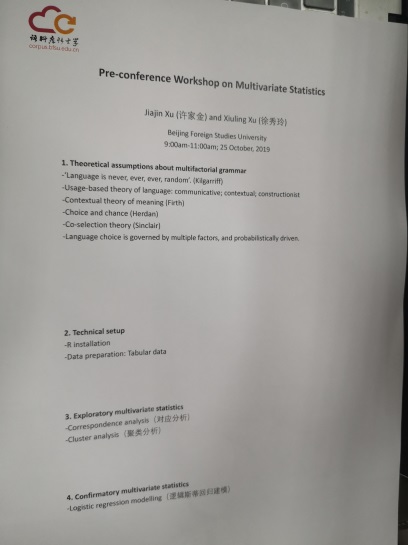
时间： 2019年10月25日星期五

地点：上海理工大学

2019年语料库大学国际大学：跨学科视角



Adam Kilgarriff 2005 langauge is never, ever

介绍1945年的书，强调语境的关注，Fires Contexual orientation

做系统的观察和记录

Multifactorial Analysis in Corpus Linguistics

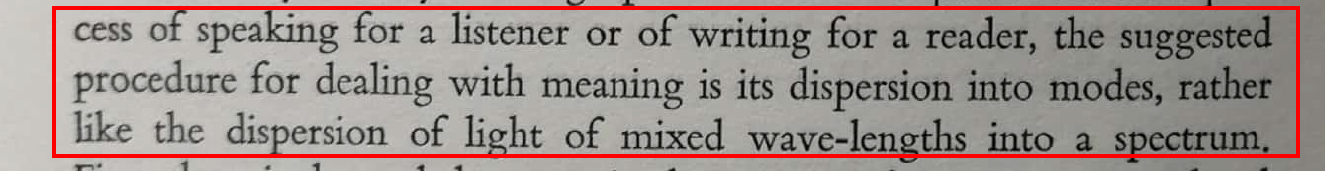
统计分析支撑语料库/语言研究

伦敦学派，Firth（弗斯）, 韩礼德。

Firth: Modes

you should know a word by the company keeps.

you should know a man by the company keeps.



Firth建议；话语事件，language event, 从高到低， descending order.

从低到高的层级排练

BerylT.Atkins 是John sinklaire的姐姐。

Beryl Atkins 1987 Semantic ID tages

相信文本： Trust the text

第二章： The search for units of meaning,

语料库语言学关注意义.

举例：naked eye。用法、含义。意思不确定，需要去扩展，意思落到短语上。

扩展意义单位，

Collocation colligation semantic preferences semantic prosody.

语料库关于意义的研究。

研究学术文本，学术话语研究，学术分学科，hyland。 研究特色这些短语起承转合，在不同的学科是否不同，跟学科是否有关联，人文，理工科。

从语言要看到相互关联的属性。

把语言理解为选择，什么情况选择什么样的词？

Language choice is governed by multiple factors, and probabilistically driven

Choice and change(Herdan,1960s)很早就有

Co-selection theory(Sinclair)

Contextual theory of meaning(Firth)

2.Technical setup

R installation

Data preparation: Tabular data

开源，自己专业的软件包？怎么查语言学的包

包1

CA 简单对应分析，需要对应

HAC 分层聚类分析（统称聚类分析）

LRM\_logistic regression modelling逻辑斯蒂回归建模（逻辑回归模型）

Log=对数

压缩包第一个文件

常用多元

*Corpus Methonds for semantics 2014*做意义研究。

卡方/Fisher, 单因素，一元分析,univariate

Collostructional 构式搭配 multivariate

Binary logistic regression，二选

动词不定式

成系列的方法。

软件里面的函数（lrm）

这三页纸，必须要学会。

软件操作：

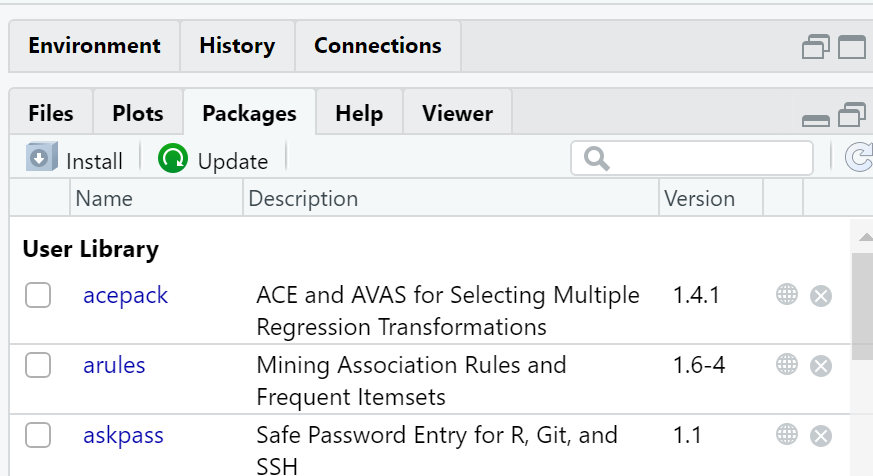
04： 安装。

MAC版本

先安装R，然后按照R-Sduido

语言研究根据自己

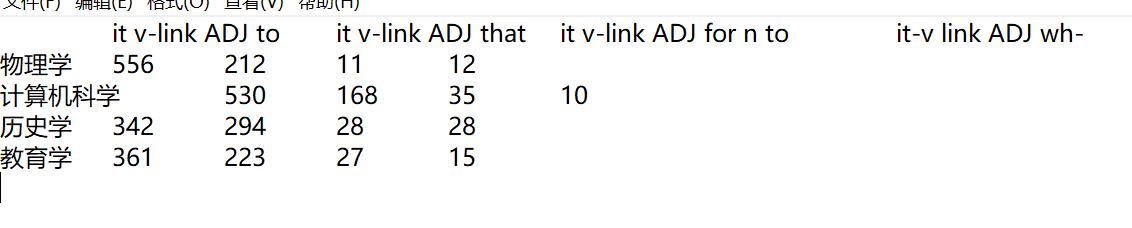
Plot命令很重要



离线安装

R专业的网站：CRAN, 看到别人书上用到什么就用什么。

Install.packages(“ca”) 安装命令



卫乃兴，2019：认识论、学科文化

对应分析

学科特征、语义特征

历史学

软学科/硬学科

评述，通过检索。

正则表达式

基础数据？

绘制data1数据

Tap箭头

文档-tap-excel

Excel是dataframe, 有行有列，常见的是行列式。

简单对应分析。 特征和学科，特征和语体的分析。

install.packages("ca")

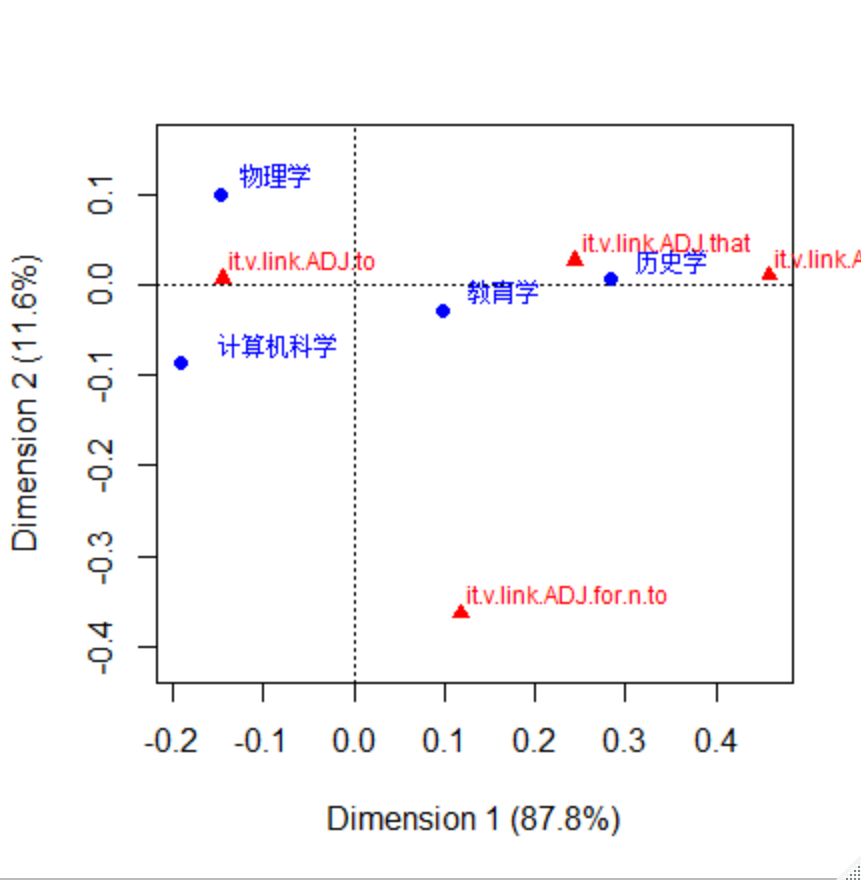
library(ca) #调用“ca”

data1<-read.table(file.choose(),header= TRUE, sep="\t", row.names= 1) #读取数据表

ca.data1<-ca(data1)

plot(ca.data1)

#header=True,第一行是表示文件抬头， sep="t",表示用tap按键隔开的，row.names=1表示第一列为名称



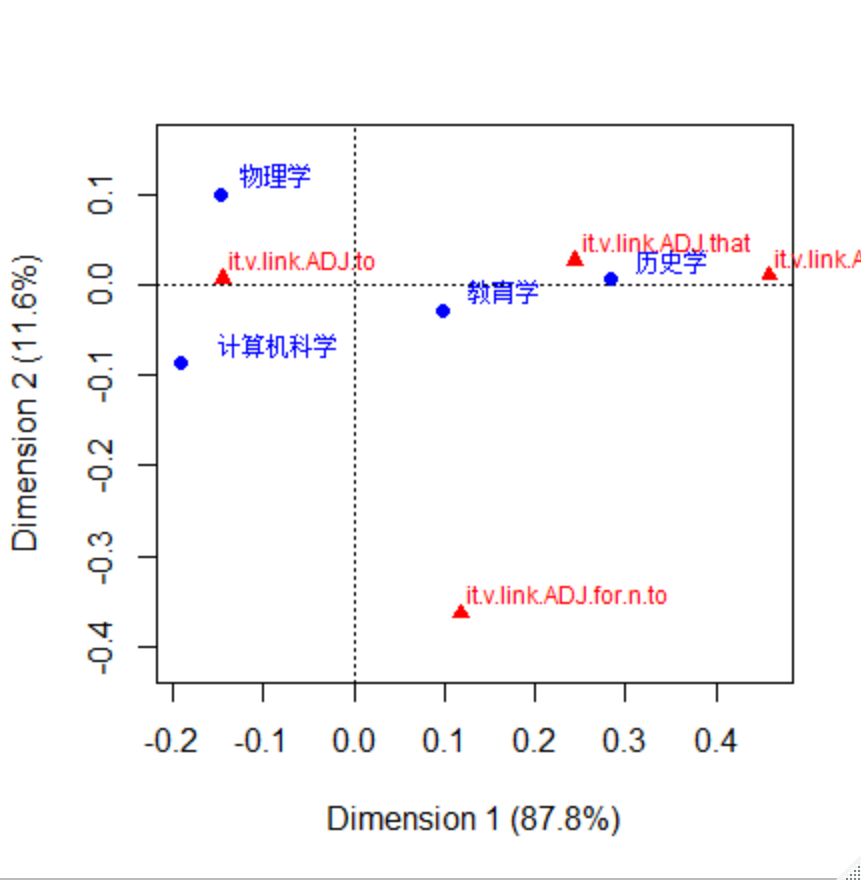
句式表达

探索空间距离。

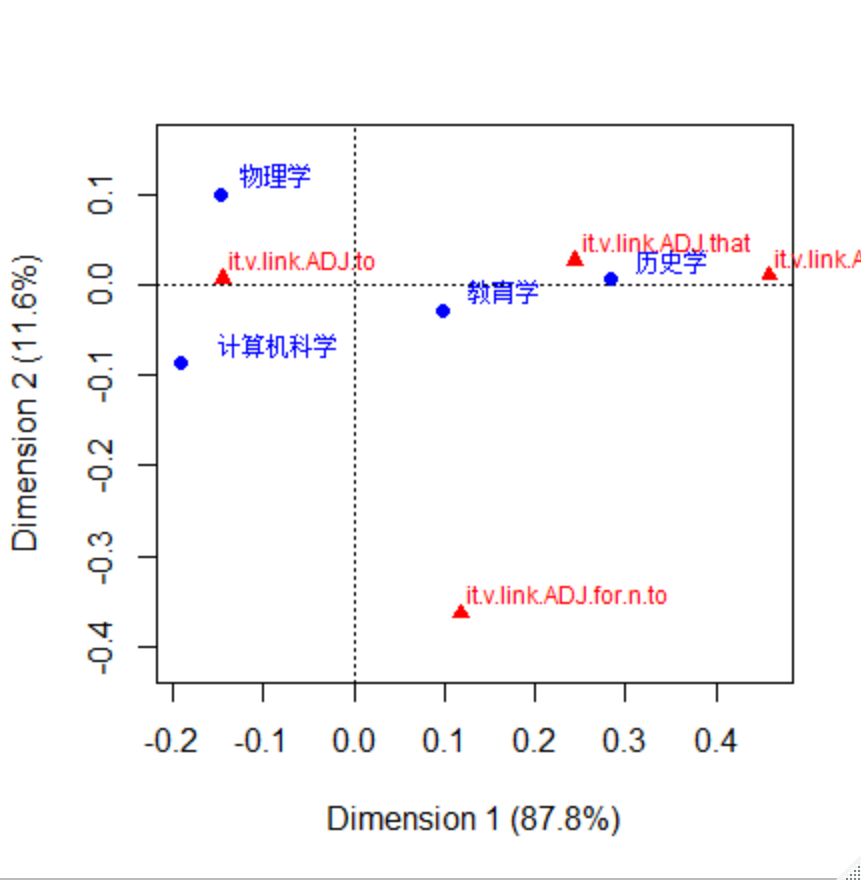
装包“pvclust”

