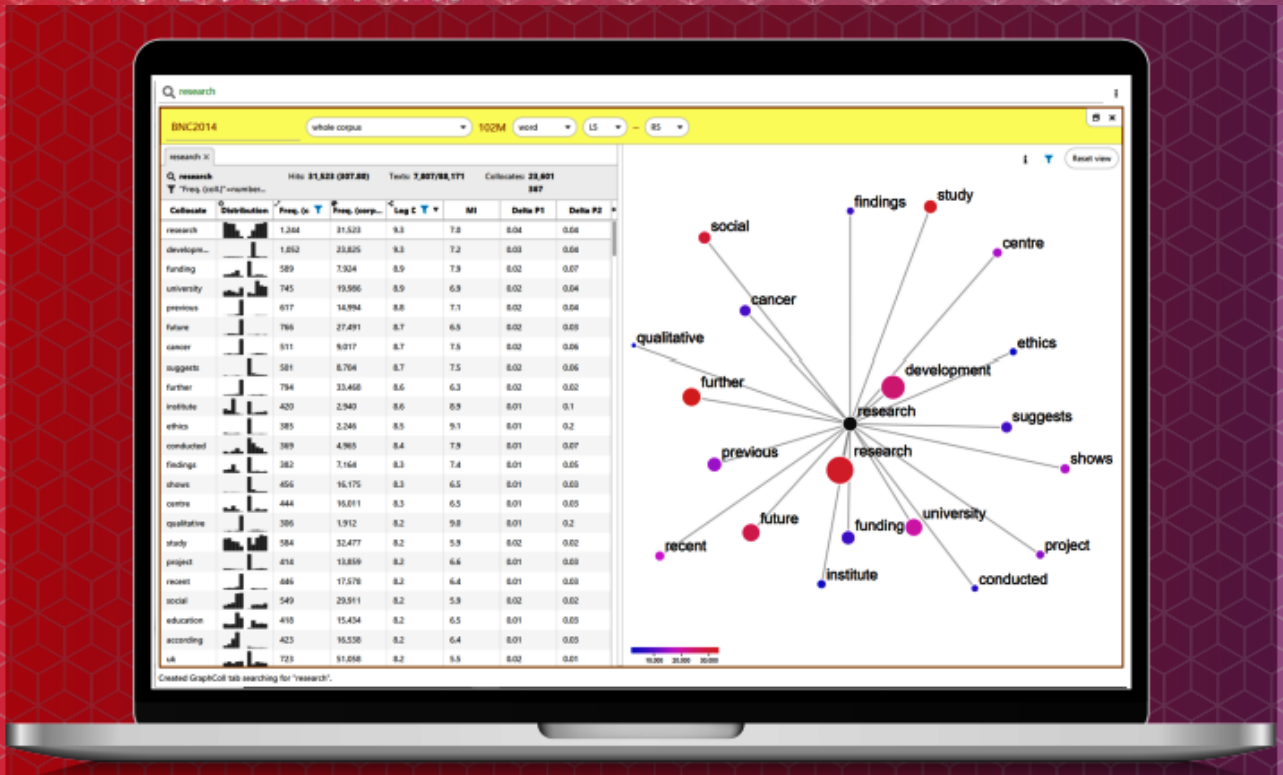
#LancsBox

创新驱动的
语料库语言学软件

#LancsBox X 是一款强大的语言分析工具：可处理亿万级别的单词。

即享免费下载。



Brezina, V., Platt, W. (2023). #LancsBox X 2.0 [software package], lancsbox.lancaster.ac.uk

The development of #LancsBox was supported by the Economic and Social Research Council (grant number EP/P001559/1, ES/K002155/1 and ES/R008906/1)

目录


Lancaster
University
Developed @ Lancaster University

#LancsBox X：许可协议

#LancsBox 采用 BY-NC-ND 知识共享许可协议，可供非商业用途免费使用。完整的许可协议可从以下网址获取：http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

1 下载和运行 #LancsBox X

#LancsBox 是款新一代的语料库分析工具。该工具的 X 版本针对 64 位操作系统（Windows 64 位、Mac 和 Linux）进行了优化设计，以实现最佳性能。

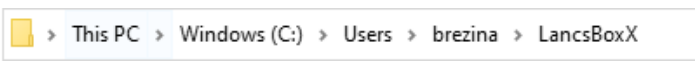
❶ 选择并下载：选择适合您操作系统的版本，并将安装程序下载到您的计算机。

或者直接点击此按钮进行下载：

❷ 运行安装程序

#LancsBox 安全可靠。您只需双击安装程序文件，并按照安装程序中的步骤进行操作即可。请务必将#LancsBox 安装到具有读写权限的文件夹，例如 Users 文件夹（默认）或桌面。对于 Windows 系统，请避免将#LancsBox 安装到 Program Files 文件夹中。

完成常规安装后，#LancsBox 将被安装在以下位置：

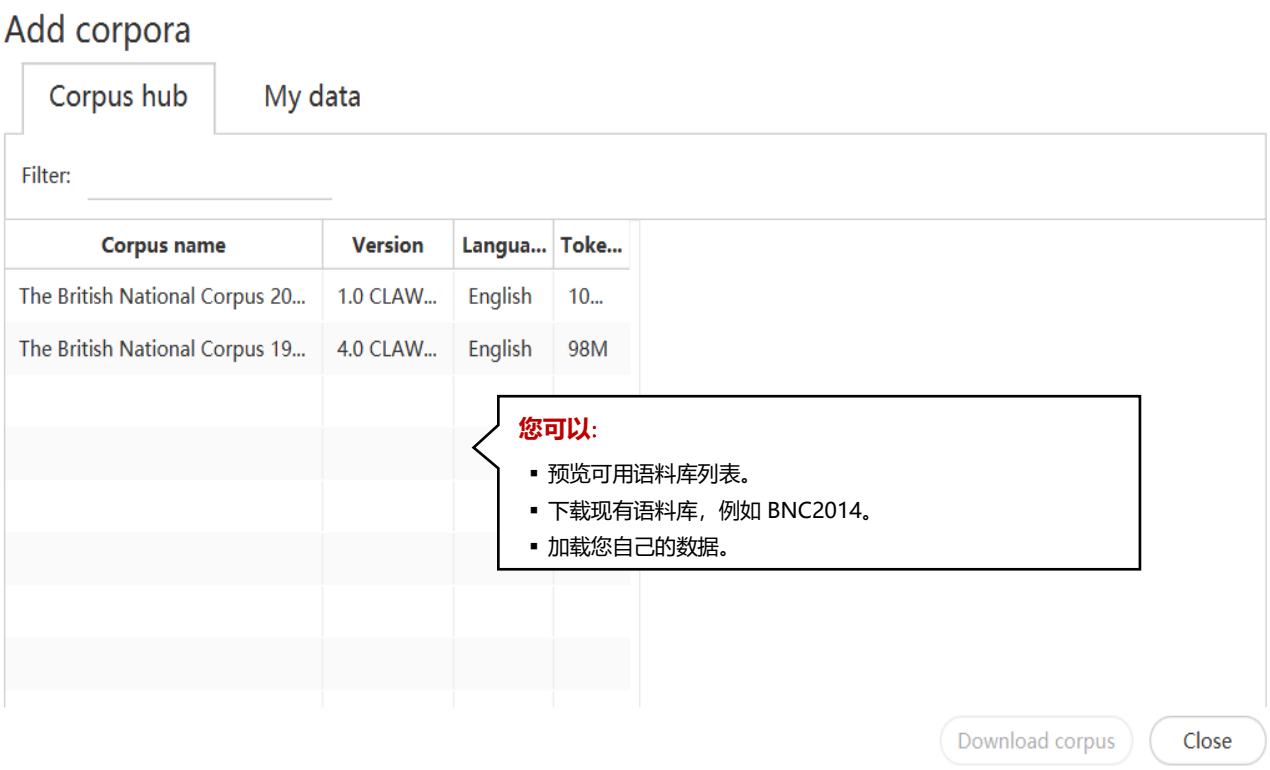
Windows 

Mac *Macintosh HD>Users>*username*>*# LancsBox X*

2 导入数据

#LancsBox X 为处理大型语料库而设计。该工具支持 XML 格式，使得使用丰富的元数据成为可能。数据可以轻松加载并导入到#LancsBox 中。

2.1 导入数据的视觉概述



提示：您可以使用键盘快捷键 Ctrl - 和 Ctrl + （Mac 上的 Cmd-和 Cmd +）来调整缩放程度。

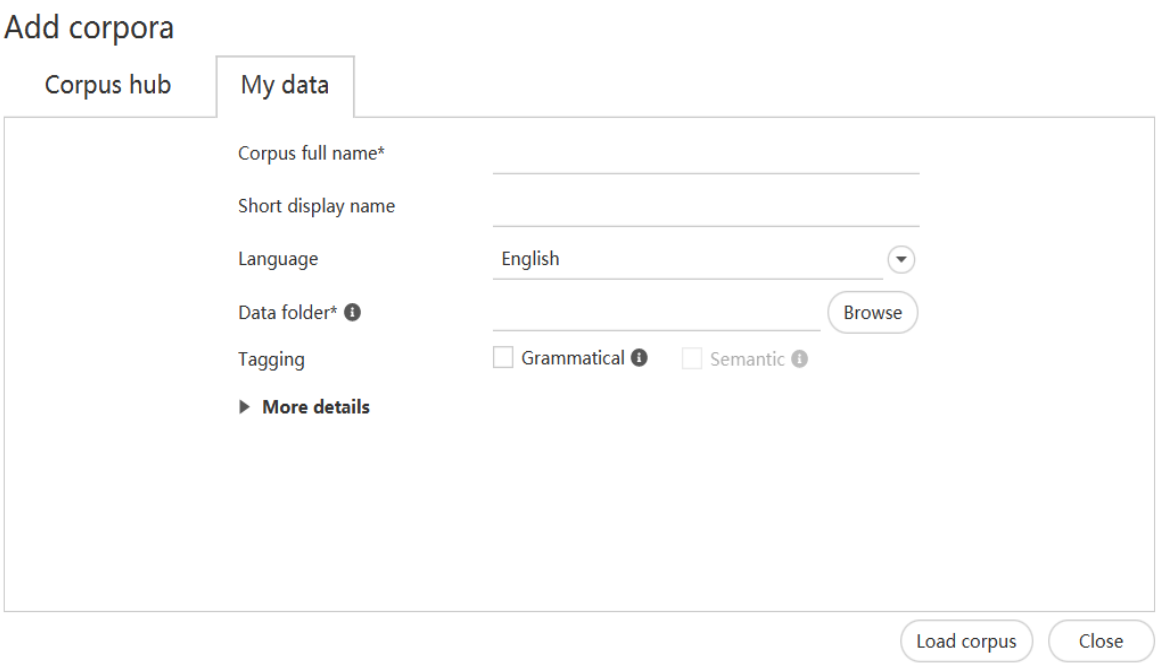
2.2 加载语料库

#LancsBox 允许您使用自己的语料库，并支持多种文件格式（如 txt、docx、pdf、pptx、xlsx 等）以及 XML 。

.txt	XML with w elements
We can pick up on the last comment. Once we are in the grip of reflective thinking it is very hard, if not impossible, for us to see our ethical justifications of our ethical concepts, say, in a genuine way: we will always be drawn to the thought that this is all local. In addition, we will no longer see such judgements as embodying any sort of knowledge.	<?xml version="1.0" encoding="utf-8"?> <text id="AcaHumBk20" mode="writing" genre="academic prose" subgenre="academic prose: humanities" subsubgenre="academic prose: humanities: NA" publication="book" section="NA" sample="end" source="NA" author="NA" pubDate="NA" words="6635"> <p n="1"><s n="1"><w pos="PPIS2" hw="we" class="PRON" usas="Z8">We</w> <w pos="VM" hw="can" class="VERB" usas="A7">can</w> <w pos="VVI" hw="pick" class="VERB" usas="M2">pick</w> <w pos="RP" hw="up" class="ADV" usas="M2">up</w> <w pos="II" hw="on" class="PREP" usas="Z5">on</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="MD" hw="last" class="ADJ" usas="N4">last</w> <w pos="NN1" hw="comment" class="SUBST" usas="Q2:1">comment</w><c>.</c></s> <s n="2"><w pos="CS" hw="once" class="CONJ" usas="Z5">Once</w> <w pos="PPIS2" hw="we" class="PRON" usas="Z8">we</w> <w pos="VBR" hw="be" class="VERB" usas="A3">are</w> <w pos="II" hw="in" class="PREP" usas="Z5">in</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="NN1" hw="grip" class="SUBST" usas="A1:1:1">grip</w> <w pos="IO" hw="of" class="PREP" usas="Z5">of</w> <w pos="JJ" hw="reflective" class="ADJ" usas="X2:1">reflective</w> <w pos="NN1" hw="thinking" class="SUBST" usas="X2:1">thinking</w> <w pos="PPH1" hw="it" class="PRON" usas="Z8">it</w> <w pos="VBZ" hw="be" class="VERB" usas="A3">is</w> <w pos="RG" hw="very" class="ADV" usas="A13:3">very</w> <w pos="JJ" hw="hard" class=

