也……

- 他\$连(路)\$也(走)不动了。



DCC 2.0 动态流通语料库

关于我们 帮助 旧版

普通检索 模式检索

他\$连(路)\$也(走)不动了。

报刊  教材  微博  口语  文学  论文

检索式 (?<也>[word=也]) <admod (?<走>[ ]>dep (?<路>[ ]>admod (?<连>[word=连])

① ……连看电视也把音量调得很小但是客厅里的椅子不听话碰到地板想了想,一,每次挪动的时候。  
连 走

② 当时11,爸爸出差2在外地,回到家后连饭也没吃,就去了医院。  
连 走

③ 扶西一下课就急急忙忙跑出去,跟同学们玩儿,连上厕所小便也忘了。  
连 走

④ 怎么今天风向变了,连你也向看他。  
连 走

⑤ 你怎么连这么简单的问题也不知道?  
连 走

⑥ 我虽然想来听课可是现在这个座位的旁边没有空位,只好坐在点儿的旁边,把个座位让给去听课。

结果 元信息

走

走 (23 samples)

- (7) 没有
- (5) 看
- (5) 说
- (3) 吃
- (2) 保留
- (2) 回
- (2) 小
- (2) 是
- (2) 选择
- (1) 买
- (1) 他
- (1) 口


© BLCU·ICALL, 2022 版权所有



# 模式检索

## ■ 介词“对”后可以跟的谓词性成分

### • 这\$对我(有利)

 DCC 2.0 动态流通语料库

[关于我们](#) [帮助](#) [旧版](#)



普通检索 模式检索

这\$对我(有利)

报刊  教材  微博  口语  文学  论文



检索式: (?<有利-[i]) >nmod prep (?<我-[i]) >case (?<对-[word=对])

- ① #一件好事星期天早上,姐姐**对我**说:“今天我们来做一些有意义的事情吧。”  
**有利**
- ② 另一个人**对他**说:“你把运动鞋穿上有什么用?”  
**有利**
- ③ **对**羚羊来说,如果它跑不过最快的狮子的话,它就会被吃掉。  
**有利**
- ④ 你**对**历史也有兴趣?  
**有利**
- ⑤ 您**对**中国历史一定很有研究呢。  
**有利**
- ⑥ 今天参观这个乾隆博物馆**对**您研究唐代历史一定会有很大帮助。

捕获 元信息

有利

有利 (118 samples)

- (38) 说
- (15) 有
- (12) 进行
- (6) 充满
- (3) 没有
- (3) 重要
- (2) 习惯
- (2) 关心
- (2) 加以
- (2) 好
- (2) 干涉
- (2) 瞎

© BLCU-ICALL, 2022 版权所有



# 检索结果的过滤与下载

- 点击语料前的 **i** 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。

语料详情

所在段落

西安是很值得参观的，你对历史也有兴趣？我也是搞历史的。

语料元信息

类型	
doc_type	textbook
textbook	今日汉语(一)(第二版)
title	31.我们是同行

捕获

元信息

date

date (825 samples)

- (4) 20091021
- (4) 20101220
- (3) 19920720
- (3) 19970416
- (3) 19970531
- (3) 19980213
- (3) 20001022



# 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<sup>[2]</sup>

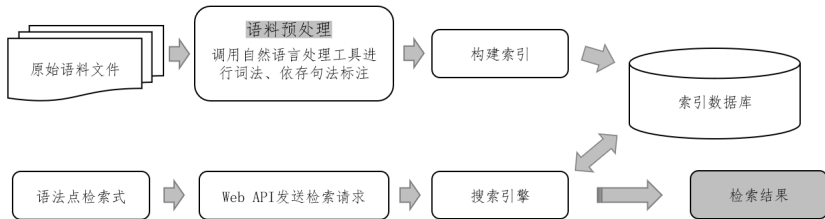
- 对象：《国际中文教育中文水平等级标准》(GF0025 - 2021) 中《附录（规范性）语法等级大纲》中的语法点。
- 语法点检索式实例

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）。

# 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!



普通检索 模式检索

[word=学习] 搜索

报刊  教材  微博  口语  文学  论文

结果下载

请您谈谈如何深入学习领会习近平总书记重要讲话精神,推进全面依法治国各项工作任务落地落实。

下一步,围绕学习贯彻习近平总书记重要讲话精神,主要抓好三个方面工作。

一是认真组织各地区各部门深入学习宣传贯彻习近平总书记重要讲话精神,通过专题学习会、报告会、党委(党组)理论学习中心组学习会等多种形式,深刻领会精神实质,把握核心要义,切实把思想和行动统一到党中央决策部署上来,增强做好全面依法治国工作的思想自觉和行动自觉。

一是认真组织各地区各部门深入学习宣传贯彻习近平总书记重要讲话精神,通过专题学习会、报告会、党委(党组)理论学习中心组学习会等多种形式,深刻领会精神实质,把握核心要义,切实把思想和行动统一到党中央决策部署上来,增强做好全面依法治国工作的思想自觉和行动自觉。

一位学生家长抱怨,在拥挤的教室里,学习受影响、安全有隐患、成长烦恼多……2015年下半年,山东提出力争用两年时间,推进解决城镇普通中小学大班额问题。

刘杰说,如今,没有了大班额,有了老师的督促,孩子的学习积极性也提高了不少。

解决大班额问题,让中小学生摆脱逼仄的学习空间,还他们一个宽松的学习环境,是加快推进新型城镇化的

共有200条结果

捕获 元信息

# 欢迎大家使用和反馈问题!



普通检索 模式检索

她\$不但(漂亮)\$而且(聪明)

报刊  教材  微博  口语  文学  论文

1 她 不但 漂亮 而且 聪明

格式式 (?<不但>[word=不但]) <advmod (?<漂亮>[ ]) >conj (?<聪明>[ ]) >advmod (?<而且>[word=而且])

1 这里不但非常干净,而且非常有趣。  
漂亮 聪明

1 人们都说中国的真丝制品和羊毛制品不但便宜,而且质量也不错,我想看看这类东西。  
漂亮 聪明

1 特别是这个炒鸡丁,不但好吃,而且样子也很好看。  
漂亮 聪明

1 不但好吃,而且好看。  
漂亮 聪明

1 那里不但非常干净,而且非常漂亮。  
漂亮 聪明

1 电影的故事情节是这样的:母亲年轻时是村子里有名的美女,不但善良,而且勇敢。

捕获 元信息

漂亮 --None--

漂亮 (105 samples)

- (1) 甘美
- (1) 简单
- (1) 繁多
- (1) 结
- (1) 绘制
- (1) 给
- (1) 耐穿
- (1) 舒服
- (1) 节日
- (1) 表现
- (1) 要命







请大家批评指正！