数据二 教材难度句法分析

关系通过树形结构，红色表的是显著性。



教材的难度



install.packages("pvclust")

library(pvclust)

data2<-read.table(file=file.choose(), sep = "\t",header = TRUE, row.names= 1)

data2

data2.pvc <- pvclust(data2, method.hclust = "ward.D2", method.dist = "canberra", nboot = 1000)

plot(data2.pvc, hang = -1)

pvrect(data2.pvc, alpha = 0.95)

That/zero

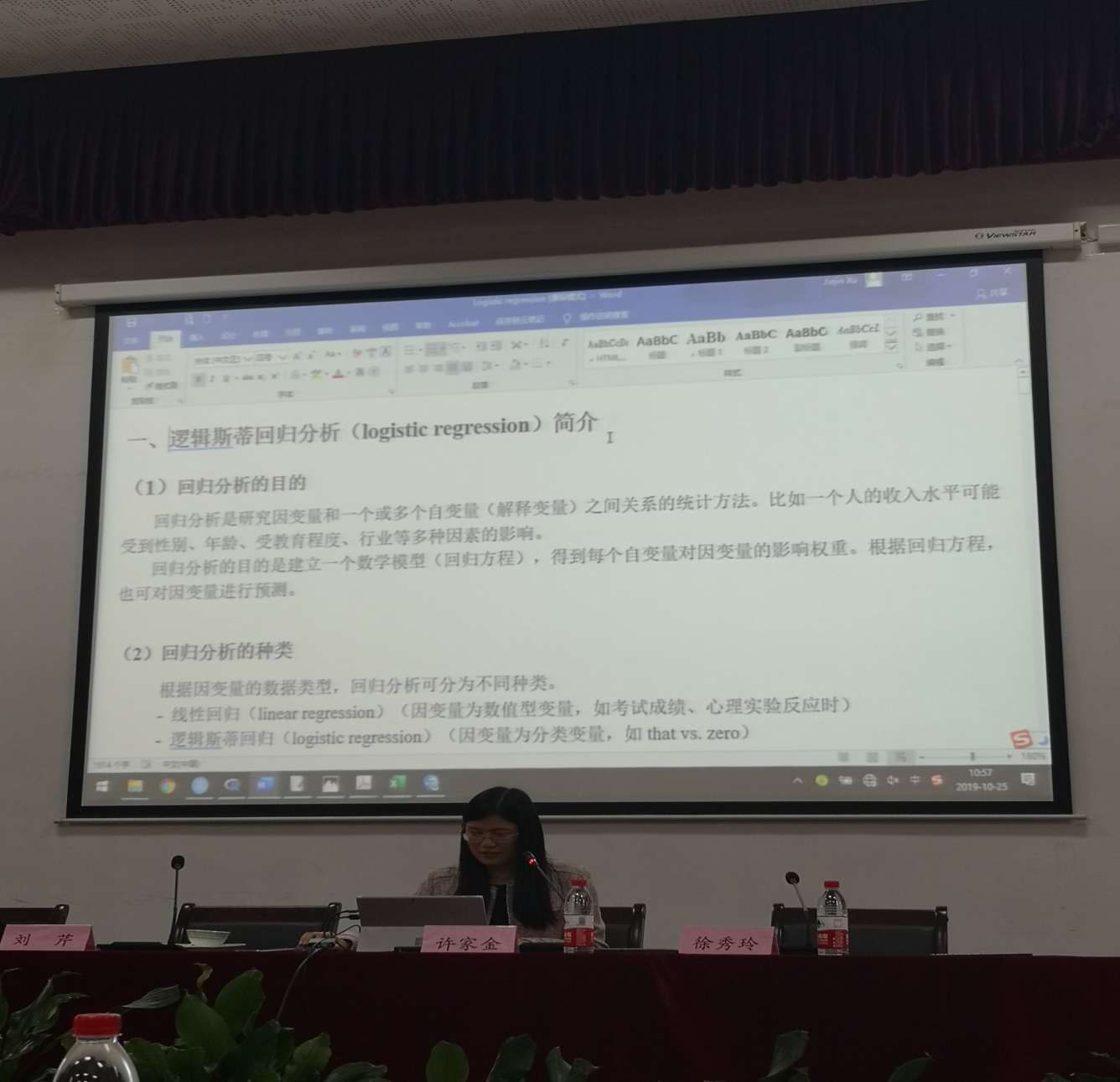
1可比语境下关系代词取舍的多因素分析，芳印杰

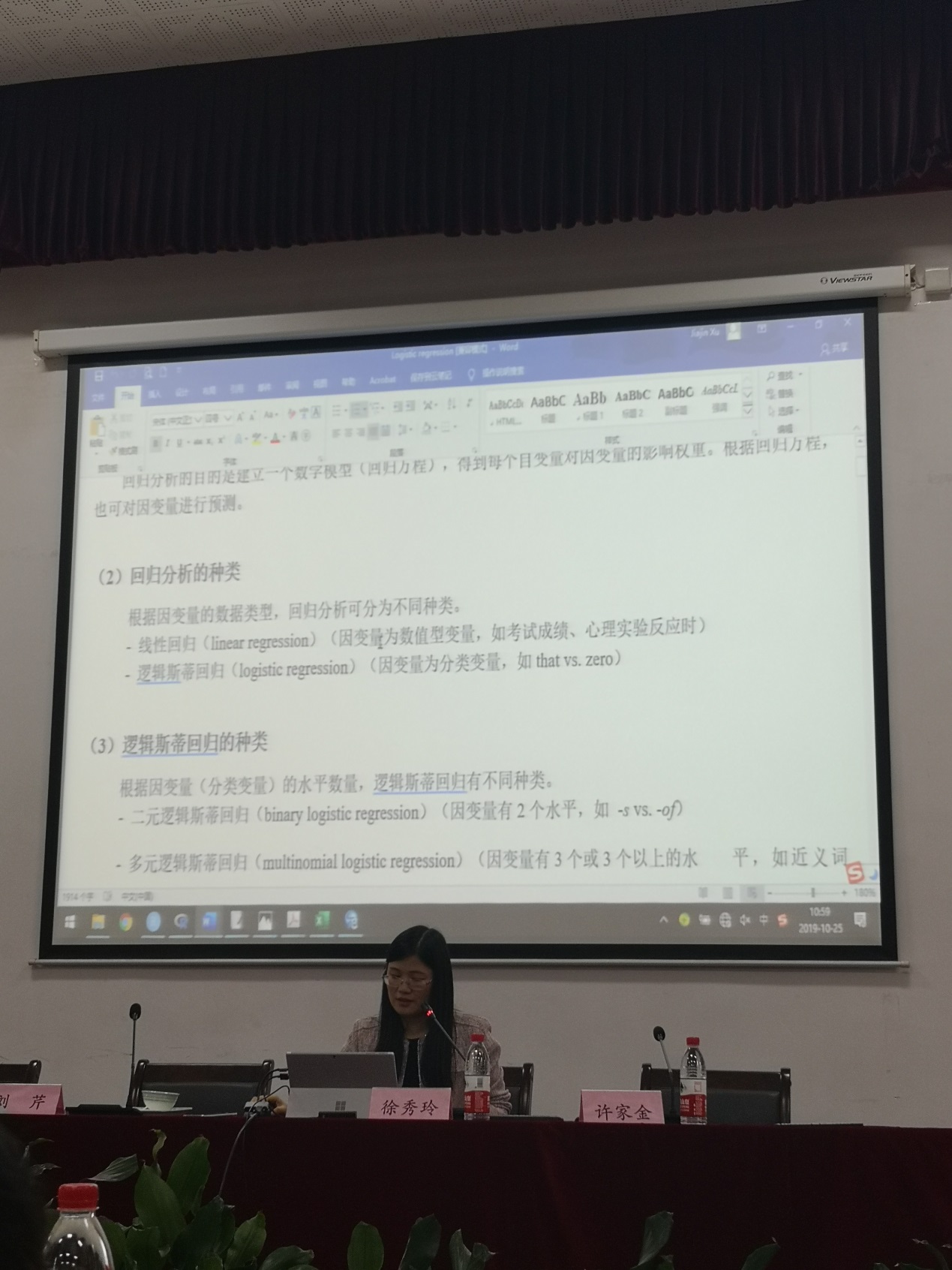
2英汉与格交替现象的多因素研究，张懂（博士论文）

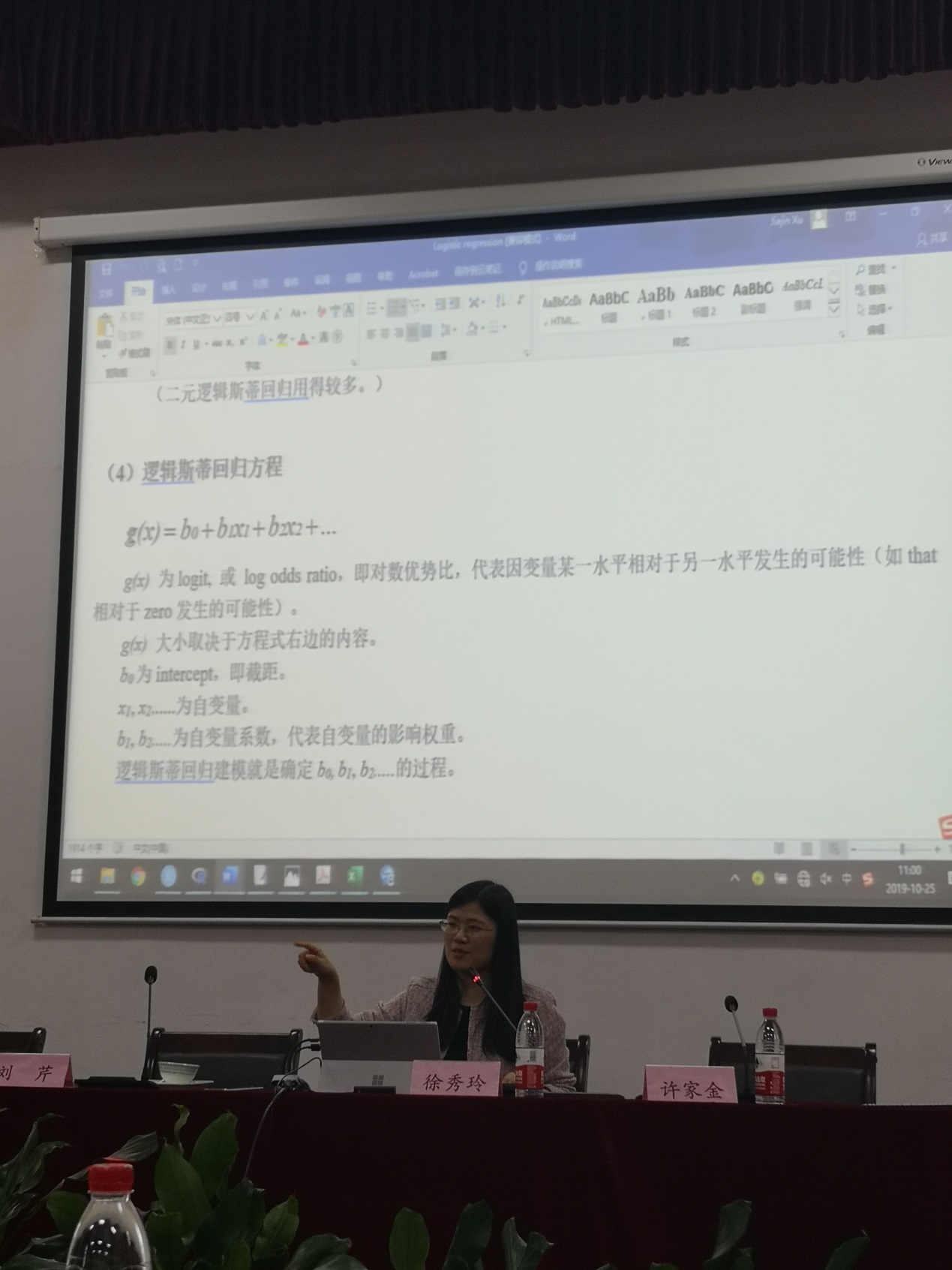
Give a book to me/give to me a book. 英汉对比