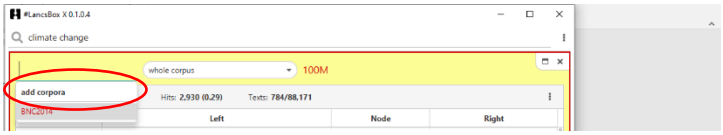
1. 准备好您的数据，并将其存放在一个文件夹中。

2. 在“My data（我的数据）”选项卡中填写语料库信息，然后单击“Browse（浏览）”以选中目标文件夹。



3. 单击右下角的“Load corpus（加载语料库）”按钮。

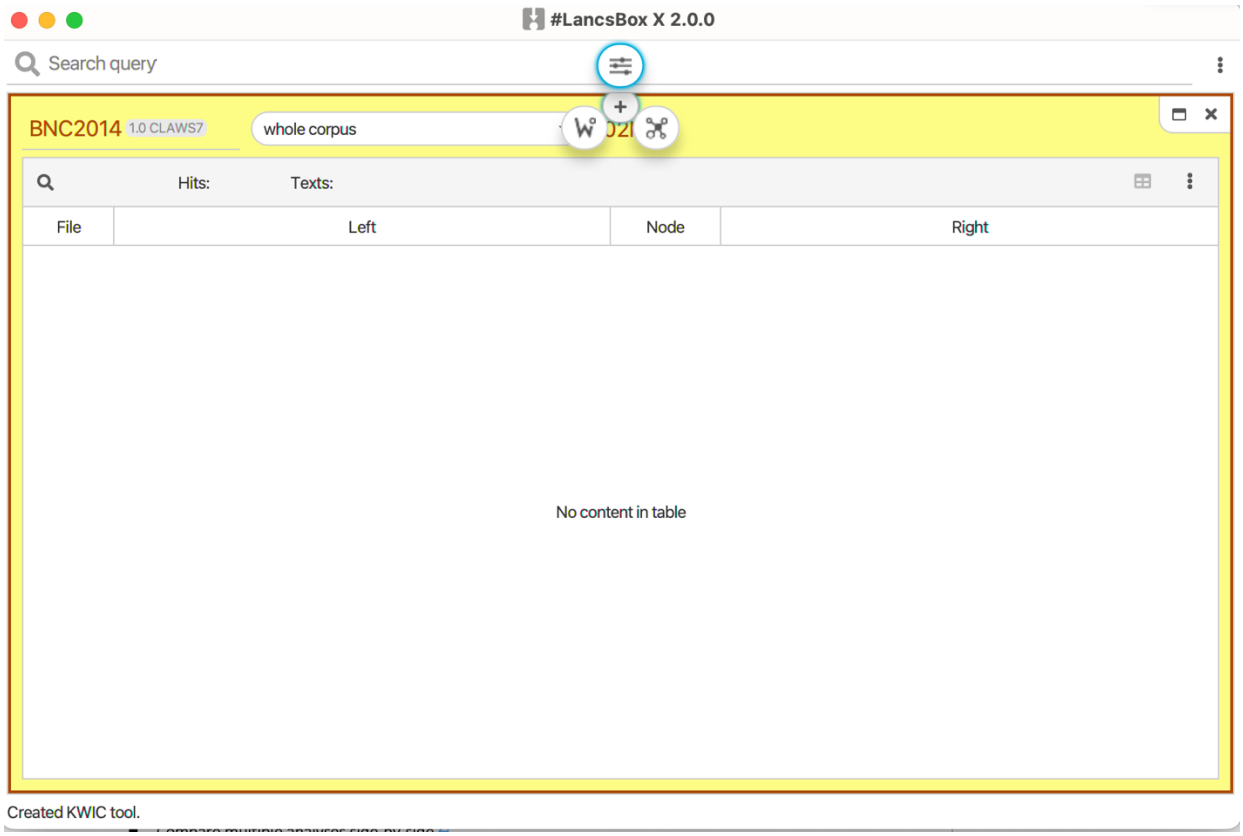
您可以通过单击语料库名称并从下拉菜单中选择“add corpora ”来添加更多的语料库。



提示：#LancsBox X 版本支持对多种语言的数据（语料库）进行自动标注。目前，可提供词性（POS）、句法（依存）和语义标注。可通过勾选相应的选项来开启此功能——请选择“grammatical（语法）”选项进行词性、句法标注和词形还原，选择“semantic（语义）”选项通过 USAS 赋码集进行语义标注。

2.3 #LancsBox 的三个主要功能

#LancsBox 的三大核心语料库分析工具包括 KWIC、GraphColl、以及 Words tool。当您将鼠标指向界面中的加号图标，三个工具的图标将会浮现出来：KWIC 位于顶部，GraphColl 位于右侧，Words tool 位于左侧。在接下来的内容中，我们将分别为您详细解读这三个工具的功能特性。



3 KWIC 工具（语境中的关键词）

KWIC 工具用于生成检索项在一个语料库中所有实例的列表，并以索引的形式呈现。它可以用于以下方面：

- 查找一个词或短语在语料库中的出现频率。
- 查找不同词类（如名词、动词、形容词）的出现频率。
- 使用"smart searches（智能搜索）"查找复杂的语言结构，如被动语态，分裂不定式。
- 对检索结果进行排序。
- 并排比较多个分析结果。

3.1 KWIC：概述

下图展示了 KWIC 工具简单且高效的设计。它提供多种强大的搜索功能，可在单个搜索框中执行。

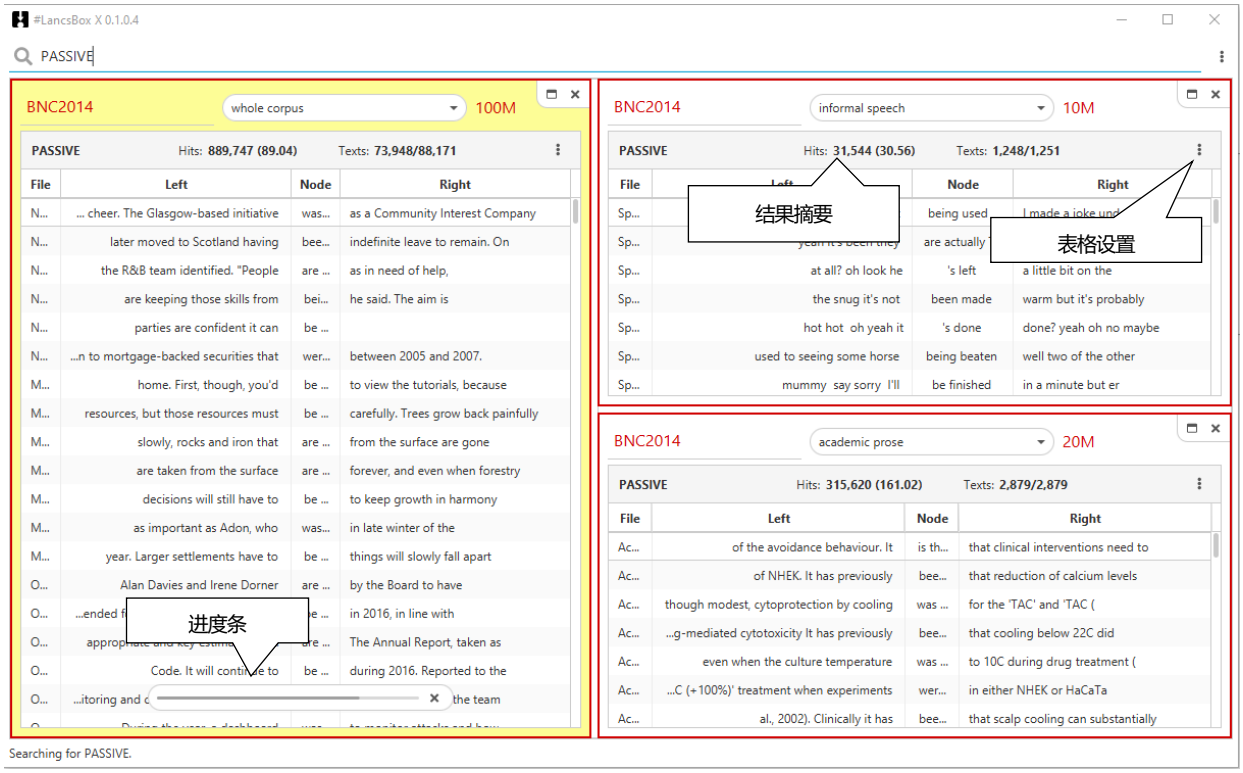
单击表格中的某一行即可选中该行。按住 Ctrl 或 Cmd 键并单击即可选择多行。选定的行可以通过使用 Ctrl+C / Cmd+C 键盘快捷键或右键单击表格并选择"Copy（复制）"选项来复制。

检索结果也可以轻松地从主菜单保存。您可以选择“Save（保存）“ 以保存活动面板（已突出显示）， 或选择“Save all（全部保存）“ 以保存所有面板。

3.2 多面板

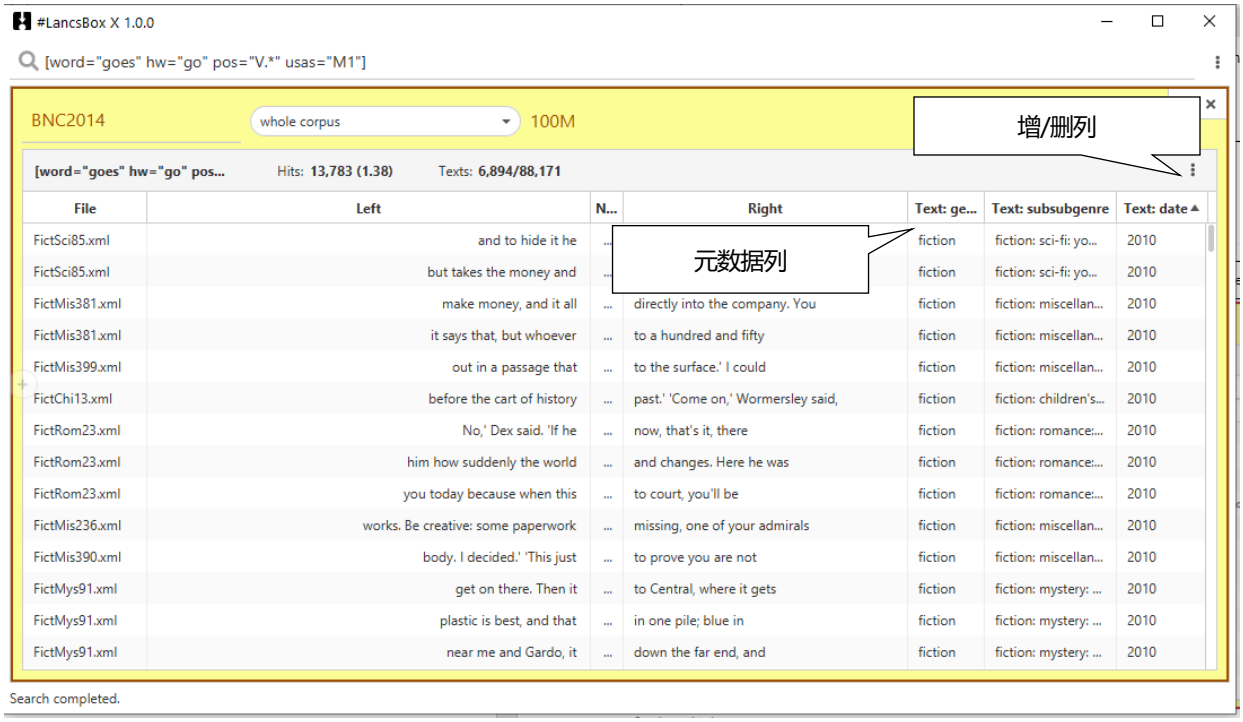
#LancsBox X 支持多面板分析。您可以通过点击并拖动窗口顶部的区域来重新排列面板。

您可以通过按住 Ctrl 或 Cmd 键并单击工具来选择多个面板。这可用于同时在多个面板中执行相同的搜索。



3.3 元数据列

#LancsBox X 致力于高效处理元数据。 KWIC 的列表可显示不同类型的元数据。您可以根据需要添加列，并通过排序和过滤来显示所需信息。如果您需要在表格中添加或删除列，请单击表格设置菜单（ ）并从“Columns（列）”子菜单中选择相关项目。

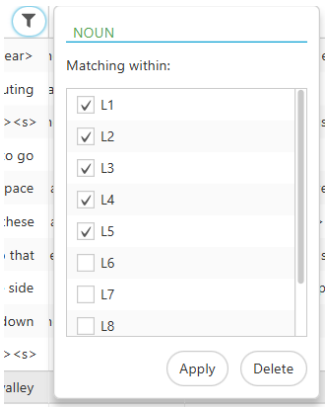


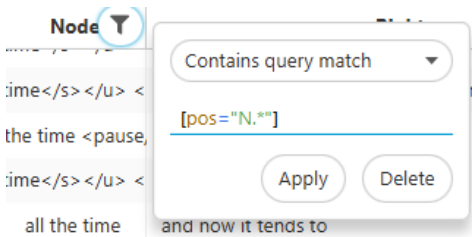
File	Left	N...	Right	Text: ge...	Text: subsubgenre	Text: date
FictSci85.xml	and to hide it he	...		fiction	fiction: sci-fi: yo...	2010
FictSci85.xml	but takes the money and	...		fiction	fiction: sci-fi: yo...	2010
FictMis381.xml	make money, and it all	...	directly into the company. You	fiction	fiction: miscellan...	2010
FictMis381.xml	it says that, but whoever	...	to a hundred and fifty	fiction	fiction: miscellan...	2010
FictMis399.xml	out in a passage that	...	to the surface.' I could	fiction	fiction: miscellan...	2010
FictChi13.xml	before the cart of history	...	past.' 'Come on,' Wormersley said,	fiction	fiction: children's...	2010
FictRom23.xml	No,' Dex said. 'If he	...	now, that's it, there	fiction	fiction: romance:...	2010
FictRom23.xml	him how suddenly the world	...	and changes. Here he was	fiction	fiction: romance:...	2010
FictRom23.xml	you today because when this	...	to court, you'll be	fiction	fiction: romance:...	2010
FictMis236.xml	works. Be creative: some paperwork	...	missing, one of your admirals	fiction	fiction: miscellan...	2010
FictMis390.xml	body. I decided.' 'This just	...	to prove you are not	fiction	fiction: miscellan...	2010
FictMys91.xml	get on there. Then it	...	to Central, where it gets	fiction	fiction: mystery: ...	2010
FictMys91.xml	plastic is best, and that	...	in one pile; blue in	fiction	fiction: mystery: ...	2010
FictMys91.xml	near me and Gardo, it	...	down the far end, and	fiction	fiction: mystery: ...	2010

3.4 筛选器

多种强大的筛选器可被应用于 i）语言和 ii）元语言数据。您只需将鼠标指针悬停在任何列标题的右侧，即可找到筛选器选项按钮 。

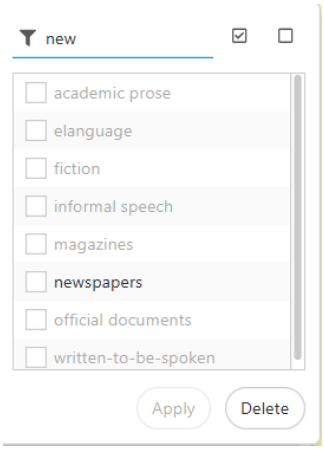
您可以使用完整的语言搜索功能来筛选语言数据，也可以对于所需语言特征应出现在左侧和右侧上下文的位置进行选择。





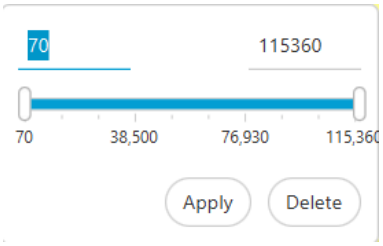
元语言数据可以根据三种数据类型进行过滤：i）类别、ii）数字和 iii）日期。

类别



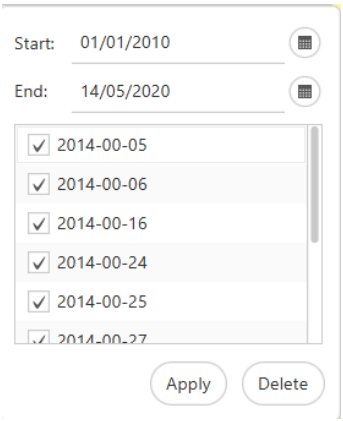
选择所需的类别，可通过勾选各类别旁边的复选框，或通过搜索类别并单击 按钮一键勾选所有被突出显示的相应类别。

数字



使用最小值和最大值或滑块选择数字范围。

日期

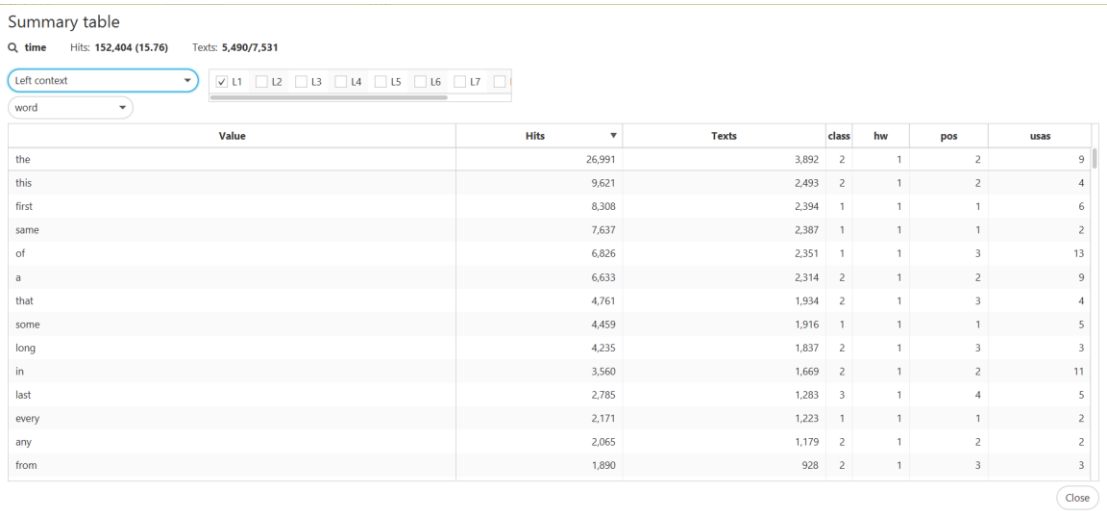


选择开始和结束日期。不符合有效的 YYYY-MM-DD 模式的日期将会显示为类别。

3.5 汇总表

在 KWIC 中显示的数据也可以使用“Summary Table（汇总表）” 功能进行总结。汇总表功能可以应用于 i）语言和 ii)元语言数据。

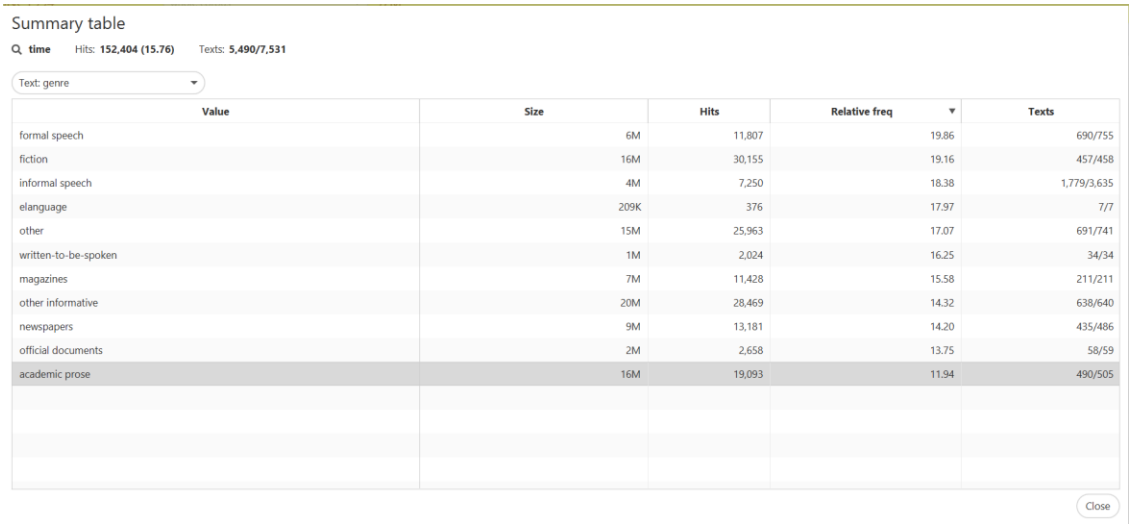
- 语言特征汇总包括以下信息：i）命中数（绝对频数），ii）出现该语言特征的文本数量，以及 iii）根据任何可用的语言标注进行细分，如 词性（POS）标签、语义标签（USAS）、词头（Headwords）等。



Value	Hits	Texts	class	hw	pos	usas
the	26,991	3,892	2	1	2	9
this	9,621	2,493	2	1	2	4
first	8,308	2,394	1	1	1	6
same	7,637	2,387	1	1	1	2
of	6,826	2,351	1	1	3	13
a	6,633	2,314	2	1	2	9
that	4,761	1,934	2	1	3	4
some	4,459	1,916	1	1	1	5
long	4,235	1,837	2	1	3	3
in	3,560	1,669	2	1	2	11
last	2,785	1,283	3	1	4	5
every	2,171	1,223	1	1	1	2
any	2,065	1,179	2	1	2	2
from	1,890	928	2	1	3	3

例如，上方的表格显示了在索引表格中，L1 位置最常见的单词是“the”，其次是“this”，“first”，“same”等。在 3,892 个不同的文本中，“the”在 L1 位置的绝对频数为 26,991 次。在该位置上，“the”共被标记为 2 个 POS 标签 AT 和 RT42，以及 9 个不同的语义 usas 标签。标签及其频率的详细信息可以使用鼠标悬停功能中的工具提示查看。

- 元数据汇总根据所选类别进行分解，包括以下信息：i）组件的大小，ii）组件中的命中次数（绝对频数），iii）组件中的相对频数，以及 iv）组件中出现语言特征的文本数量及其占组件中所有文本的比例。

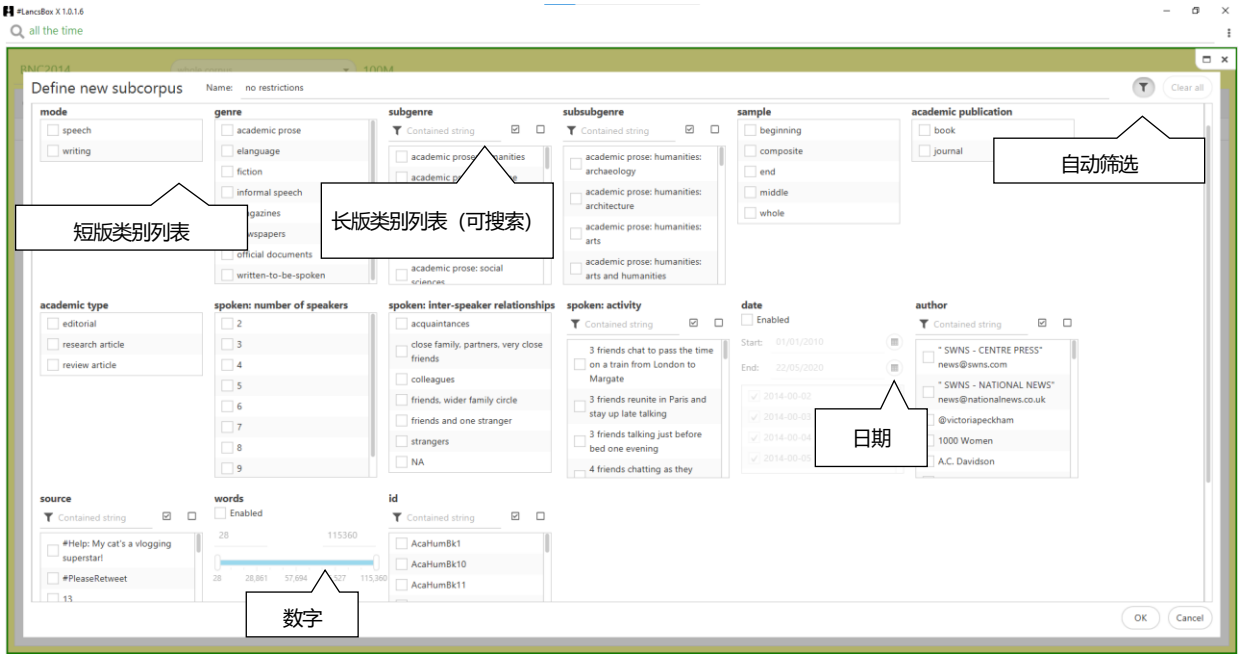


汇总表可以进行复制、粘贴或保存；保存还将包括通过工具提示显示的各个标签的详细信息。

3.6 使用子语料库

#LancsBox X 允许用户定义子语料库，并以此将搜索限制在语料库的特定部分中。要定义新的子语料库，请先单击子语料库下拉菜单，然后选择“New Corpus（新建子语料库）”选项。