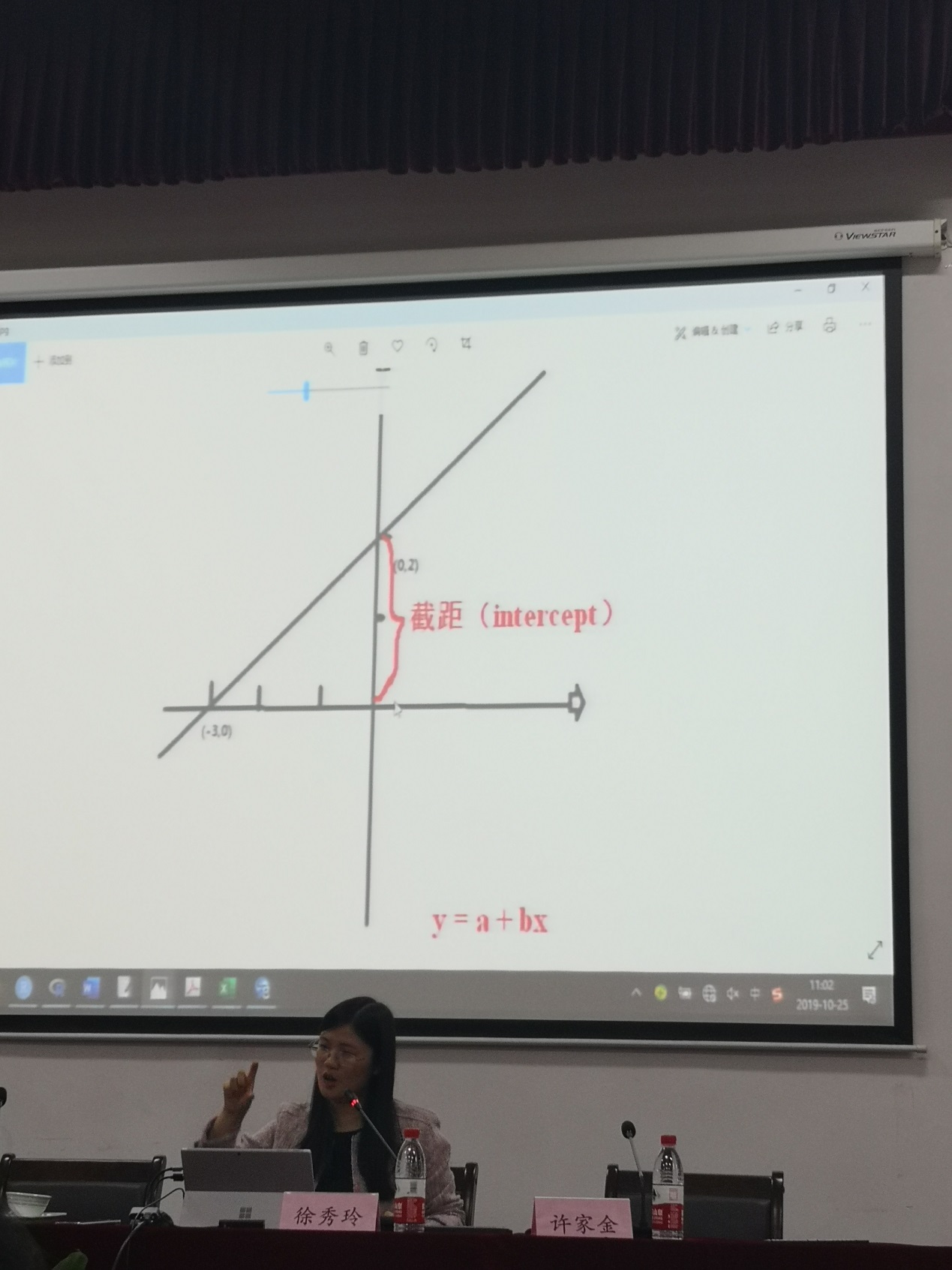
有的情况要用双宾，有的用

逻辑斯蒂，到底什么是最重要的。











1．逻辑斯蒂回归的原理

2．案例分析

1. 回归分析

探讨自变量和因变量的关系，目的是建立一个方程：

Y = kx + b

线性回归（因变量为数值型变量，如考试成绩、心理实验反应时）

逻辑斯蒂回归（因变量是分类变量，如that vs zero）

逻辑斯蒂回归的种类（二元、多元）

根据因变量（分类变量）的水平数量

G(x) = b0 + b1x1+b2xx

B1,b2是影响因素，B0是截距，

参考文献Levshina(2015)第12章

左边是因变量，右边是字变量

第三部分代码

install.packages("MASS")

library(MASS)

data3<-read.table(file.choose(),sep="\t", header= TRUE)

head(data3) #查看数据

str(data3) #告诉数据的结构

summary(data3)#描述统计

m1<-glm(Aux~Causation + EPTrans + Country, data = data3, family = binomial)

summary(m1) #描述逻辑斯蒂回归的总结

levels(data3$Aux) #因变量的水平，laten相对于doen的概率

m1<-glm(Aux~Causation + EPTrans + Country, data = data3, family = binomial)

因变量Aux，自变量Causation + EPTrans + Country, binomil二元

estimate系数，标准误，Z value, Pr

Coefficients:

Estimate Std. Error z value

(Intercept) -2.5715 0.3842 -6.693

CausationInducive 3.3725 0.3741 9.015

CausationPhysical -0.4661 0.6275 -0.743

CausationVolitional 3.7373 0.4278 8.735

EPTransTr 1.2952 0.3394 3.816

CountryNL 0.7085 0.2841 2.494

Pr(>|z|)

(Intercept) 2.18e-11 \*\*\*

CausationInducive < 2e-16 \*\*\*

CausationPhysical 0.457575

CausationVolitional < 2e-16 \*\*\*

EPTransTr 0.000136 \*\*\*

CountryNL 0.012633 \*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimate系数， Error残差，

安装：MASS

数据的解释

Rms表示的是

多因素分析一定要做的。

逻辑斯蒂回归一定要学，标注流程。

模型好的评估，工作坊和培训按照0起点，然后找选题，找数据，怎么标注，怎么拟合检验。

好书： 2015 how to do linguistics with R.

Natalia Levshina (2015) How to do Linguistics with R #非常好的，操作步骤很详细，有配套的

R. H. Baayen (2008) Analyzing Linguistic Data #统计水平比较高

Dylan Glynn & Justyna Robinson (2014) Corpus Methods for Semantics #这本书有5-600页

#没有一本好的中文文献，语言类的，可以出一本书。

R的学习笔记

首选读英文的文献。聚类和MDS多维尺度分析，研究视角，语言变体：英语变体，张炜炜（鲁汶大学毕业）

做变体/中介语

学术英语，不同题材，做变异比较多，哪些因素影响了变体，同时在一个方程里考虑。

哪个是最重要的，可以评估处理，认知和理论

首选读英文的文献。聚类和MDS多维尺度分析，研究视角，语言变体：英语变体，张炜炜（鲁汶大学）

研究的难点：包含了很多泪水，怎么标注，标准几千句，要标注每个句子标注10几个特征。

标注+计算机

研究问题做完后的解释，这个决定就是why，有多元统计，语境因素共同制约。

前期的标准

语言学知识和高级统计做一个平衡。

语料库不仅仅是一套

语言是一个chance and choice?

自变量的初步筛查？

答：基于文献，包含最重要的变量，标注中是否有遗漏的，在标注的环节是可以补充和添加的。

如果在统计学上实现。

共现性检查：

初步筛查。

张宝林老师：全球中介语语料库可以用了

学习者用处不大，而且借鉴太多。

**李文中：Exploratory corpus research：from data to text**

浙江工商大学

* 近三年关注：探索+可视化分析+统计

文科的统计不太好。

卡方检验。

肉眼观察+大数据

Firth:语料库的目标：研究意义

Sinclaire:语言使用的意义

观点：探究性研究，从数据到文本

一、如何更好的观察、组织数据

二、对文本特征研究关系

Data preparation:

Tidying data for tidy data and looking at text from outside

R学习是一个必选项。 2007年自学R， 数据基本结构

统计分析的前提：对数据有很好的分析。

视觉语法

library(tidyverse) 清洗数据的包

library(readx1)

library(cvd)

library(vcdExtra)

library(FactotMineR)

library(factoextra)

library(data.table)

#setwd(“”)改变工作的路径

探究式研究：在还没有想法的时候，对文本的整体特征，进行分析，并做可视化。

数据准备：

It is often said that 80% of data analysis is spent

Import→ tidy→ transform→ communication

如何输入数据，如何读入数据，如何整理数据

定性数据，定量数据

类型变量，男女，高中低

定量变量：

BNC学术语料库

列联表，一个变量放在列，一个变量放在行里面。

Dataframe： 数据框

基于数据框，然后做列联表，然后做转换。

R常用的三种数据结构

1. 个例数据框 data frame of cases

最原始的表

常用R的包

tidyverse

ggplot

Xtabs:? 计算频数，

Visualize dataframe of cases

语料库的比例关系，值不值得做。

* Methods applicable

Mosaic() 马赛克图，展示多维可视化手段

Barplot()

Assoc() ： 利用残差检验

Statistics: chisq.test(), assocstats()