在随即出现的覆盖层中，您可以选择定义子语料库的条件并为其选择一个名称。当选择好所有标准后，请单击“OK（确定）”。您的新子语料库将被选中。



您可以使用下拉菜单来更改子语料库。下拉菜单中的编辑和删除按钮允许您更改或删除已定义的子语料库。

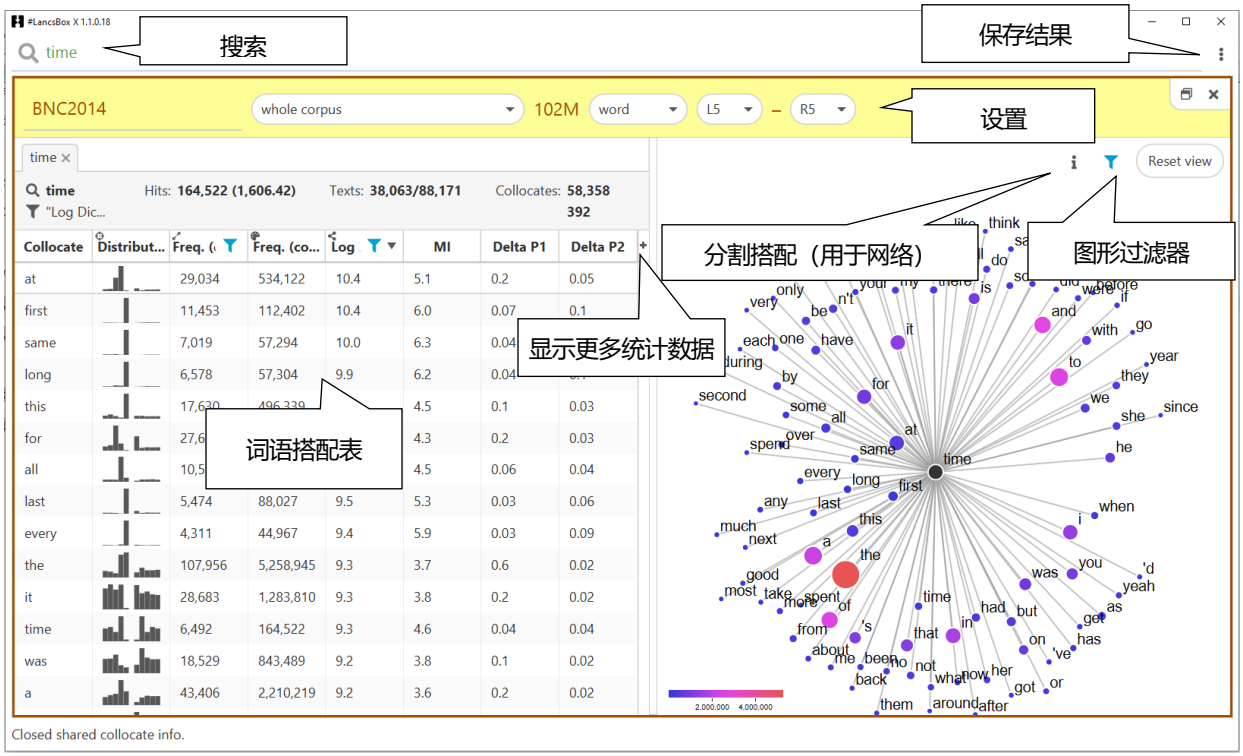
4 GraphColl

GraphColl 工具可以识别词语搭配(collocations)，并以表格、词语搭配图或词语搭配网络的形式显示。

例如，GraphColl 可以用于以下几种情况：

- 查找单词或短语的搭配词(collocates)。
- 查找语法范畴的共现，类链接(colligation)。
- 可视化词语搭配和类联接。
- 识别单词或短语的共享搭配词(collocates)。
- 用“关于性”(aboutness)的术语总结话语(discourse)。

4.1 GraphColl: 概述



4.2 生成词语搭配图

GraphColl 工具可以即时生成词语搭配表格和图形。选择合适的设置后，您可以开始搜索节点及其搭配词。

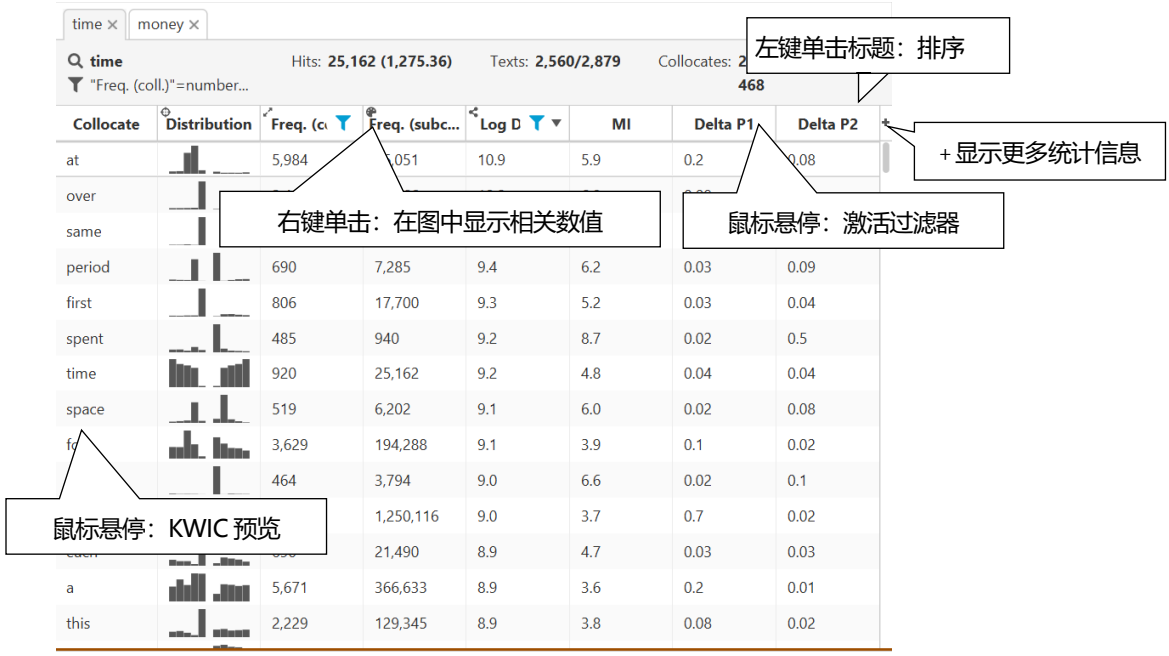
1. 选择适当的词语搭配搜索设置：

   i) 语料库和子语料库：选择现有的或定义新的（子）语料库。
   ii) 单位：用于搭配词的单位（例如单词、词目（hw）、词性（POS）、词元和词位）。
   iii) 范围：搜索范围，即在节点（检索项）左侧（L）和右侧（R）各包括多少个单词。
2. 在搜索框中输入检索项，并按 Enter 键。
3. 这将产生一个词语搭配表格（左侧）和一个词语搭配图（右侧）。

4.3 解读词语搭配表

词语搭配表是传统的显示搭配的方式。在 GraphColl 中，该表显示每个词语搭配的以下信息：i) 分布，ii) 搭配频率， iii) 该词汇搭配在语料库中的频率， iv) 所有相关的统计量。默认情况下，该表按默认的词汇搭配统计量“log Dice”排序（从大到小），并应用适当的频率过滤器。

1. 下面是关于词语搭配表的视觉描述。

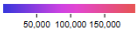


2. 每一列的含义如下：
   i) Collocate:显示所查询的搭配词。
   ii) Distribution: 显示条形图，指示该搭配词在文本中的位置（例如在 L5-R5 的范围内）。
   iii) Freq (coll): 显示词语搭配的频率（节点+搭配词的频率）。
   iv) Freq (corpus): 显示搭配词在语料库中出现的频率。
   v) Stats (names): 显示所选择关联度量值；同时计算所有可用的测量值。若要显示更多或更少的值，点击"+"按钮。

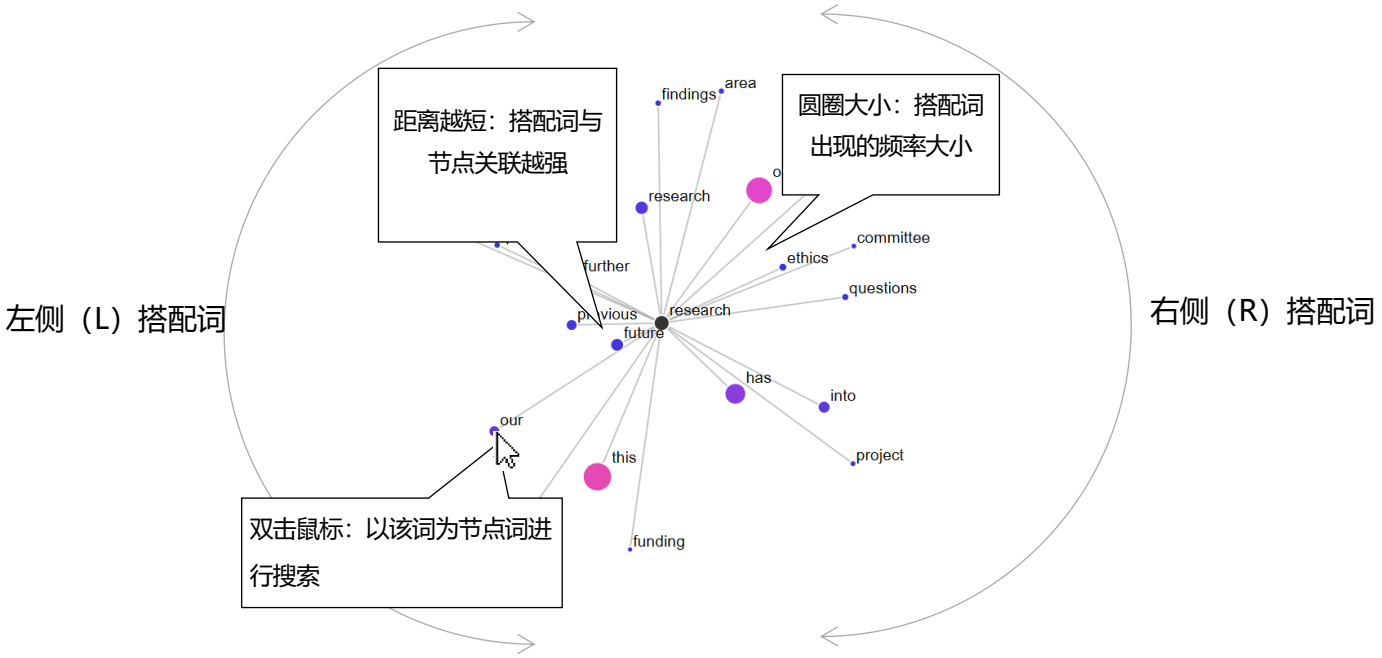
4.4 解读词语搭配图

图表根据表格设置显示多个维度（右键单击表头以将图表值分配给列）。要了解关于某个搭配词更多的信息，请将鼠标悬停在其上以获取索引行（KWIC 预览），其中搭配词与节点共现。

1. 线的长度: 默认情况下，边（线）的长度会根据默认的关联度量值进行分配，以表达搭配的强度。越接近节点的搭配词与节点的关联越强（"磁效应"）。
2. 大小: 每个搭配词圆圈的大小默认以频率（Freq (coll)）进行分配。 该词汇搭配出现的频率越高，圆圈越大。
3. 颜色: 每个圆圈的颜色默认根据搭配词出现在语料库中的频率（Freq (corpus)）来分配。频率范围显示在图例中。



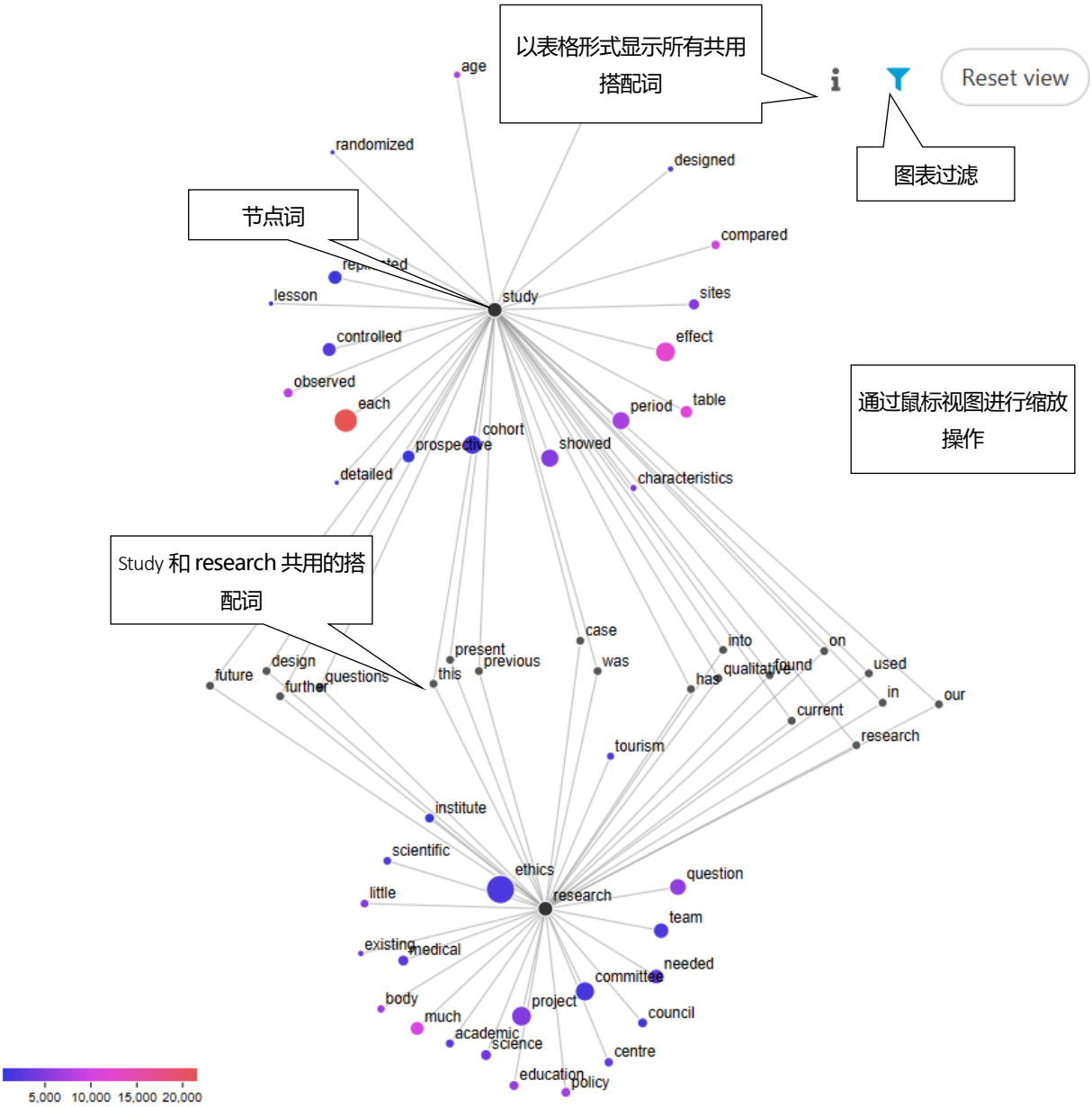
4. 位置: 图表中搭配词围绕节点词的位置反映了它们在文本中的位置关系：一些搭配词主要出现在节点词的左侧，一些更容易出现在右侧；搭配词出现在左侧和右侧的频率相似时，便会显示在图表中的中间位置。为了方便显示，如果多个搭配词出现在相似的位置并重叠，该工具会稍微拉开一些这些词汇之间的距离。