A data frame of counts计数数据框

把列连表编程数据

A mosaic plot

TOFEL作文，把关键信息写到文件名里面。

文件命名：Letter\_man\_1960.001.txt

R语言的文件包，专门处理文件 quanteda包，搭配、文本特征分析

查看词汇特征，最常用的，通过统计手段提取出。

主题河：Streamgraph作图

**English-Corpora.org**

主讲人：Mark Davies

介绍BYU, COCA

iWeb

新网址：语料移动走了，不在BYU了。

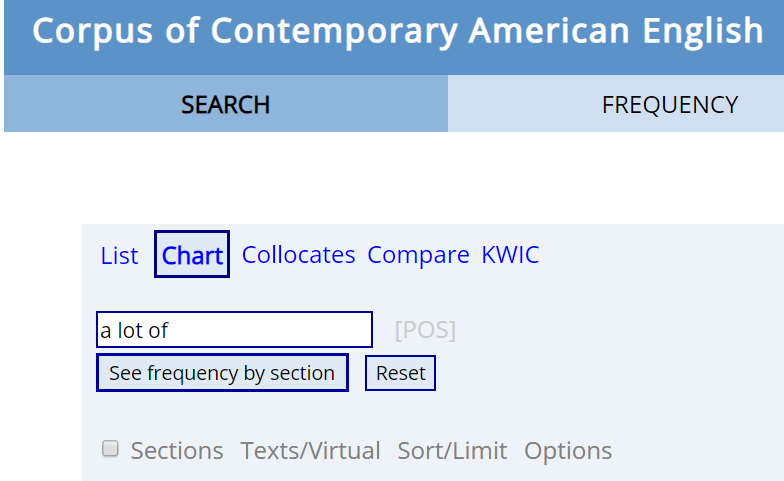
<https://www.english-corpora.org/>

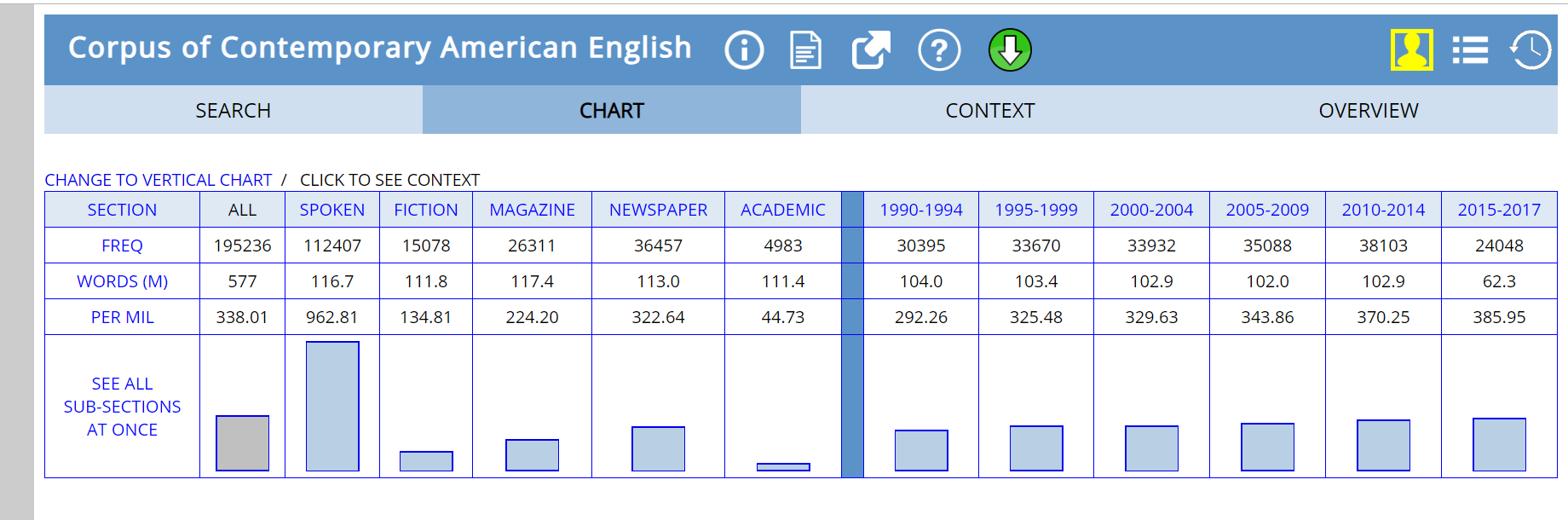
为什么创立语料库

主题：关注variation （genere difference）

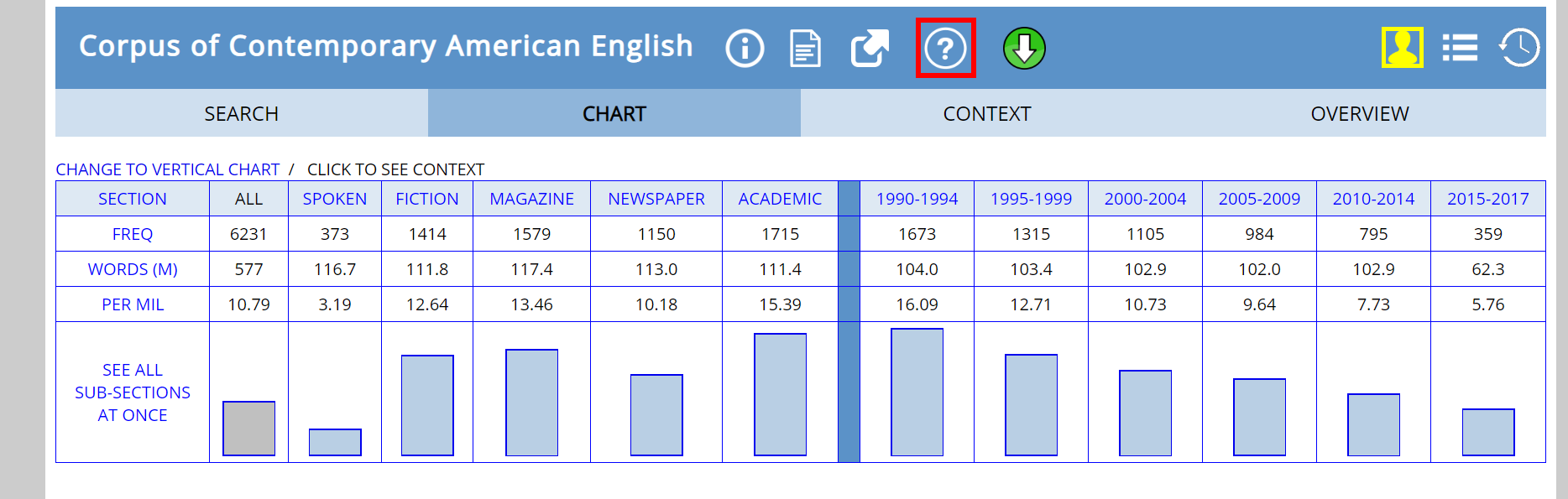
[Corpus of Contemporary American English (COCA)](https://www.english-corpora.org/coca/)

Corpus of Contemporary American English(COHA)