4.5 将搭配图扩展成搭配网络

词语搭配网络是扩展的词语搭配图，显示了 i）共用的搭配词和 ii）多个节点词之间的交叉关联。

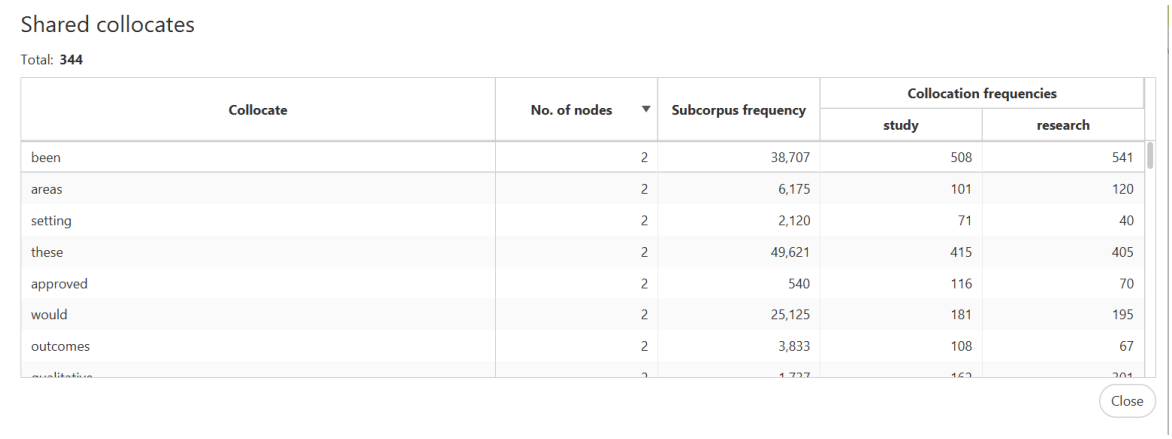
1. 要将简单的词语搭配图扩展为词语搭配词汇网络，请搜索更多的节点或在图中某个搭配处双击鼠标左键。
2. 词语搭配网络显示具有独特搭配的节点词（图的外部）和共用的搭配词（图的中间）。



4.6 共用搭配词

共用搭配词是指图表中至少有两个节点词共用的搭配词。共用搭配词会显示在图表中央，并链接到相关的节点词。

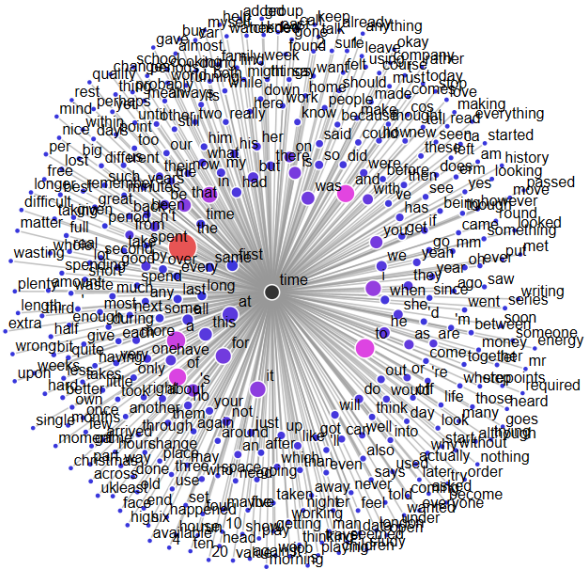
1. 点击“i”图标 ，可以获得完整的共用搭配词列表。
2. 共用搭配列表以表格形式显示。



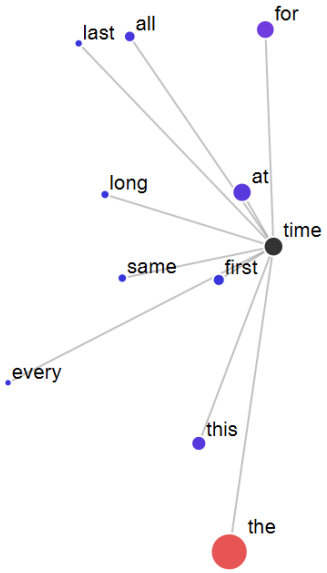
4.7 当图信息过载时

如果一个词语搭配图或网络包含了过多的节点和搭配词，那么它将变得难以阅读，这被称为过度拥挤的图表/网络。为了解决这个问题，可以在表格中更改筛选条件并使阈值更为严格，或者对图表应用筛选器。

以下图示展示了左侧是信息过载时生成的图形，右侧是更易于解读的图形。



含有 392 个搭配词的图表



显示前 10 个搭配词的图表

Choose the maximum number of collocates to show from each query. They will be selected by edge length variable.

Non-shared collocates per query 10

Shared collocates per query 10

Apply Delete

4.8 报告搭配词：CPN

重要的是要意识到，没有任何搭配词集合是一定确定的：不同的统计程序和阈值会突出不同的搭配词集合。因此，我们需要使用称为词搭配参数符号表示法（Collocation Parameters Notation（CPN））的标准符号来报告识别词语搭配所涉及的统计选择。当保存结果时，GraphColl 会以 CPN 的形式保存设置。

Brezina et al. (2015) 提出 CPN 作为一种用于准确描述搭配过程和复制结果的特定符号表示法。下列参数将被报告：

Statistic ID	Statistic name	Statistic cut-off value	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
4b	MI2	3	L5-R5	5	1	Function words removed
4b-MI2(3), L5-R5, C5-NC1; function words removed						

▶ 你知道吗？

GraphColl 的名字是“graphical collocations tool”的缩写。GraphColl 是 #LancsBox (v.1.0) 中的第一个模块，后来才添加了其他工具。Collocation 网络的图形展示是受到了 Phillips (1985)的启发，他在小型专业语料库中展示了“词汇网络”（Phillips 对词语搭配网络的术语）的概念。GraphColl 进一步发展了这一思想，在小型和大型语料库中实时生成词语搭配网络，并提供了不同的统计选择。

Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.

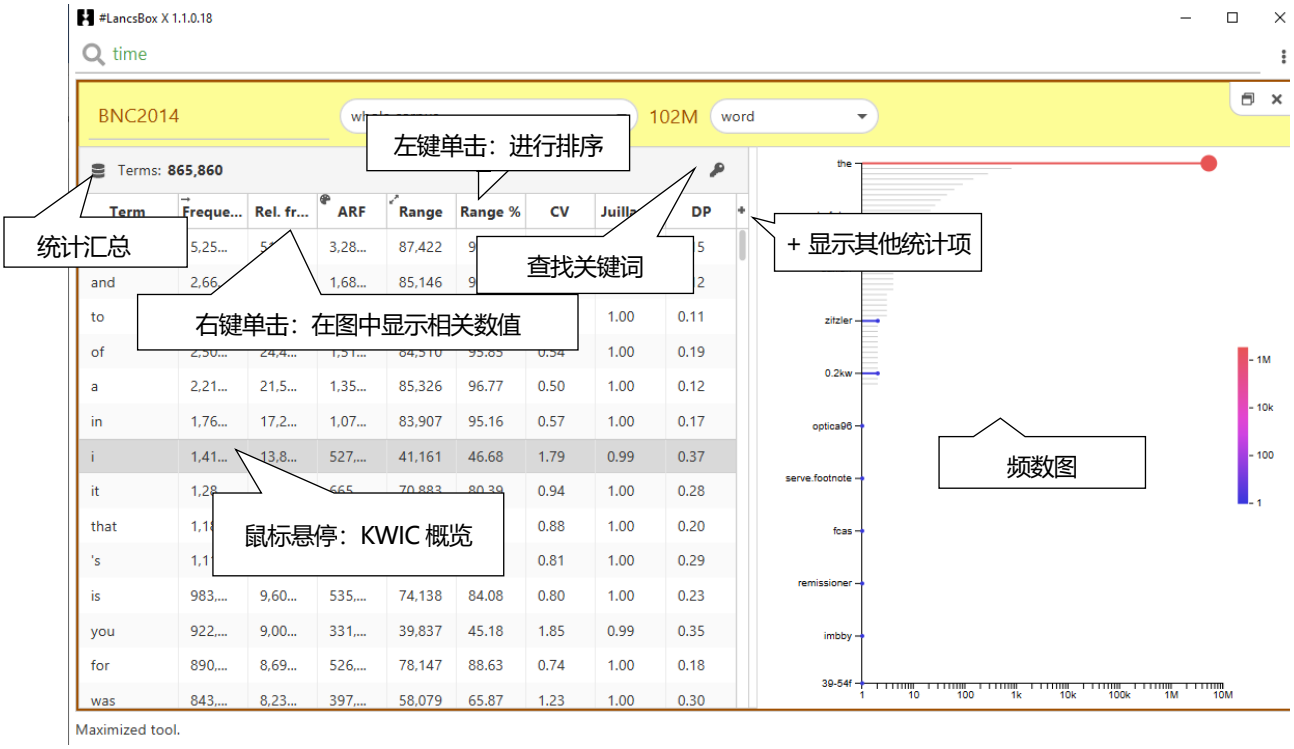
5 Words tool

词语工具可以深入分析词语频数、语法范畴和语义范畴的频数，并通过分析关键词来比较语料库。

例如，可以用于：

- 计算频数和离散度指标。
- 可视化语料库中的频数和离散度。
- 通过分析关键词来比较语料库。

5.1 Words 概览



左侧：生成词频表，计算离散度和关键词。

右侧：频数可视化

5.2 生成词频表

当打开该工具时，Words tool 会根据默认语料库和默认设置显示词频表。可轻松更改相关设置，以生成不同的词频表。

1. 以下是词频表的设置：

i) 语料库和子语料库：选择已存在或重新设定

ii) 单位：词频表统计的单位（例如单词、词目（hw）、词性（POS）、词元、词位）

2. 一次性计算出所有频数和离散度指标。
3. 可使用搜索框（顶部）搜索词频表。
4. 可左键单击各列标题对词频表进行排序。
5. 可在各列使用过滤器对词频表进行筛选。

提示：请注意在#LancsBox X 版本中，词频表被预先计算和存储供后续使用。如果您首次创建词表，过程可能会需要一些时间，具体取决于语料库的大小和注释的复杂性（被统计单位的数量）。

5.3 生成关键词

Words 模块使用所选的统计指标来计算和比较两个语料库/词表之间的频数。

1. 点击表格右上角的钥匙图标 。
2. 选择合适的参照语料库。
3. 根据您偏好的关键词统计程序对数据进行排序和/或筛选（默认使用 Simple Maths 来排序）。

Keywords

Reference corpus: BNC2014 whole corpus

Terms: **865,860**

Term	Focus rel. freq. (...	Reference rel. fr...	Simple maths ▼	Log likelihood	% difference	Log ratio
et	2,615.35	516.57	4.40	NaN	406.29	2.34
al.	1,991.15	383.75	4.32	NaN	418.87	2.38
fig.	1,120.67	215.91	3.86	688,915.67	419.06	2.38
studies	921.47	203.08	3.37	630,539.84	353.74	2.18
data	1,419.01	353.43	3.35	NaN	301.49	2.01
study	1,294.72	317.11	3.34	NaN	308.29	2.03
analysis	925.53	220.50	3.20	NaN	319.73	2.07
e.g.	514.51	102.49	3.03	NaN	401.99	2.33

Close

▶ 你知道吗？

统计分析关键词的技术最初由 Mike Scott（1997）开发，并在 WordSmith Tools 中实施。它依赖卡方检验（chi-squared test）或对数似然比（log-likelihood）检验来比较语料库。正如 Kilgarriff 所指出，卡方检验和对数似然比检验并不完全合适于此。Kilgarriff 在 Sketch Engine 中提出的解决方案是使用“Simple Maths”程序来比较语料库，即比较两个语料库中词汇相对频数的简单比率。除“Simple Maths”外，#LancsBox 还提供其他类型的解决方案来比较语料库。

Scott, M. (1997). PC analysis of key words—and key key words. System, 25(2), 233-245.
Kilgarriff, A. (2009, July). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference. Liverpool, UK*.