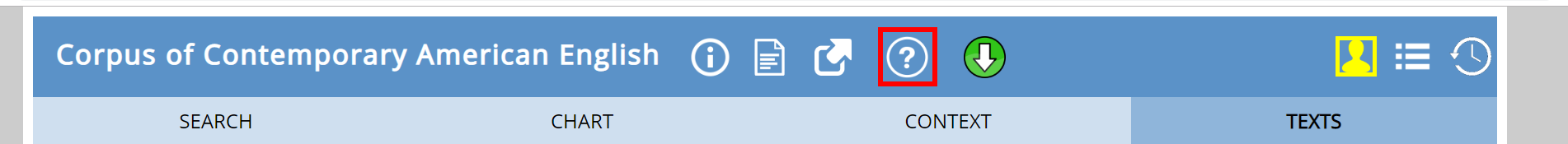




seldom



点击文本查询文本

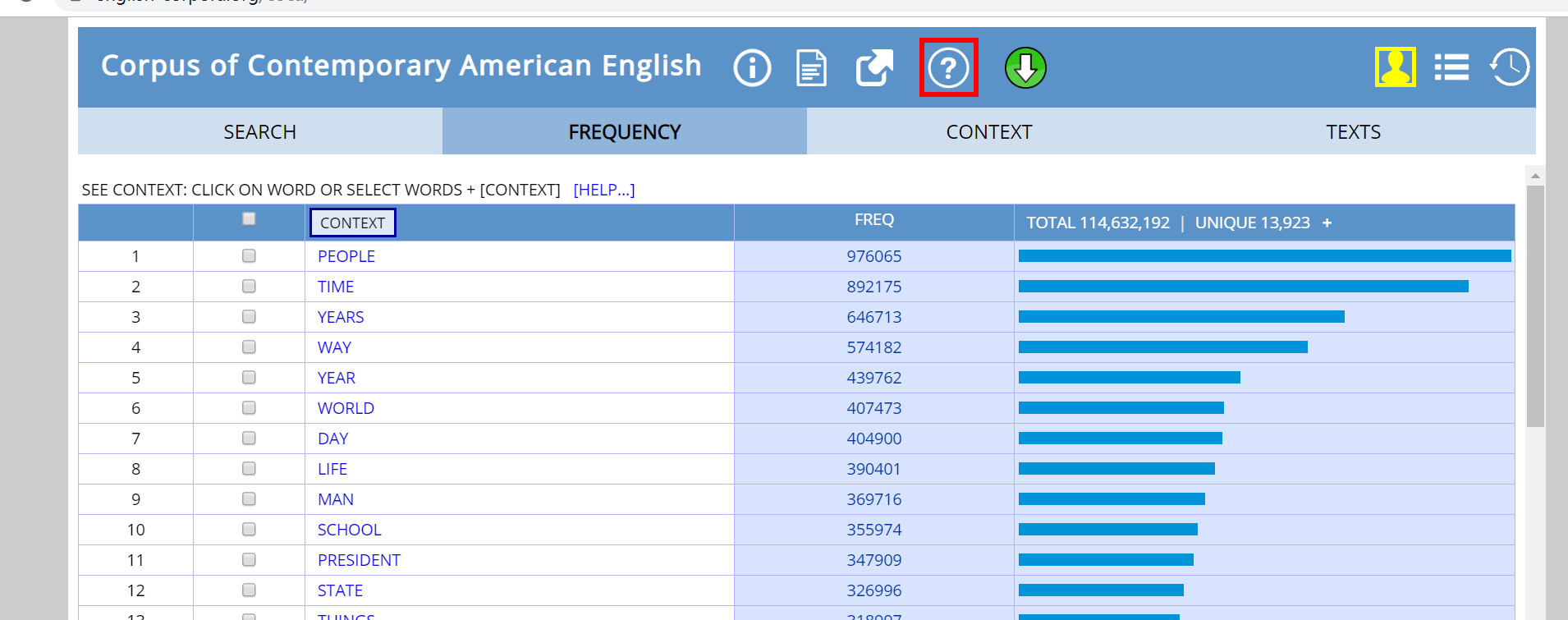


Frequency list

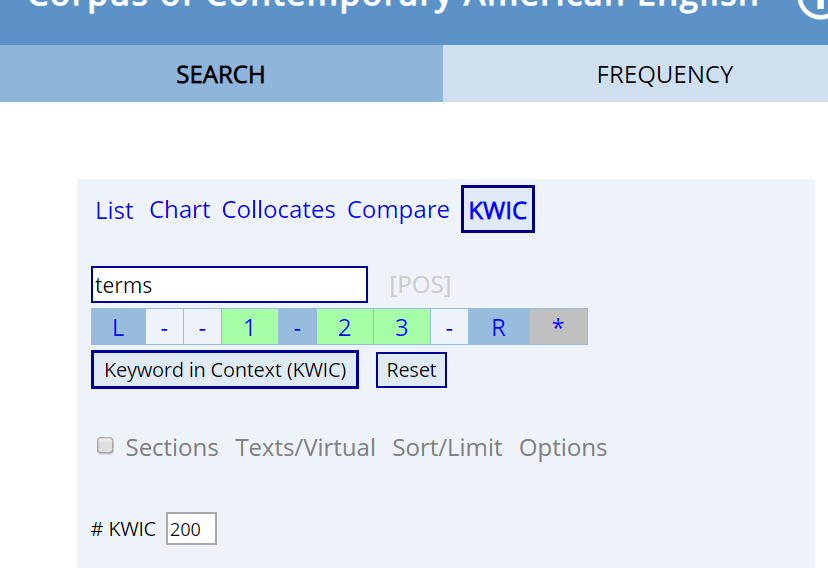
Search

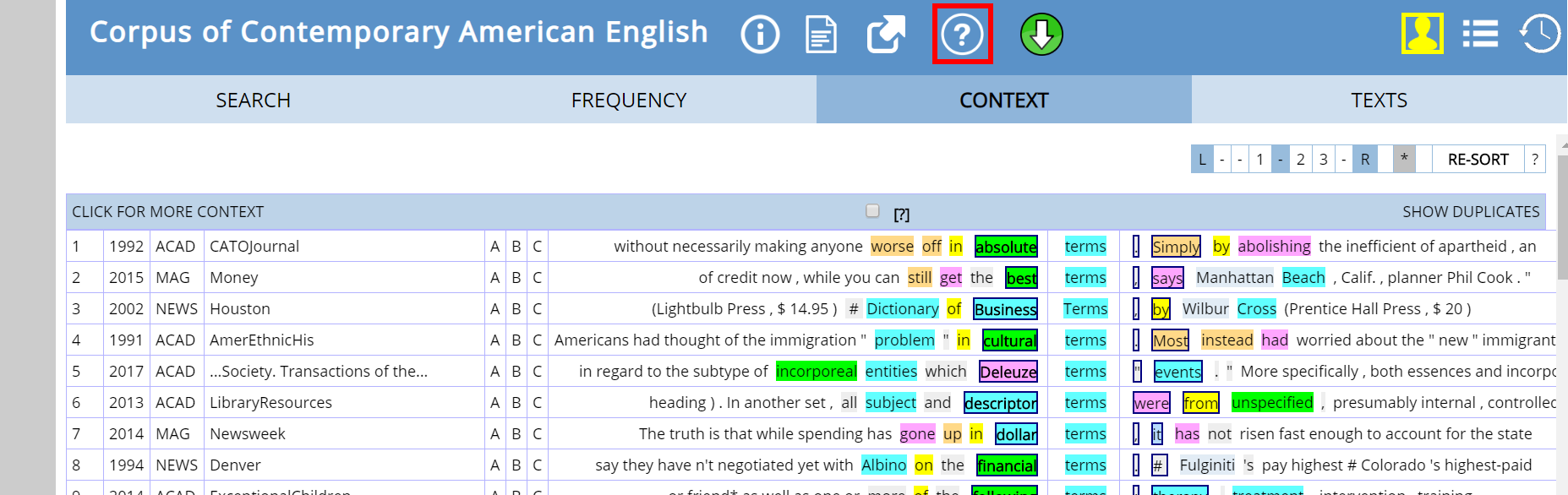
NOUN

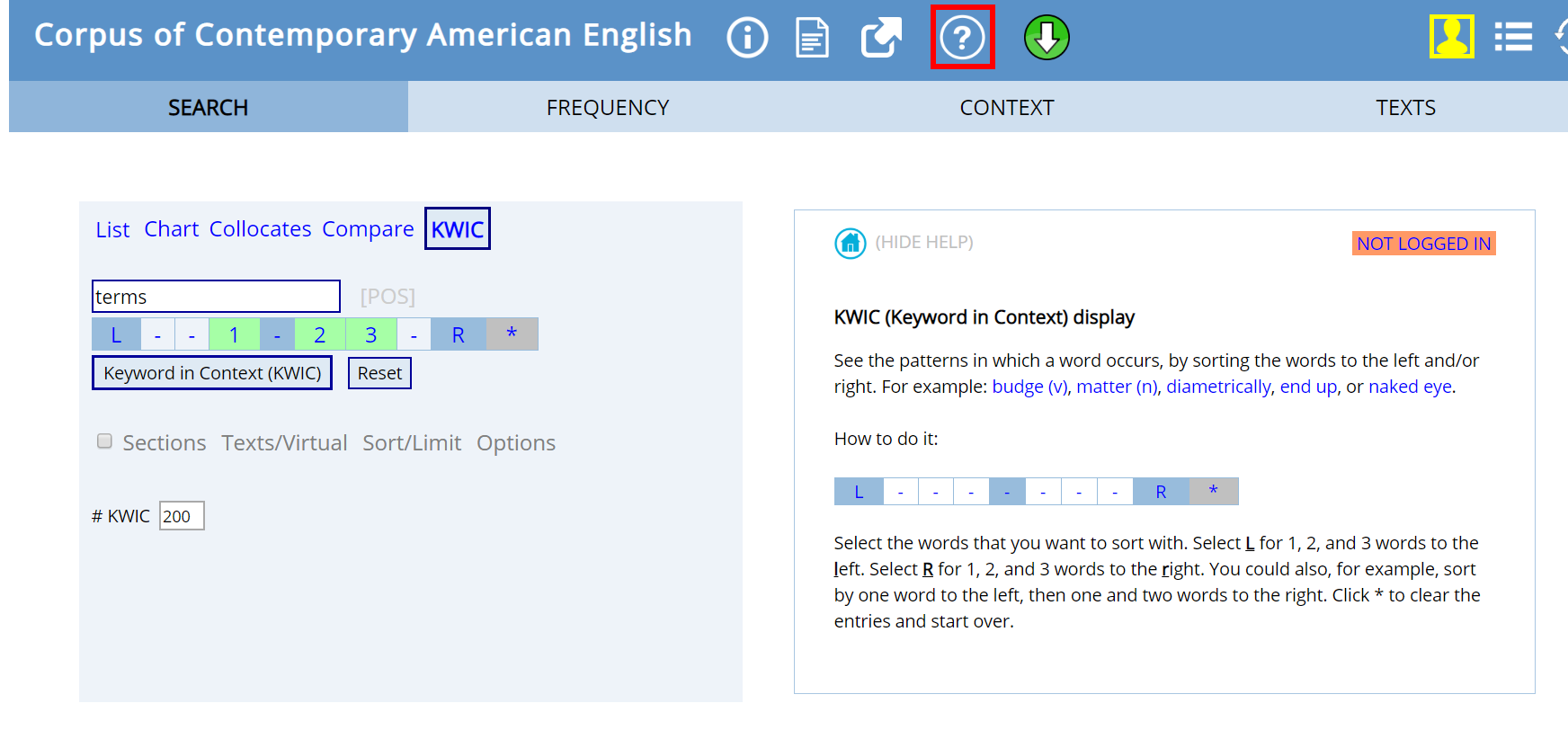


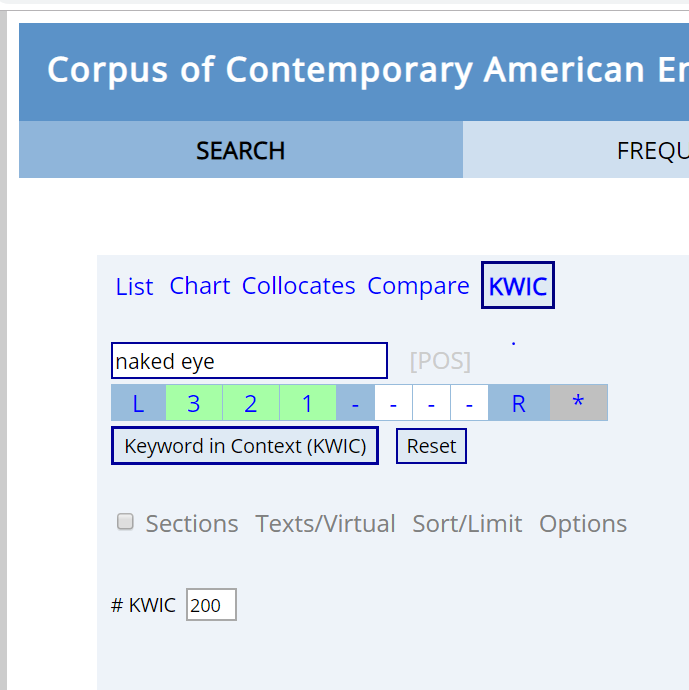


KWIC，terms左1右23

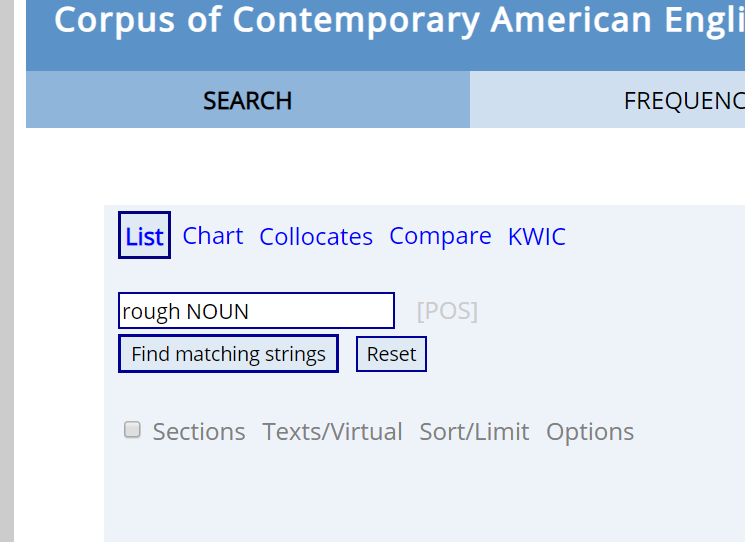




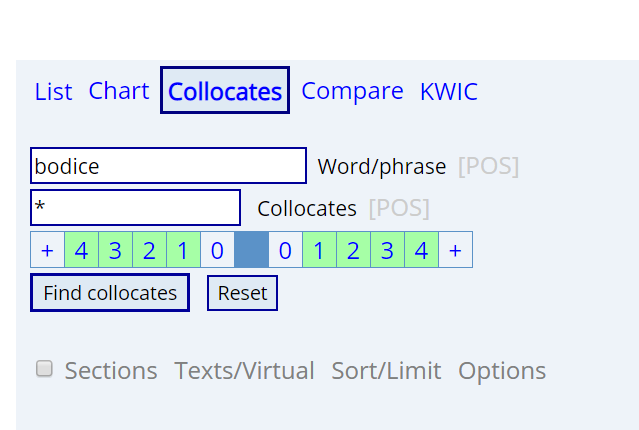




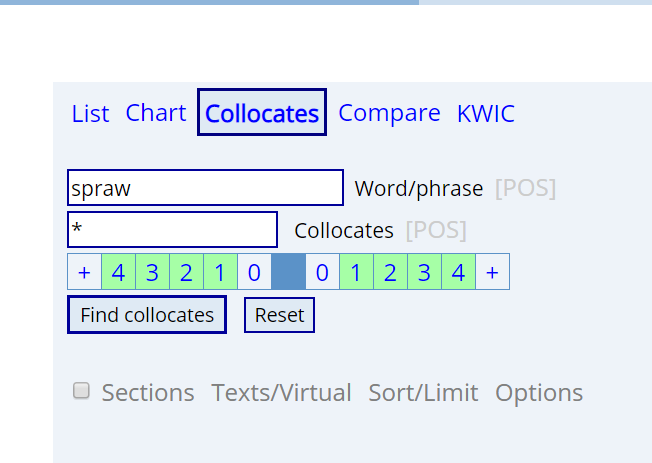
Collocate



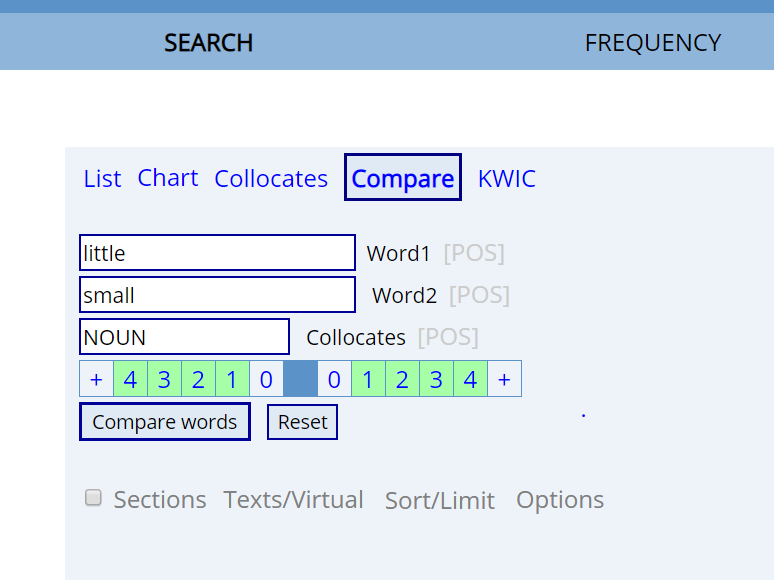
Bodice

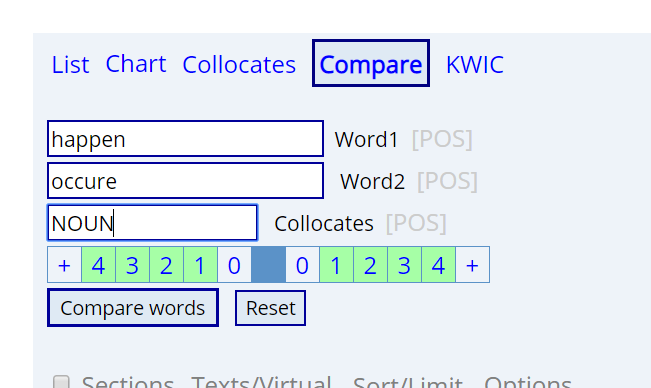


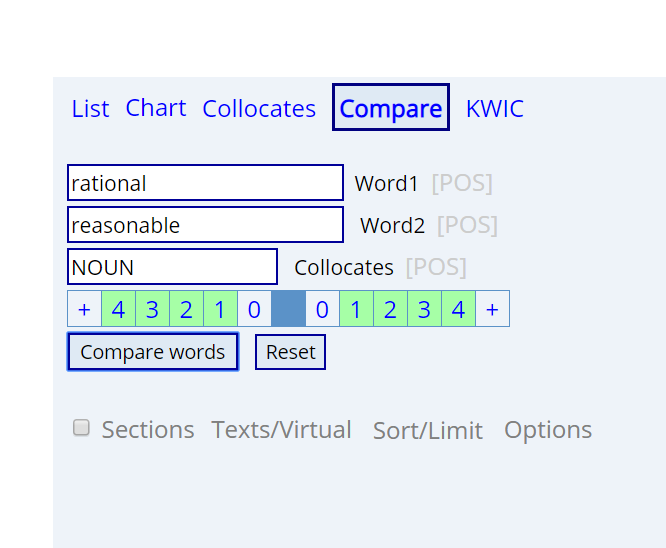
Spraw, look like a city,百度会呈现图片

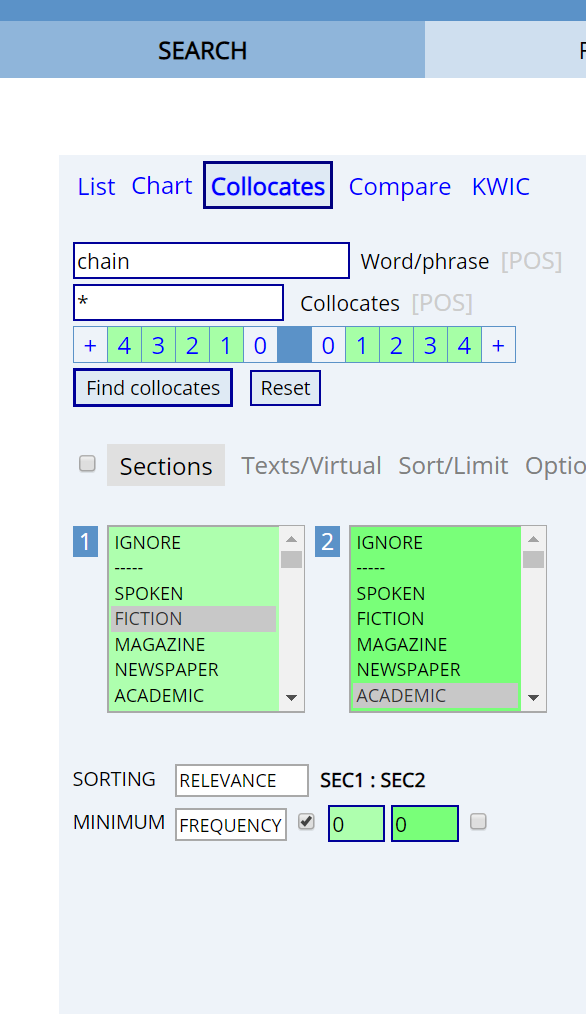


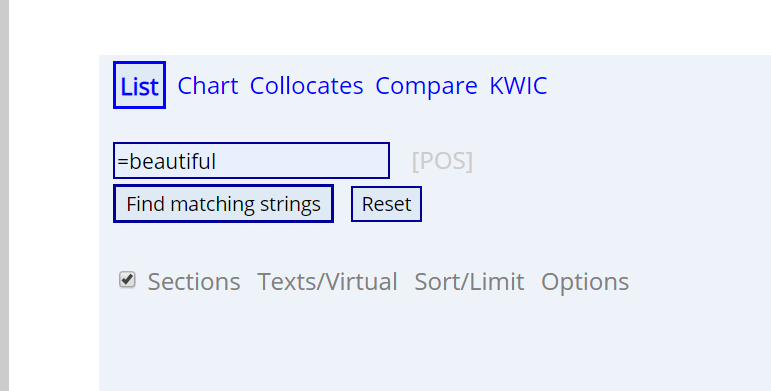
区别近义词



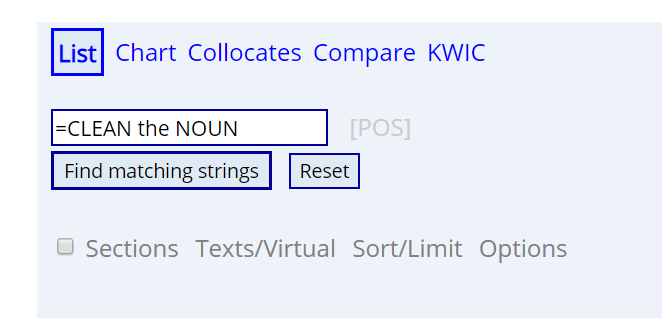


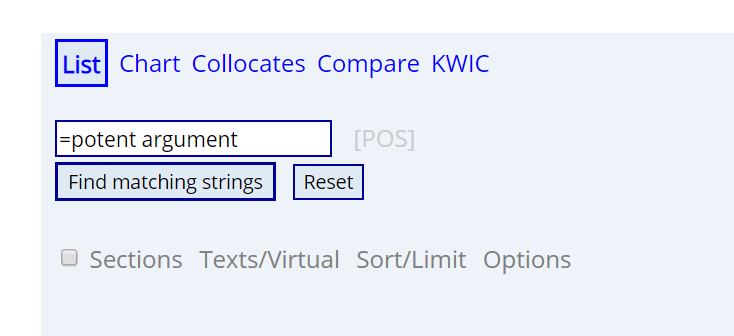






大写字母表示单词的各形式

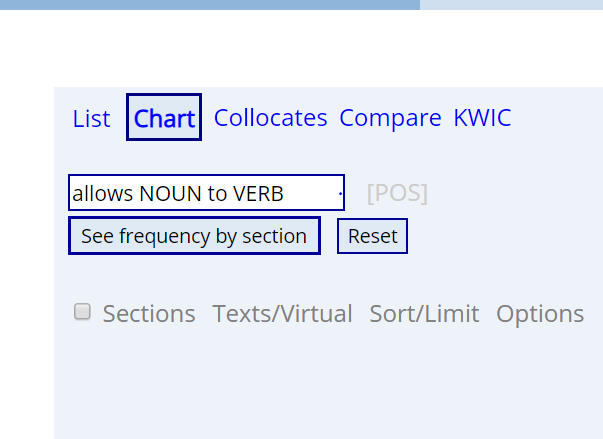


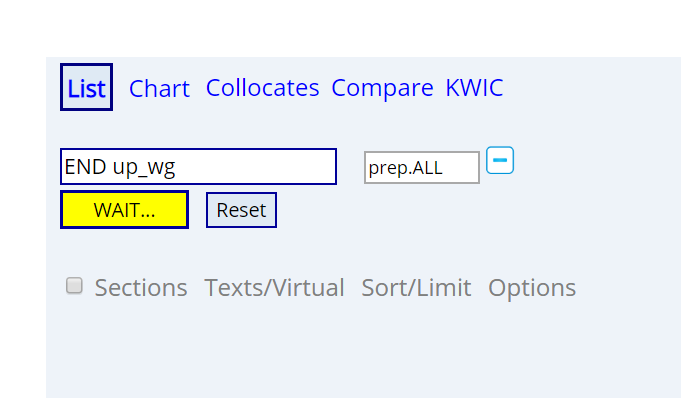


形态检索



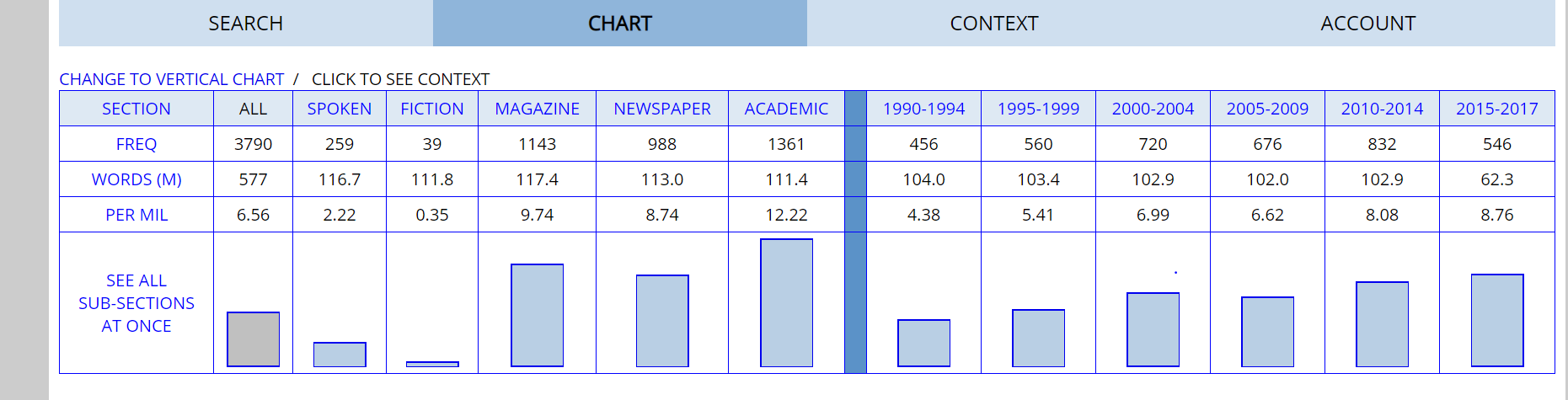


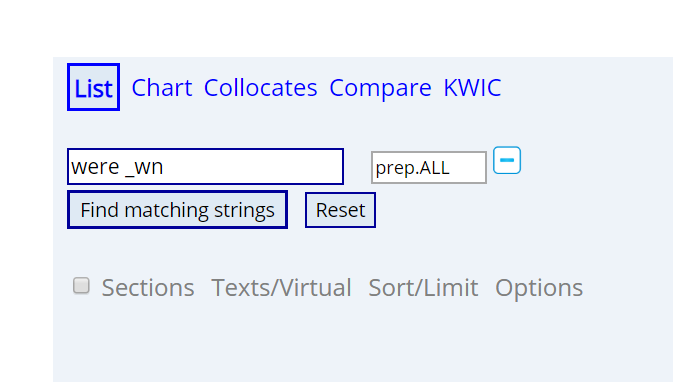






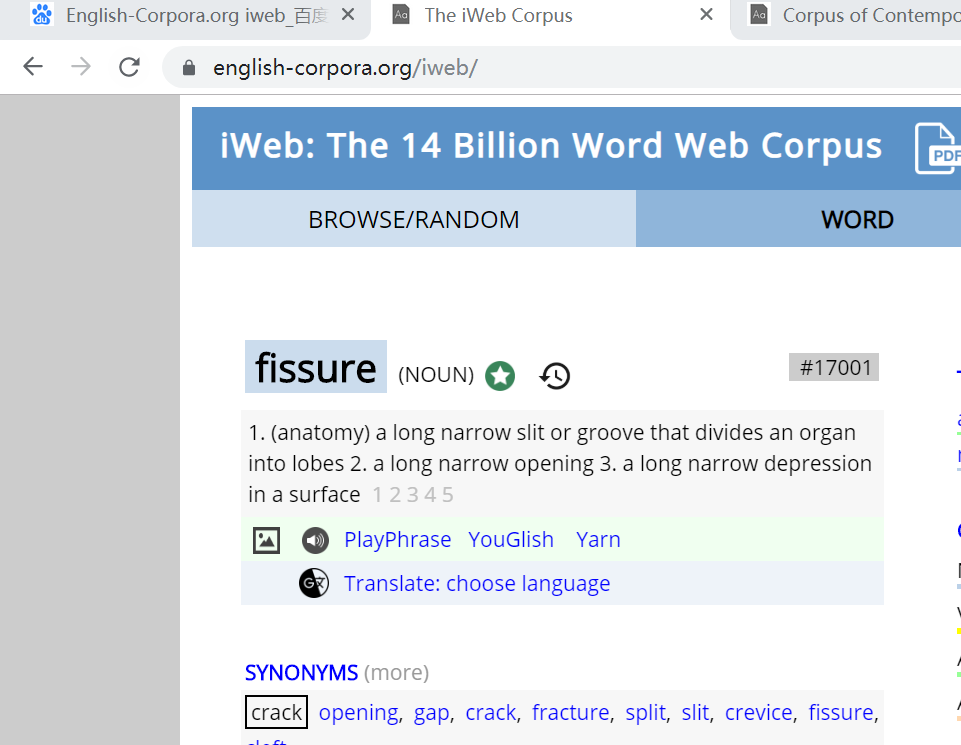
CONJ PRON BE like

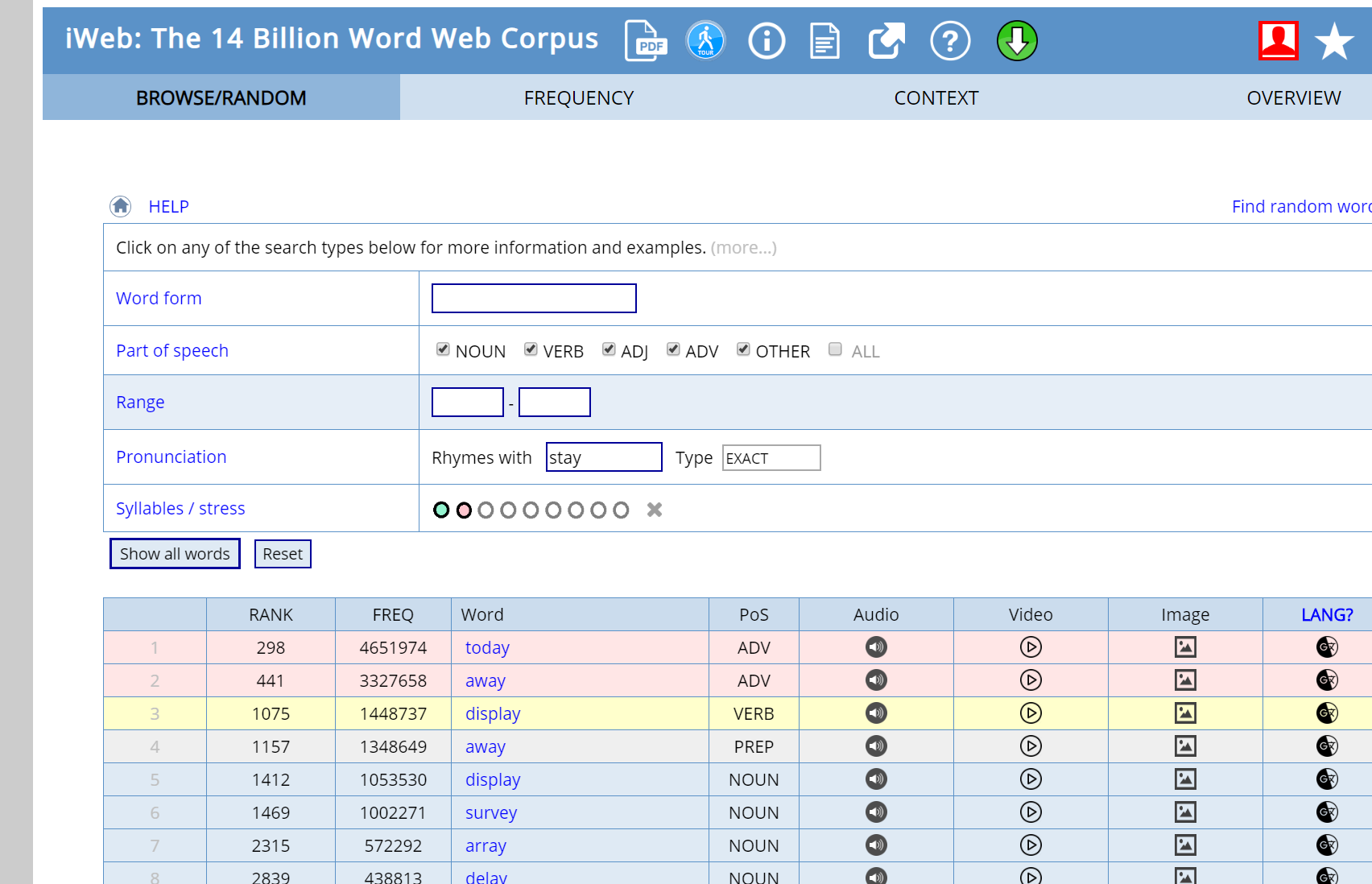