6 在#LancsBox 中检索

#LancsBox 有强大的检索功能，能检索语料库不同层次的注释，包括：i) 简单检索 ii) 通配符检索 iii) 标点符号检索 iv) 智能检索 v) CQL（语料库检索语言）检索。

1. 简单检索是对特定单词（例如"new"）或短语（例如"New York Times"）进行文字检索。简单检索不区分大小写，这意味着检索"new"、"New"、"NEW"、"*NeW*"等将得到相同结果。

2. 通配符检索是使用星号（*）作为特殊字符的检索。

特殊字符	含义	示例
*	0 或以上字符 a 任何词[有空格]	new* [*new, news, newly, newspaper…*] new *[*new car, New York, new ideas…*]

3. 标点符号检索
   检索标点符号，请使用以下示例中斜杠（/）符号。
   /?/
   hello /,/

4. 智能检索是用户使用预定义的检索工具简便地进行复杂检索，是#LancsBox 独有的功能。可用于检索词类（名词、动词等）、复杂的语法形式（被动、分裂不定式等）和语义类别（地点副词）。

以下智能检索适用于英语：

ADJECTIVE 形容词
ADVERB 副词
BE 动词
BODY 身体
BOOSTER 强化
COLLECTIVE_NOUN 集体名词
COLOUR 颜色
COMPARATIVE 比较级
COMPLEX_NOUN_PHRASE 复杂名词短语
CONDITIONAL 条件句
CONNECTOR 连接词
CONTRACTION 缩写形式
DEGREE ADVERB 程度副词
DETERMINER 限定词
DO 动词
DOWNTONER 减弱

EMOTION 情感
EXISTENTIAL_THERE, There 表存在
FEMALE 女性
FOOD 食物
GERUND 动名词
HAVE 动词
HYPHENATED_WORD 含连字符的单词
INDEFINITE_PRONOUN 不定代词
INFINITIVE 不定式
INTERJECTION 感叹词
LINKING_ADVERB 连接副词
LONG_WORD 长词
MALE 男性
MEDIA 媒体
MODAL 情态动词
NEGATION 否定

NOMINALIZATION 名词化
NOUN 名词
NUMBER 数字
PARTICLE 小品词
PASSIVE 被动语态
PAST_PARTICIPLE 过去分词
PAST_TENSE 过去时态
PEOPLE 人
PERFECT_INFINITIVE 不定式完成式
PHRASAL_VERB 动词短语
PLACE_ADVERB 地点副词
PLANET 行星
PREPOSITIONAL_PHRASE 介词短语
PRESENT_TENSE 现在时
PRONOUN 代词
PROPER_NOUN 专有名词
REFLEXIVE_PRONOUN 反身代词
SHORT_WORD 短词
SPLIT_INFINITIVE 分裂不定式
SUPERLATIVE 最高级
SUPERNATURAL 超自然
SWEARWORDS 脏话
TECHNOLOGY 技术
TIME 时间
TIME_ADVERB 时间副词
VERB 动词

5. CQL（语料库检索语言）检索。#LancsBox 支持使用 CQL 进行强大的搜索.

可用于定义对不同层次注释的复杂检索。
注释层次和句法取决于语料库的标注方式，但对 XML 语料库，通常有以下标注：i）单词，ii）词头/词目（hw），iii）词性（POS），iv）用户定义的标签。例如，可在 CQL 中检索单个形符，如下所示：

[word="goes" hw="go" pos="V.*" sem="M1"]

这将对应所有包含单词"goes"、词目为"go"、词性标签为 V.*（动词）以及 usas 标签为 M1（移动、来去）的实例。若无指定注释层次，则对该层次上不会附加任何限制。双引号中的所有内容都被解释为不区分大小写的正则表达式。

要进行区分大小写的检索，请使用双等号（==），示例如下:

[word=="US"]

要检索标点符号，请使用斜杠（/）和属性名 punc，示例如下。请注意，特殊字符（如问号或句号）需要通过反斜杠符号（\）进行标注。

/punc="\?|\.|,|;"/

可按顺序列出多个形符。空方括号[]将匹配任何形符。可使用句法{X}表示形符重复 X 次，且可使用句法{Y, Z}表示形符重复 Y 到 Z 之间的次数。{0, 1}的简写是一个问号。因此，例如以下 CQL 表达式：

[pos="VB.*"] []{0,3} [pos="V.N"]?

表示"to be"动词（VB.*）后面跟着 0 到 3 个未限定的形符（[]{0,3}），且其后可选择跟过去分词（V.N）。

检索部分也可用括号()括起来，可将量词（如{1,2}）置于形符序列，例如([pos="N.*"] [word="and"]){2}。单词、短语和智能检索可检索任意 CQL 形符，例如 very{2} ADJECTIVE{1,2} [hw="year"]。

CQL 还支持检索 XML 结构。该检索可匹配每个<u></u>元素，<u/>表示口语。以下示例匹配数量为 1 且国籍属性为英国或美国的口语实例：

<u n="1" nationality="British|American"/>

此类元素的检索可使用 within 句法与其他类型的检索相结合：

```
[pos="D.*"] green NOUN within <text genre="newspapers"/>
```

该检索匹配在报纸文本中，限定词后面紧跟着"green"，再紧跟着一个名词的各实例。Within 检索的左右两侧可添加任意内容，亦可为其他的 within 检索。

```
(<emoji/> within please) within (<e/> within <text genre="elanguage"/>)
```

7 CLAWS 赋码集（C7）

原文链接 Source: http://ucrel.lancs.ac.uk/claws7tags.html

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our) 形容词性物主代词
AT	Article (e.g. the, no) 冠词
AT1	Singular article (e.g. a, an, every) 单数冠词
BCL	before-clause marker (e.g. in order (that),in order (to)) 从句前标记词
CC	coordinating conjunction (e.g. and, or) 并列连词
CCB	adversative coordinating conjunction (but) 转折连词
CS	subordinating conjunction (e.g. if, because, unless, so, for) 从属连词
CSA	as (as conjunction) 连词 as
CSN	than (as conjunction) 连词 than
CST	that (as conjunction) 连词 that
CSW	whether (as conjunction) 连词 whether
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same) 起代词作用的后位限定词
DA1	singular after-determiner (e.g. little, much) 单数后位限定词
DA2	plural after-determiner (e.g. few, several, many) 复数后位限定词
DAR	comparative after-determiner (e.g. more, less, fewer) 比较级后位限定词
DAT	superlative after-determiner (e.g. most, least, fewest) 最高级后位限定词
DB	before determiner or pre-determiner capable of pronominal function (all, half) 起代词作用的前位限定词
DB2	plural before-determiner (both) 复数前位限定词
DD	determiner (capable of pronominal function) (e.g. any, some) 限定词（可起代词作用）
DD1	singular determiner (e.g. this, that, another) 单数限定词
DD2	plural determiner (these,those) 复数限定词
DDQ	wh-determiner (which, what) wh-限定词
DDQGE	wh-determiner, genitive (whose) wh-限定词（所有格）
DDQV	wh-ever determiner (whichever, whatever) wh-ever 限定词
EX	existential there, there 表存在
FO	Formula 公式
FU	unclassified word 未分类词
FW	foreign word 外语单词
GE	germanic genitive marker (' or's) 所有格标记