介绍iweb

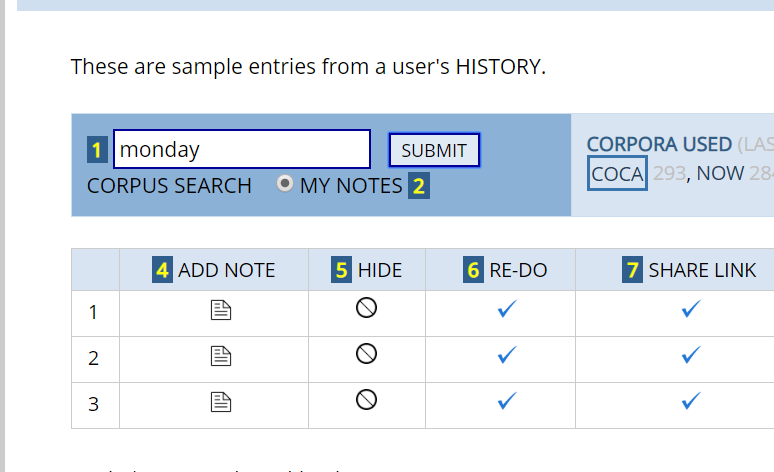
<https://www.english-corpora.org/iweb/>



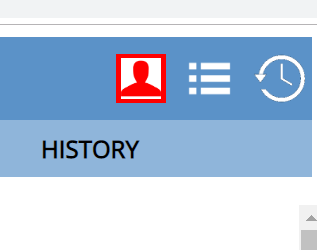


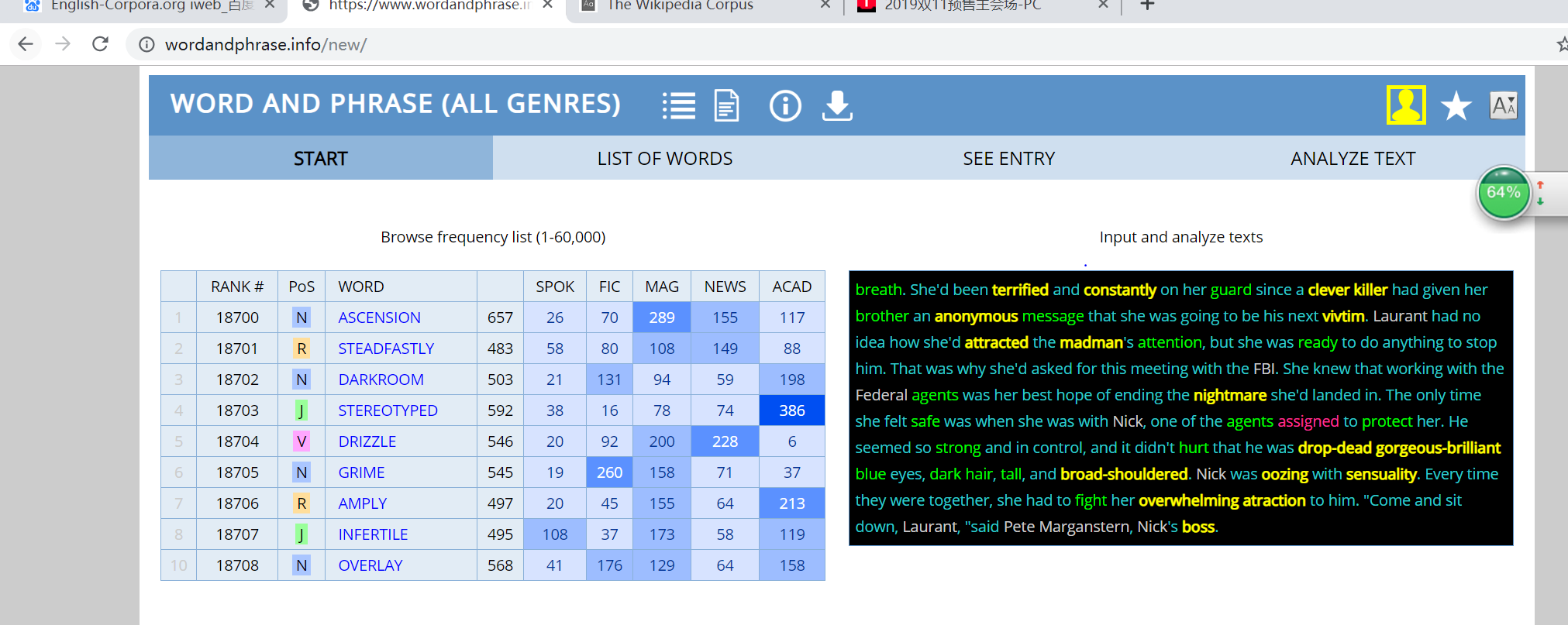
点击时钟，出现搜索的历史HISTORY功能

[www.wordandphrase.info/new](http://www.wordandphrase.info/new)



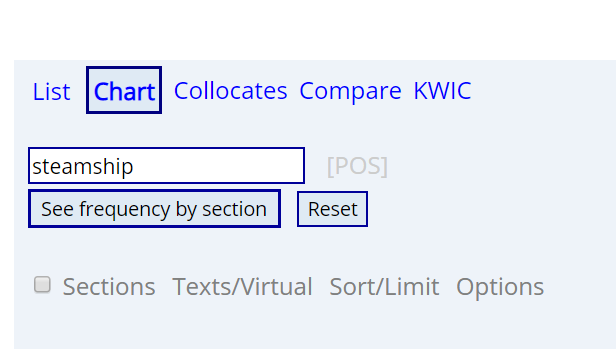
做笔记





COHA

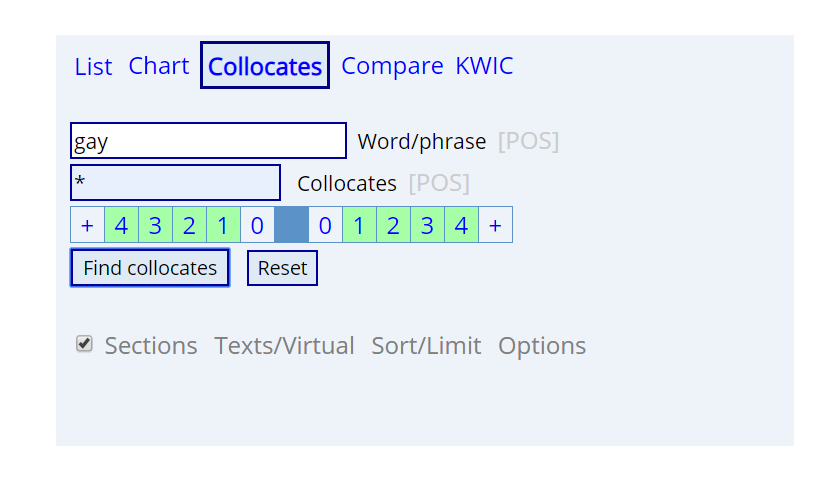
<https://www.english-corpora.org/coha/>

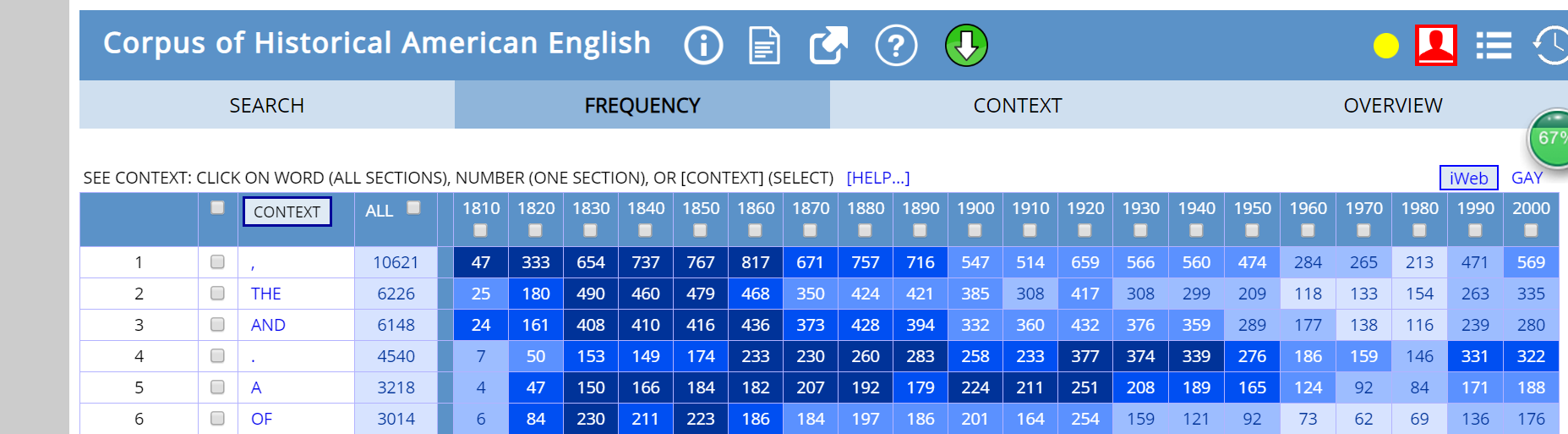


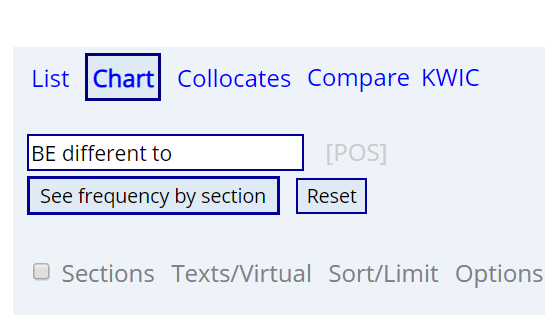


两个年代比









<https://www.english-corpora.org/tv/>

informative talk.

How to use coca. 30 minutes

Q: individual corpus

SPOKEN COCA 不如BNC

上外音响出版社：国外硕博士论文库

2019年10月26日星期六

地点：上海理工大学综合楼学术报告厅

**Keynote speech**

**Making very large corpora useful: creating virtual corpora in the BYU corpora**

**Mark Davies, BYU**

**Corpus Linguistics: an Interdisplinary perspective**

**Shanghai, China**

**October 2019**

[www.English-Corpora.org](http://www.English-Corpora.org)

Virtual Corpora, Compare Virtual Corpora

English for Specific Purposes(ESP)

[www.wordfrequency.info](http://www.wordfrequency.info)

wordandphrase.info

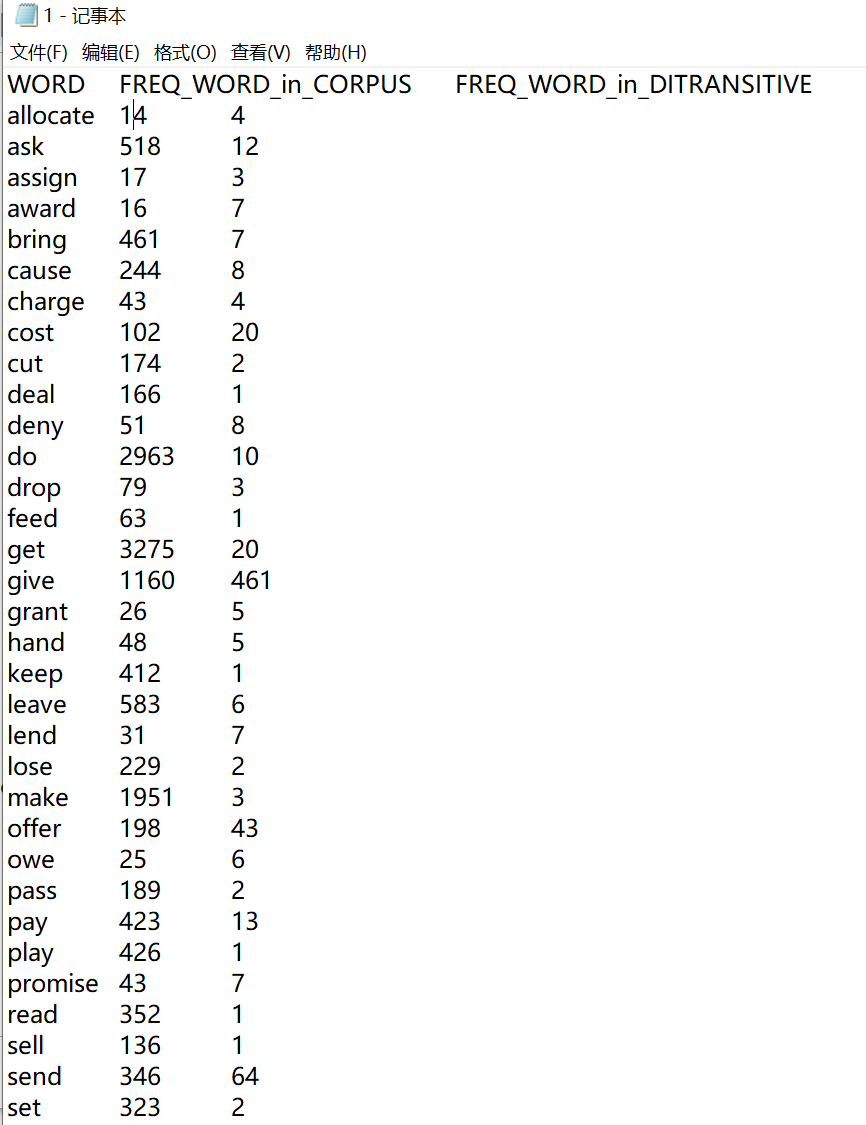
The BootCat Frontend ???? academic

演示：Creative Virtul Corpora

<http://www.stgries.info/teaching/groningen/index.html>

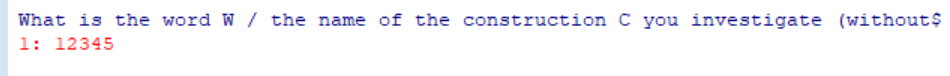
数据

<http://www.stgries.info/teaching/groningen/index.html>

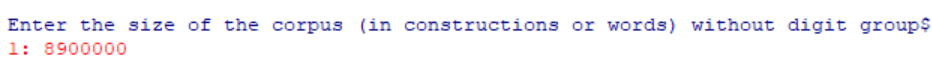


文件格式用txt，里面的格式不能改

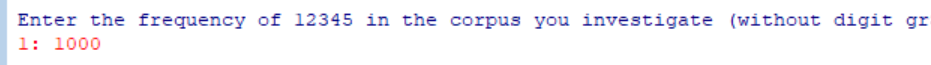
构式在整个的频率



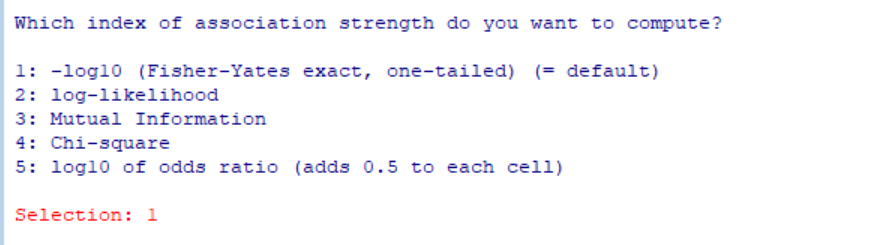
输入名称



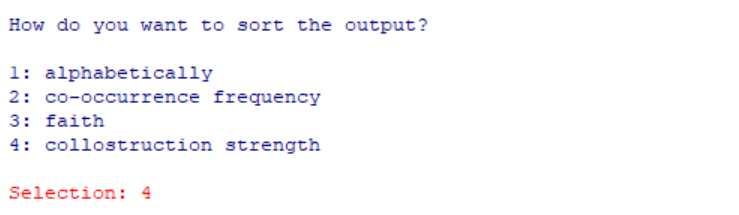
输入语料库的大小



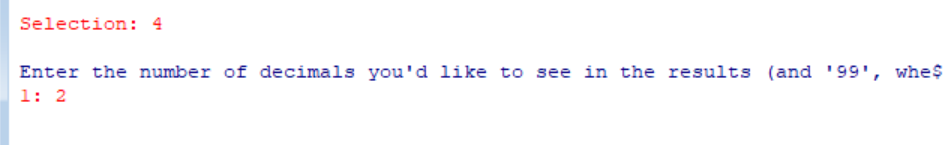
选择在构式中的频率



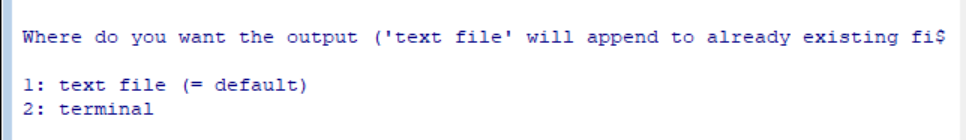
选择方法，一般选择1



选择4，表示搭配强度



保留小数



结果在终端还是在文档。

多因素分析：比较火

* 研究背景

学术读写素养范式下的学术英语读写教学研究

上海理工大学， 陈征

学术英语

学术英语教学

学术英语教学是为了提高学术英语交流采用的教学体系

理论：新读写理论

课程：学术英语读写课程，4课时/周，总64，通过四级，非英语专业

大一、大二，教材？

实践流程：学术阅读、学术写作、学术实践

* 研究设计：

学术写作论文？搜集？

* 研究工具
* 研究结果

学术英语语篇读写相关性分析。

软件词汇结构？

* 言据性分析学术学术论证模式的发展特点

访谈：

被试：大一下的学生

展望：用研一和大二的学生

怎么调动学生的积极性？

提交1500字的英文论文。

张金福：英语专业本科毕业论文结语词汇口语化特征分析

怎么界定词汇口语化特征？？

建立了一个库，60多所高校。收集了2500篇，标注了的？

主要看结论，得到桂诗春先生的观点。

Conclusition抄袭要少些？

Biber（1988）

上海交通大学学术语料库。

Qaquot(2010:150) 列出了口语特征

怎么判别口语特征？

24个口语特征 Paquot(2010)

口语化特征的选择标准参照了Paquot(2010)确认的24个口语化词汇项

《中国高水平英语学习者学术写作中的口语化特征》

收集学生的毕业论文的结语？

卡方检验： P值。

粱茂成、刘霞（2014）语篇内部的短语学特征分布模式探索——以学术论文为例

比较研究，建立了一个本科论文库。

结语总库70万。

国外本科论文，吉林大学，李亮博士

吕颖：温州医科大学，中外学者医学论文

立场标记:知识立场、态度性立场、言谈风格

立场标记理论发展脉络

摘要：说服性、广告性文本，吸引编辑和读者

Hyland学术立场标注

学术语篇作者立场标记框架

**We/us/our**

**没有找到I，可能医学论文是2-3个人。**

结论：

1中西传统文化影响。2国内外医学摘要的体裁差异。3两类作者语言能力差异。

作者立场框架：

自己的设想？

Q如果语篇库容不够大，词和频率不够高，文本标注的时候，这个词很可能落入另外一个marker，增强语/模糊语，怎么区分呢

语步标记？

Should这个词? 增强语， should模糊语，需要根据具体文本。

元话语需要标记。

功能越细

According to .

Q：分类基于自己的立场，怎么避免？

* **倪锦诚，理工类硕士学位论文摘要动词Reduce的语义韵研究**

摘要是精华和学术？

从词汇层面研究语义韵。

Provide, reduce, increase

摘要的立场

摘要的语义韵

We/我们（在汉语言中的使用）

留学生的结尾

结语的对比？

You shall know a word by the company it keeps: Firth(1937:) (J. R. Firth 1957: 11)

收集1600篇硕士论摘要

分析软件Antconc

为什么选择reduce，高频使用较多的。

卡方检验

李晶洁：建议：1倾向于自己建库，获取语料库的方法不是爱多，自己建库被质疑，库的代表性？库够不够大，是不是一个可比语料库？ Mark Davies

如果有现成的，可以解决工作量

语义韵的问题，时间很久，

语义韵研究库，细致

中国学者是overuse, prosody.

2可用出现reduce的某个用法泛化，可能代表某个，over use是否合适。

杨成：中外高影响因子期刊ESP论文摘要词块使用研究

1.语篇研究现状综述

微观：摘要语法意义、语用功能的研究

分类Hyland（2008a）

杂志的来源：语言学类的核心

工具：软件Antconc3.4

研究结果的参数：结构性特征，功能性特征，结构—功能关联性研究

卡方检验

频数分布于语义指向分布

论文写作中，可不可以用第一人称？

硬科学：不乏第一人称的使用

“we”使用最为频繁。（Kuo 1999:121）

CARE语料库，北航建立的，国内最大的，学术英语语料库。

抽取482篇文章，280万。

语料库的介绍。

语料：7000条，逐条每一行

数学和物理还倾向使用we.

遵循学科规范，无意识的行为，使用we。

研究的颗粒化的重要性。

We的分类。

**上海理工大学，陈钰：模糊限制语**

Hedges

Corpus Application in EFL Academic wit

什么是模糊限制语： Hyland&Milton,情态助动词，情态动词情态形容副词及名称

3语料库教学法

DDL, Tim Johns

你觉得有帮助吗？

⼀个爸爸带着他的⼉⼦冲进医院要求做⼿术，⼿术医⽣说：“我做不了这个⼿术，他是我⼉

⼦！

• 为什么？

答：因为医生是女的

偏见：机器是男的

机器学习的步骤：

Training data are collected and annotated → model is trained → Media are filtered, ranked, aggregated, or generated → People see output

土耳其的机械工人的分布图：Bias in Data Selection does not reflect a random sample

数据的分布没有和正在分布一样

大脑对信息的重视程度：

数据的偏移。

人对大脑的有意识的删除会造成比较多的问题

梁茂成：Easy Word2Vec

Linguistics naïve,

.66

Techqiue sophisticated

Word Similiarity Plot

* 韩素音翻译大赛

Large scale

基于Doc2vec

韩戈玲：Vice-Dean

新技术+

新技术不是为每个人准备的。

工具开发的应用者

真问题？

新文科：技术+传统文科的结合，是一种趋势。

John Sinclair是技术与文科结合的典范。

语料库语言学：任何人不能阻挡。

光做技术不行，光做语言不行？