IF for (as preposition) 介词 for

II general preposition 一般介词

IO of (as preposition)介词 of

IW with, without (as prepositions) 介词 with, without

JJ general adjective 一般形容词

JJR general comparative adjective (e.g. older, better, stronger) 一般形容词比较级

JJT general superlative adjective (e.g. oldest, best, strongest) 一般形容词最高级

JK catenative adjective (able in be able to, willing in be willing to) 链接形容词

MC cardinal number,neutral for number (two, three..) 基数词（不分单复数）

MC1 singular cardinal number (one) 单数基数词

MC2 plural cardinal number (e.g. sixes, sevens) 复数基数词

MCGE genitive cardinal number, neutral for number (two's, 100's) 所有格基数词（不分单复数）

MCMC hyphenated number (40-50, 1770-1827) 带连字符的数字

MD ordinal number (e.g. first, second, next, last) 序数词

MF fraction,neutral for number (e.g. quarters, two-thirds) 分数（不分单复数）

ND1 singular noun of direction (e.g. north, southeast) 单数方向词

NN common noun, neutral for number (e.g. sheep, cod, headquarters) 普通名词（不分单复数）

NN1 singular common noun (e.g. book, girl) 单数普通名词

NN2 plural common noun (e.g. books, girls) 复数普通名词

NNA following noun of title (e.g. M.A.) 称谓名词后成分

NNB preceding noun of title (e.g. Mr., Prof.) 称谓名词前成分

NNL1 singular locative noun (e.g. Island, Street) 单数方位名词

NNL2 plural locative noun (e.g. Islands, Streets) 复数方位名词

NNO numeral noun, neutral for number (e.g. dozen, hundred) 数字名词（不分单复数）

NNO2 numeral noun, plural (e.g. hundreds, thousands)复数数字名词

NNT1 temporal noun, singular (e.g. day, week, year) 单数时间名词

NNT2 temporal noun, plural (e.g. days, weeks, years) 复数时间名词

NNU unit of measurement, neutral for number (e.g. in, cc) 测量单位（不分单复数）

NNU1 singular unit of measurement (e.g. inch, centimetre) 单数测量单位

NNU2 plural unit of measurement (e.g. ins., feet) 复数测量单位

NP proper noun, neutral for number (e.g. IBM, Andes) 专有名词（不分单复数）

NP1 singular proper noun (e.g. London, Jane, Frederick) 单数专有名词

NP2 plural proper noun (e.g. Browns, Reagans, Koreas) 复数专有名词

NPD1 singular weekday noun (e.g. Sunday) 单数星期几名词

NPD2 plural weekday noun (e.g. Sundays) 复数星期几名词

NPM1 singular month noun (e.g. October) 单数月份名词

NPM2 plural month noun (e.g. Octobers) 复数月份名词

PN indefinite pronoun, neutral for number (none) 不定代词（不分单复数）

PN1 indefinite pronoun, singular (e.g. anyone, everything, nobody, one) 单数不定代词

PNQO objective wh-pronoun (whom) 宾语 wh-代词

PNQS subjective wh-pronoun (who) 主语 wh-代词

PNQV wh-ever pronoun (whoever) wh-ever 代词

PNX1 reflexive indefinite pronoun (oneself) 反身不定代词

PPGE nominal possessive personal pronoun (e.g. mine, yours) 名词物主人称代词

PPH1 3rd person sing. neuter personal pronoun (it) 第三人称单数中性人称代词

PPHO1 3rd person sing. objective personal pronoun (him, her) 第三人称单数宾语人称代词

PPHO2 3rd person plural objective personal pronoun (them) 第三人称复数宾语人称代词

PPHS1 3rd person sing. subjective personal pronoun (he, she) 第三人称单数主语人称代词

PPHS2 3rd person plural subjective personal pronoun (they) 第三人称复数主语人称代词

PPIO1 1st person sing. objective personal pronoun (me) 第一人称单数宾语人称代词

PPIO2 1st person plural objective personal pronoun (us) 第一人称复数宾语人称代词

PPIS1 1st person sing. subjective personal pronoun (I) 第一人称单数主语人称代词

PPIS2 1st person plural subjective personal pronoun (we) 第一人称复数主语人称代词

PPX1 singular reflexive personal pronoun (e.g. yourself, itself) 单数反身人称代词

PPX2 plural reflexive personal pronoun (e.g. yourselves, themselves) 复数反身人称代词

PPY 2nd person personal pronoun (you) 第二人称代词

RA adverb, after nominal head (e.g. else, galore) 名词后的副词

REX adverb introducing appositional constructions (namely, e.g.) 引出同位结构的副词

RG degree adverb (very, so, too) 程度副词

RGQ wh- degree adverb (how) wh-程度副词

RGQV wh-ever degree adverb (however) wh-ever 程度副词

RGR comparative degree adverb (more, less) 比较级程度副词

RGT superlative degree adverb (most, least) 最高级程度副词

RL locative adverb (e.g. alongside, forward) 方位副词

RP prep. adverb, particle (e.g about, in) 介词性副词（小品词）

RPK prep. adv., catenative (about in be about to) 介词性副词（链接词）

RR general adverb 一般副词

RRQ wh- general adverb (where, when, why, how) wh-一般副词

RRQV wh-ever general adverb (wherever, whenever) wh-ever 一般副词

RRR comparative general adverb (e.g. better, longer) 比较级一般副词

RRT superlative general adverb (e.g. best, longest) 最高级一般副词

RT quasi-nominal adverb of time (e.g. now, tomorrow) 准名词的时间副词

TO infinitive marker (to) 不定式标记

UH interjection (e.g. oh, yes, um) 感叹词

VB0 be, base form (finite i.e. imperative, subjunctive) be 原形

VBDR Were

VBDZ was

VBG being

VBI be, infinitive (To be or not... It will be ..) be 不定式

VBM am

VBN been

VBR are

VBZ is

VD0 do, base form (finite) do 原形

VDD did

VDG doing

VDI do, infinitive (I may do... To do...) do 不定式

VDN done

VDZ does

VH0 have, base form (finite) have 原形

VHD had (past tense) had（过去式）

VHG having

VHI have, infinitive have 不定式

VHN had (past participle) had（过去分词）

VHZ has

VM modal auxiliary (can, will, would, etc.) 情态助动词

VMK modal catenative (ought, used) 情态类链接动词

VV0 base form of lexical verb (e.g. give, work) 实义动词原形

VVD past tense of lexical verb (e.g. gave, worked) 实义动词过去式

VVG -ing participle of lexical verb (e.g. giving, working) 实义动词-ing 分词

VVGK -ing participle catenative (going in be going to) 作-ing 分词的链接动词

VVI infinitive (e.g. to give... It will work...) 不定式

VVN past participle of lexical verb (e.g. given, worked) 实义动词过去分词

VVNK past participle catenative (e.g. bound in be bound to) 作过去分词的链接动词

VVZ -s form of lexical verb (e.g. gives, works) 实义动词-s 形式

XX not, n't

ZZ1 singular letter of the alphabet (e.g. A,b) 字母单数

ZZ2 plural letter of the alphabet (e.g. A's, b's) 字母复数

8 USAS 赋码集

原文链接: http://ucrel.lancs.ac.uk/usas

A1 GENERAL AND ABSTRACT TERMS
一般和抽象术语
A1.1.1 General actions, making etc.
一般动作
A1.1.2 Damaging and destroying
破坏
A1.2 Suitability
适合
A1.3 Caution
小心
A1.4 Chance, luck
机会，幸运
A1.5 Use
用处
A1.5.1 Using
使用
A1.5.2 Usefulness
有效性
A1.6 Physical/mental
生理/心理
A1.7 Constraint
限制
A1.8 Inclusion/Exclusion
包括/排斥
A1.9 Avoiding
避免
A2 Affect
影响
A2.1 Affect:- Modify, change
影响：调整，改变
A2.2 Affect:- Cause/Connected
影响：原因/联系
A3 Being 存在
A4 Classification
分类
A4.1 Generally kinds, groups, examples
一般分类和列证
A4.2 Particular/general; detail
特殊/一般;细节
A5 Evaluation
评价
A5.1 Evaluation:- Good/bad
评价：好/坏
A5.2 Evaluation:- True/false
评价：正/误
A5.3 Evaluation:- Accuracy
评价：准确性
A5.4 Evaluation:- Authenticity
评价：真实性
A6 Comparing
比较
A6.1 Comparing:- Similar/different
比较：相似性
A6.2 Comparing:- Usual/unusual
比较：稀有性
A6.3 Comparing:- Variety
比较：多样性
A7 Definite (+ modals)
确定（+情态动词）
A8 Seem
看来
A9 Getting and giving; possession
取予；拥有
A10 Open/closed; Hiding/Hidden; Finding; Showing
开/关 隐藏 寻找 展示
A11 Importance
重要性
A11.1 Importance: Important
重要性：重要
A11.2 Importance: Noticeability
重要性：显著
A12 Easy/difficult
容易 困难
A13 Degree
程度
A13.1 Degree: Non-specific
程度：不确定
A13.2 Degree: Maximizers
程度：最大化
A13.3 Degree: Boosters
程度：增强
A13.4 Degree: Approximators
程度：估计
A13.5 Degree: Compromisers
程度：妥协
A13.6 Degree: Diminishers
程度：减弱
A13.7 Degree: Minimizers
程度：最小化
A14 Exclusivizers/particularizers
排斥化/特殊化
A15 Safety/Danger
安全/危险
B1 Anatomy and physiology 解剖与生理

B2 Health and disease
健康与疾病
B3 medicines and medical treatment
医药与治疗
B4 Cleaning and personal care
清洁与个人护理
B5 Clothes and personal belongings
衣物和私人用品
C1 Arts and crafts
艺术与工艺
E1 EMOTIONAL ACTIONS, STATES AND PROCESSES General
一般情感类行为、状态和过程
E2 Liking
喜爱
E3 Calm/Violent/Angry
平静/强烈/愤怒
E4 Happy/sad
喜与悲
E4.1 Happy/sad: Happy
喜与悲：高兴
E4.2 Happy/sad: Contentment
喜与悲：满足
E5 Fear/bravery/shock
恐惧/勇敢/震惊
E6 Worry, concern, confident
担忧, 关注, 信心
F1 Food
食物
F2 Drinks
饮品
F3 Cigarettes and drugs
香烟与毒品
F4 Farming & Horticulture
农业与园艺
G1 Government, Politics and elections
政府, 政治和选举
G1.1 Government etc.
政府
G1.2 Politics
政治
G2 Crime, law and order
犯罪,法律和秩序
G2.1 Crime, law and order: Law and order
犯罪,法律和秩序：法律和秩序
G2.2 General ethics
一般伦理
G3 Warfare, defence and the army; weapons
战争, 国防, 军队;武器
H1 Architecture and kinds of houses and buildings
建筑学和各类房屋
H2 Parts of buildings
建筑局部
H3 Areas around or near houses
建筑周围
H4 Residence
住房
H5 Furniture and household fittings
家具和家装
I1 Money generally
金钱：总体
I1.1 Money: Affluence
金钱：富裕
I1.2 Money: Debts
金钱：负债
I1.3 Money: Price
金钱：价格
I2 Business
商业
I2.1 Business: Generally
商业：总体
I2.2 Business: Selling
商业：销售
I3 Work and employment
工作与就业
I3.1 Work and employment: Generally
工作与就业：总体
I3.2 Work and employmeny: Professionalism
工作与就业：职业化
I4 Industry
工业
K1 Entertainment generally
娱乐：总体
K2 Music and related activities
音乐及相关活动
K3 Recorded sound etc.
录音等
K4 Drama, the theatre and showbusiness
戏剧与演艺业
K5 Sports and games generally
体育和游戏：总体
K5.1 Sports
体育
K5.2 Games
游戏
K6 Childrens games and toys
儿童游戏 玩具
L1 Life and living things
生命与生物
L2 Living creatures generally
生物：总体
L3 Plants
植物
M1 Moving, coming and going
移动, 来去
M2 Putting, taking, pulling, pushing, transporting &c.
放置, 推拉, 传送
M3 Vehicles and transport on land
陆地交通

M4 Shipping, swimming etc. 船运，游泳
M5 Aircraft and flying 飞行器和飞行
M6 Location and direction 位置和方向
M7 Places 地点
M8 Remaining/stationary 静止
N1 Numbers 数字
N2 Mathematics 数学
N3 Measurement 测量
N3.1 Measurement: General 测量：总体
N3.2 Measurement: Size 测量：大小
N3.3 Measurement: Distance 测量：距离
N3.4 Measurement: Volume 测量：体积
N3.5 Measurement: Weight 测量：重量
N3.6 Measurement: Area 测量：面积
N3.7 Measurement: Length & height 测量：长和高
N3.8 Measurement: Speed 测量：速度
N4 Linear order 线性顺序
N5 Quantities 数量
N5.1 Entirety; maximum 全部;最大值
N5.2 Exceeding; waste 超过;浪费
N6 Frequency etc. 频数等
O1 Substances and materials generally 实体与材料：总体
O1.1 Substances and materials generally: Solid 实体与材料：固体
O1.2 Substances and materials generally: Liquid 实体与材料：液体
O1.3 Substances and materials generally: Gas 实体与材料：气体
O2 Objects generally 物体：总体
O3 Electricity and electrical equipment 电与电气设备
O4 Physical attributes 实体特性
O4.1 General appearance and physical properties 总体外观与实体属性
O4.2 Judgement of appearance (pretty etc.) 对外观的评判（如美丽）
O4.3 Colour and colour patterns 色彩和颜色图案
O4.4 Shape 形状
O4.5 Texture 质地
O4.6 Temperature 温度
P1 Education in general 教育：总体
Q1 LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION 语言行为, 状态和过程;传播
Q1.1 LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION 语言行为, 状态和过程;传播
Q1.2 Paper documents and writing 书面文件和写作
Q1.3 Telecommunications 电信交流
Q2 Speech acts 演讲
Q2.1 Speech etc:- Communicative 演讲：交流
Q2.2 Speech acts 演讲活动
Q3 Language, speech and grammar 语言，言语和语法
Q4 The Media 媒体
Q4.1 The Media:- Books 媒体：书籍
Q4.2 The Media:- Newspapers etc. 媒体：报刊
Q4.3 The Media:- TV, Radio and Cinema 媒体：影视
S1 SOCIAL ACTIONS, STATES AND PROCESSES 社会行为, 状态和过程
S1.1 SOCIAL ACTIONS, STATES AND PROCESSES 社会行为,状态和过程
S1.1.1 SOCIAL ACTIONS, STATES AND PROCESSES 社会行为,状态和过程
S1.1.2 Reciprocity 交互性
S1.1.3 Participation 参与性
S1.1.4 Deserve etc. 应当等
S1.2 Personality traits 个人特性
S1.2.1 Approachability and Friendliness 亲近和友好程度

S1.2.2 Avarice
贪婪
S1.2.3 Egoism
自我中心
S1.2.4 Politeness
礼貌
S1.2.5 Toughness; strong/weak
坚强；强/弱
S1.2.6 Sensible
理智
S2 People
人
S2.1 People:- Female
人：女人
S2.2 People:- Male
人：男人
S3 Relationship
人际关系
S3.1 Relationship: General
人际关系：总体
S3.2 Relationship: Intimate/sexual
人际关系：亲密/性
S4 Kin
亲戚
S5 Groups and affiliation
团体与从属
S6 Obligation and necessity
义务与需要
S7 Power relationship
权力关系
S7.1 Power, organizing
权力, 组织
S7.2 Respect
尊敬
S7.3 Competition
竞争
S7.4 Permission
批准
S8 Helping/hindering
促进/阻碍
S9 Religion and the supernatural
宗教与超自然
T1 Time
时间
T1.1 Time: General
时间：总体
T1.1.1 Time: General: Past
时间：总体：过去
T1.1.2 Time: General: Present; simultaneous
时间：总体：现在
T1.1.3 Time: General: Future
时间：总体：未来
T1.2 Time: Momentary
时间：瞬间
T1.3 Time: Period
时间：片段
T2 Time: Beginning and ending
时间：始与终
T3 Time: Old, new and young; age
时间：新旧；年龄
T4 Time: Early/late
时间：早与迟
W1 The universe
宇宙
W2 Light
光
W3 Geographical terms
地理术语
W4 Weather
天气
W5 Green issues
环境问题
X1 PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES
心理活动, 状态和过程
X2 Mental actions and processes
精神活动和过程
X2.1 Thought, belief
想法, 信念
X2.2 Knowledge 知识
X2.3 Learn
学习
X2.4 Investigate, examine, test, search
调查，测试，查找
X2.5 Understand
理解
X2.6 Expect
期望
X3 Sensory
感官
X3.1 Sensory:- Taste
感官：味觉
X3.2 Sensory:- Sound
感官：听觉
X3.3 Sensory:- Touch
感官：触觉
X3.4 Sensory:- Sight
感官：视觉
X3.5 Sensory:- Smell
感官：嗅觉
X4 Mental object
心理物体
X4.1 Mental object:- Conceptual object
心理物体：概念物体
X4.2 Mental object:- Means, method 心理物体：方式方法
X5 Attention
注意力
X5.1 Attention
注意力
X5.2 Interest/boredom/excited/energetic
兴趣/厌烦/兴奋/有精力
X6 Deciding
决定
X7 Wanting; planning; choosing
想要, 计划；选择
X8 Trying
尝试

X9 Ability
能力
X9.1 Ability:- Ability, intelligence
能力：技能与智力
X9.2 Ability:- Success and failure
能力：成功与失败
Y1 Science and technology in general
科技常规类
Y2 Information technology and computing
信息技术与计算机
Z0 Unmatched proper noun
未分类的专有名词
Z1 Personal names
人名
Z2 Geographical names
地名
Z3 Other proper names
其他专有名词
Z4 Discourse Bin
话语垃圾箱
Z5 Grammatical bin
语法垃圾箱
Z6 Negative
否定
Z7 If
如果
Z8 Pronouns etc.
代词等
Z9 Trash can
垃圾箱
Z99 Unmatched
未分类

9 智能检索定义

ADJECTIVE 形容词	[pos="J.*"]
ADVERB 副词	[pos="R.*"]
BE BE 动词	[pos="VB.*"]
BOOSTER 强化	[hw="absolutely|altogether|completely|enormously|entirely|extremely|fully|greatly|highly|intensely|perfectly|strongly|thoroughly|totally|utterly|very"]
COLLECTIVE_NOUN 集体名词	[hw="a" pos="D.*"][hw="aerie|album|ambush|anthology|archipelago|argument|argumentation|armada|army|array|arsenal|ascension|assembly|aurora|badelynge|bag|bale|band|bank|banner|barrel|barren|bask|basket|batch|battery|bazaar|bed|bellowing|belt|bench|bevy|bew|bill|bind|bits|blessing|bloat|block|blush|board|bob|body|boil|boll|bond|book|bouquet|bowl|brace|branch|brew|brigade|brood|bubble|budget|building|bunch|bundle|bury|business|cache|canteen|caravan|cartload|cast|caste|catalogue|catch|cavalcade|celebration|cete|chain|charm|chatter|chattering|chest|chine|choir|chorus|circle|circus|clamour|clan|clash|clashing|class|clattering|clew|clique|cloud|clowder|cluck|clump|cluster|clutch|clutter|coalition|coil|collection|colony|column|comb|commonwealth|communion|community|company|compendium|confab|conflagration|confraternity|confusion|congregation|congress|conspiracy|constellation|converting|convocation|convoy|copse|cornucopia|corps|cortege|cost|cote|coterie|coven|cover|covert|covey|cowardice|cran|crash|crate|creche|crew|crop|crowd|cry|culture|death|deceit|deck|den|descent|desert|destruction|dicker|disguising|dissimulation|diving|division|doading|dole|dopping|dout|down|doyft|draft|draught|dray|drift|dropping|drove|drum|dule|durante|dynasty|earth|eleven|embarrassment|equivocation|erst|escargatoire|exaltation|faculty|faggot|fall|family|farrow|fellowship|fesnying|fesnyng|festival|fesynes|fidget|field|fine|fitting|fixie|flange|flap|fleet|flick|flight|fling|flink|float|flock|flotilla|flourish|flush|fluther|flutter|fold|forest|fraunch|fun|gaggle|galaxy|gam|gang|garland|garrison|gathering|gatling|gaze|generation|giggle|glaring|gleam|glide|glint|glitter|glory|glossary|grist|group|grove|gulp|hail|hand|haras|harem|harvest|haul|head|heap|heard|hedge|herd|hill|hive|holiness|horde|host|house|hover|huddle|hunt|hurtle|husk|illusion|implausibility|index|infestation|intrusion|invention|kaleidoscope|kendle|kennel|kettle|kindle|kine|kingdom|knab|knob|knot|labour|lamentation|layer|lead|leap|leash|lepe|library|line|list|litter|lodge|loft|lounge|loveliness|machination|malapertness|marvel|mask|mass|match|melody|memory|menagerie|mess|mews|miller|mischief|mob|mouthful|movement|multiply|murder|murmuration|muscle|muster|mustering|mutation|mute|necklace|nest|neverthriving|nide|nosegay|nuisance|number|nursery|nye|obesiance|observance|obstinacy|orchard|orchestra|ostentation|outfit|pace|pack|packet|paddling|pair|panel|panes|pantheon|parade|parcel|parel|park|parliament|party|passel|patrol|peal|peep|pencil|piddle|pile|pint|pit|piteousness|pitying|plague|platoon|plump|pocket|pod|ponder|pontification|pool|posse|pounce|poverty|prattle|prettying|prickle|pride|prudence|puddling|pump|punnet|purse|quabble|quarrel|quire|quiver|rabble|radiance|raffle|raft|rafter|rag|rainbow|rake|rangale|range|rayful|ream|reel|regiment|rhumba|richesse|ring|roll|romp|rookery|roost|rope|rouleau|round|rout|route|row|royalty|rumble|rump|rumpus|run|rush|salvo|sarcasm|sault|scatter|school|scold|scorn|scourge|screech|scurry|sea|sect|sedge|sequitur|series|serving|set|setting|sheaf|shelf|shimmer|shitload|shoal|shower|shrewdness|shuffle|siege|singular|sizzle|skein|skirl|skulk|slate|sleuth|slew|slither|sloth|smack|snarl|snatch|sneak|sord|sounder|soviet|sowse|span|spawn|spinney|spring|sprinkle|squad|squadron|stable|stack|staff|stage|stalk|stand|staple|stare|state|stench|stick|stock|storytelling|streak|stream|string|stud|suit|suite|superfluity|sute|swarm|swirl|tassel|team|tenement|thought|threatening|thunder|tiding|tittering|toil|tok|torment|totter|tower|trace|train|trembling|tribe|trimming|trip|troop|troubling|troupe|truss|tuft|tumult|turn|ubiquity|unkindness|venue|vineyard|volery|wad|waddle|wake|walk|warren|watch|wealth|wedge|weyr|wheel|whiteness|whoop|wing|wisdom|wisp|wolfpack|wrack|wreath|yap|yoke|zap|zeal|zoo"][hw="of"][pos="NN.*"]{1,2}
COMPARATIVE 比较级	[pos="JJR|RGR|RRR"]
COMPLEX_NOUN PHRASE 复杂名词短语	[pos="J.*"]{1,5}[pos="NN.*"]
CONDITIONAL 条件句	[hw="if|unless"]
CONNECTOR 连接词	[pos="I.*|CS|CC"]

CONTRACTION 缩写形式	[][word="'(s|re|ve|d|m|em|ll)|n't" pos="[^G].*"]
DEGREE_ADVERB 程度副词	[hw="very|really|too|quite|exactly|right|pretty|real|more|relatively" pos="R.*"]
DETERMINER 限定词	[pos="D.*"]
DO DO 动词	[hw="do" pos="VV.*"]
DOWNTONER 减弱	[hw="almost|barely|hardly|merely|mildly|nearly|only|partially|partly|practically|scarcely|slightly|somewhat"]
EXISTENTIAL_THERE There 表存在	[pos="EX"]
GERUND 动名词	[hw="(?!(.*thing|evening|morning|viking)).{2,}ing" pos="NN[12]"]
HAVE HAVE 动词	[pos="VH.*"]
INFINITIVE 不定式	[pos="TO"][pos="V.*"]
HYPHENATED_WORD 含连字符的单词	[word=".*-.*"]
INDEFINITE_PRONOUN 不定代词	[hw="anybody|anyone|anything|everybody|everyone|everything|nobody|none|nothing|nowhere|somebody|someone|something"]
INTERJECTION 感叹词	[pos="UH"]
LINKING_ADVERB 连接副词	[hw="then|so|anyway|though|however|e\.?g\.?|i\.?e\.?|therefore|thus|nevertheless|nonetheless" pos="R.*"]
LONG_WORD 长词	[word=".{15,}"]
MODAL 情态动词	[pos="MD"]
NEGATION 否定	[word="not|.*n't|no|neither|nowhere|never|nor|none|nobody|nothing"]
NOMINALIZATION	[word=".{3,}(tion|tions|ment|ments|ness|nesses|ity|ities)"]

名词化	
NOUN 名词	[pos="N.*"]
NUMBER 数字	[pos="M.*"]
PARTICLE 小品词	[pos="RP"]
PASSIVE 被动语态	[pos="VB[^0].*"][pos="R.*"]{0,3}[pos="V.N"]
PAST_TENSE 过去时态	[pos="V.D.?"]
PAST_PARTICIPLE 过去分词	[pos="V.N"]
PERFECT_INFINITIVE 不定式完成式	[pos="TO"][pos="VH.*"][pos="V.N"]
PHRASAL_VERB 动词短语	[pos="VV."][pos="PP.*"]{0,1}[pos="RP"]
PLACE_ADVERB 地点副词	[hw="aboard|above|abroad|across|ahead|alongside|around|ashore|astern|away|behind|below|beneath|beside|downhill|downstairs|downstream|east|far|hereabouts|indoors|inland|inshore|inside|locally|near|nearby|north|nowhere|outdoors|outside|overboard|overland|overseas|south|underfoot|underneath|uphill|upstairs|upstream|west"]
PREPOSITIONAL_PHRASE 介词短语	[pos="I.*|CS"][pos="J.*|PP.*|CC|D.*|RR|M.*|GE|N.*"]{0,5}[pos="N.*"]
PRESENT_PARTICIPLE 现在分词	[pos="V.GK?"]
PRESENT_TENSE 现在时	[pos="V.Z"]
PRONOUN 代词	[pos="P.*"]
PROPER_NOUN	[pos="NP.*"]
REFLEXIVE_PRONOUN 反身代词	[hw=".*sel(f|ves)" pos="P.X."]
SHORT_WORD 短词	[word=".{1,3}"]
SPLIT_INFINITIVE	[pos="TO"][pos="R.*"][pos="V.*"]

分裂不定式	
SUPERLATIVE 最高级	[pos="DAT|JJT|RGT|RRT"]
SWEARWORDS 脏话	[hw="arse|arsehole|bastard|bellend|bint|bitch|bloodclaat|bloody|bollocks|bugger|bullshit|clunge|cock|crap|cunt|damn|dick|dickhead|fanny|feck|fuck.*|gash|git|god|goddam|jesus|minge|minger|motherfucker|munter|piss|prick|punani|pussy|shit|sod|tit|twat"]
TIME_ADVERB 时间副词	[hw="afterwards?|again|earlier|early|eventually|formerly|immediately|initially|instantly|late|lately|later|momentarily|now|nowadays|once|originally|presently|previously|recently|shortly|simultaneously|soon|subsequently|today|tomorrow|tonight|yesterday"]
VERB 动词	[pos="V.*"]
PEOPLE 人	[sem="S2|S2:1|S2:2|S3|S3:1|S3:2|S4"]
MALE 男性	[sem="S2:2"]
FEMALE 女性	[sem="S2:1"]
SUPERNATURAL 超自然	[sem="S9"]
EMOTION 情感	[sem="E|E1|E2|E3|E4|E4:1|E4:2|E5|E6"]
TIME 时间	[sem="T1|T1:1|T1:1:1|T1:1:2|T1:2|T1:3|T2|T3|T4"]
PLANET 行星	[sem="W1|W2|W3|W4|W5|L1|L2|L3"]
COLOR 颜色	[sem="O4:3"]
COLOUR 颜色	[sem="O4:3"]
BODY 身体	[sem="B1|B2|B3"]
FOOD 食物	[sem="F1|F2"]
TECHNOLOGY 技术	[sem="Y1|Y2"]
MEDIA 媒体	[sem="Q4|Q4:1|Q4:2|Q4:3|K1|K2|K3|K4"]

10 术语

绝对频数（或原始频数）Absolute (or raw) frequency –指某一语言特征在整个语料库或其部分中出现的次数，或检索项在语料库的出现次数。

类联接 Colligation –指在文本中统计识别到的语法范畴（如词性标注）系统性共现。

搭配词 Collocate –指与节点（所检索的词或短语，即检索项）系统性出现的词。

词语搭配 Collocation –指在文本中统计识别到的词语系统性共现。

索引行 Concordance line –指在 KWIC（语境中的关键词）表中的单个行，通常包含节点（检索匹配项）及节点前后的若干词（左右侧语境）。

索引 Concordance –指一种显示在语料库中检索到的语言使用实例的常见形式，通常将节点（检索匹配项）位于中间，显示左右侧构成其语境的若干词。索引有时亦被称为“KWIC（display）”，显示语境中的关键词。

语料库 Corpus（复数为 corpora）–指可被计算机检索的语言数据集合。

频数 Frequency –指在语料库中检索项匹配文本的出现次数。注意区分绝对频数（单纯的出现次数）和相对频数（每若干数量词中该词的出现次数）。

KWIC- “keyword in context” 的缩写。为一种显示语料库中检索到的实例的常见形式，其中节点（所检索的词或短语）位于中间，显示其左则和右侧语境的若干词。KWIC 有时亦称“索引”。

左侧语境 Left context –指位于特定检索匹配项（节点）之前的词。左侧语境的各个位置分别被称为 L1（紧邻节点之前的位置）、L2、L3 等。

词目 / 词头 Lemma / Headword –指属于同一个词根的所有屈折变化形式。例如，词目 “go”包括以下词形（类符）：“go”，“goes”，“went”，“going” 和“gone”。

节点 Node –指所检索的单词、短语或语法结构；检索项匹配的文本。

词性（POS）Part-of-speech –指语法范畴、词类。自动标注词性的过程常被称为词性标注（见下文）。

词性标注（POS 标注）Part-of-speech tagging –指向文本或语料库中每个单词添加有关其语法范畴信息的过程。例如以下例句词性已标注：自动_RB 为_IN 数据_NNS 标注_VBZ 词性_NN。

正则表达式 Regular expressions (regex) –是一种特殊的元语言，允许高水平用户同时检索多个字符串。

相对频数（或标准化频率）Relative (or normalized) frequency (RF) –是检索词的绝对频数除以被检索的总词数（语料库或子语料库的总词数）。频数的标准化通常会乘以一个适当的基数（例如 10,000）。

右侧语境 Right context –指位于特定检索匹配项（节点）之后的词。右侧语境中的各个位置分别被称为 R1（紧跟节点后的位置），R2，R3 等。

子语料库 Subcorpus（复数为 subcorpora）–指语料库中被用户限定检索的部分，可包括整个文本或多个文本的部分。在 LancsBox 的 X 版本，子语料库使用 XML 结构以定义。

标注 Tagging -自动或半自动地向文本或语料库中的词语添加语言信息的过程。参见词性标注。

文本 Text - 语料库的基本单位; 语料库是多个文本的集合。

形符 Token -文本或语料库中一个词形的单个出现。

XML-可扩展标记语言的缩写，在文档中写入机器可读信息，并为其提供结构和注释。在语料库中，XML 可为单词添加词性信息，为文本提供结构信息，例如分节和段落。


Developed @ Lancaster University
Lancaster University
chemistry