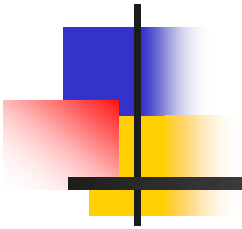


第六讲　语料库

詹卫东

http://ccl.pku.edu.cn/doubtfire/

提纲

- 1 什么是语料库
- 2 语料库的发展简史
- 3 语料库的设计
- 4 语料库的加工
- 5 语料库的应用

1 什么是语料库

在今天，仅仅将语料库视为存放语言材料的仓库，是令人无法忍受的观点。新一代的兆亿级的大规模语料库可以作为语言模型的训练和测试手段，来评价一个语言模型的质量；此外，诸如困惑度之类的统计方法也可利用语料库来评估一个语法模型对语料的解释能力。

—— Geoffrey Leech, *The State of The Art in Corpus Linguistics*, 1991, In Aijmar, K. and Altenberg, B. , eds. , *English Corpus Linguistics: Studies in Honor of Jan Svartvik*, London: Longman, 1991.

关于语料库的三点基本认识

- 语料库中存放的是在语言的实际使用中真实出现过的语言材料；
- 语料库是以电子计算机为载体承载语言知识的基础资源；
- 真实语料需要经过加工（分析和处理），才能成为有用的资源；

语料库的分类

- 口语语料
- 书面语料

- 共时语料
- 历时语料

- 平衡语料
- 专门语料

- 监控语料
- 样本语料

- 单语
  - 词性标注语料
  - 树库语料
  - …
- 双语
- 多语
  - 平行语料库 parallel
  - 比较语料库 comparable

语料库示例（一）

北京大学计算语言所富士通人民日报标注语料库样例：

历史/n 将/d 铭记/v 这个/r 坐标/n ：/w 北纬/b ４１．１/m 度/q 、/w 东经/b １１４．３/m 度/q ；/w 人们/n 将/d 铭记/v 这/r 一/m 时刻/n ：/w １９９８年/t １月/t １０日/t １１时/t ５０分/t 。/w

……

[中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n，/w 并/c 按照/p “/w 一国两制/j ”/w 、/w “/w 港人治港/l ”/w 、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an 。/w

语料库示例（二）

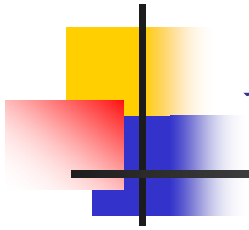
London-Lund英语口语语料库样例

```
^what a_bout a cigar\ette# .                    /
*((4 sylls))*                                   /
*I ^w\on't have one th/anks#* - - -             /
^aren't you .going to sit d/own# -              /
^[/\m]# -                                       /
^have my _coffee in p=eace# - - -               /
^quite a nice .room to !s\it in ((actually))#   /
*^\isn't* it#                                   /
*^y/\es#* - - -                                 /
```

转引自Tony McEnery & Andrew Wilson, 1996, Corpus Linguistics, p55,

London-Lund英语口语语料库部分标记

标记	含义
#	语调群的结束 (end of tone group)
^	语音开始 (onset)
/	上升型核心语调 (rising nuclear tone)
\	下降型核心语调 (falling nuclear tone)
^	先升后降型核心语调 (rise-fall nuclear tone)
_	平型核心语调 (level nuclear tone)
[ ]	不完整的词语和音节符号 (enclose partial words and phonetic symbols)
.	标准重音 (normal stress)
!	高音高于前一个音节的重音 (booster: higher pitch than preceding prominent syllable)
=	高音跟前一个音节相当的重音 (booster: continuance)
(( ))	不清晰的音节 (unclear)
* *	同步发音 (simultaneous speech)
-	一个重音单位的停顿 (pause of one stress unit)



语料库与语言知识库

语言知识库 （Linguistic Database）

语料库（corpora/corpus）

2 语料库发展简史

- 20世纪50年代 Chomsky的影响
- 第一代（1970－80年代）
- 第二代（1980－90年代）
- 第三代（1990年代）
- ？ 第四代 （21世纪）

第一代语料库

- Brown语料库
- LOB语料库
- LLC语料库

1960年代初，美国Brown大学，100万词次，当代美国英语，根据系统性原则采样，……

1970年代初，英国Lancaster大学，挪威Oslo大学，挪威Bergen大学，当代英国英语，……

百万词级

以语言研究为导向

1960年代初，由London大学Randolph Quirk主持，收集2000小时的谈话和广播等口语素材并整理成书面材料，由瑞典Lund大学J. Svartvik主持全部录入计算机，1975年建成……

第二代语料库

建于1980年代，由英国Birmingham大学与Collins出版社合作完成，规模达2000万词次，基于该语料库出版的Collins Cobuild词典（1987）受到了广泛的好评

- COBUILD语料库
- Longman语料库

千万词级

词典编纂 — 应用导向

建于1980年代，包括三个语料库：
LLELC语料库（Longman/Lancaster英语语料库）
LSC语料库（Longman口语语料库）
LCLE（Longman英语学习语料库）
目标是编撰英语学习词典，为外国人学习英语服务，词典规模达5000万词次

第三代语料库

美国计算语言学会倡议发起“数据采集计划”（Data Collection Initiative），由宾州大学M.Liberman主持，保存语料原始文本形式以及SGML标注信息

- ACL/DCI语料库
- UPenn树库
- LDC

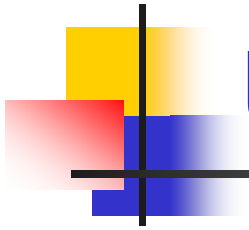
超大规模（上亿词级）

标准编码体系

深度标注/多语种

NLP应用

- BNC （英国国家语料库）
- ……



UPenn Treebank

- 美国Pennsylvania大学1980年代末开始发起
- 由该校计算机系M.Marcus主持
- 1993年，完成了对近300万英语词的句子语法结构标注
- 2000年发布中文树库（第一版）
  10万词，4185个句子，325 data files（新华社语料）
- 2004年发布中文树库 4.0版
  404,156 words, 664,633 Hanzi, 15,162 sentences, and 838 data files（大陆、香港、台湾语料）

宾州大学中文树库示例

他还提出一系列具体措施和政策要点。

他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ 措施/NN 和/CC 政策/NN 要点/NN 。/PU

```
(IP (NP-SBJ (PN 他))
    (VP (ADVP (AD 还))
        (VP (VV 提出)
            (NP-OBJ (QP (CD 一)
                        (CLP (M 系列)))
                    (NP (NP (ADJP (JJ 具体))
                            (NP (NN 措施)))
                        (CC 和)
                        (NP (NN 政策)
                            (NN 要点))))))
    (PU 。))
```

3 语料库的设计

语料库三方面	属性	值
A. 语料本身	规模	百万词级 | 千万词级 | 亿万词级 | …
	领域	政治 | 经济 | 体育 | 心理学 | …
	体裁	文学 | 应用文 | 新闻 | …
	时代	共时 | 历时
	语体	书面语 | 口语
	语种	单语 | 双语 | 多语 双语平行语料库 | 双语比较语料库
	语言层次	语音（音节，韵律） | 语法（词，句，…）
B. 语料加工	数据形式	Text文本 | HTML文本 | 数据库 | …
	编码体系	TEI标准 | 自定义编码体系 | …
	加工层次	词 性 | 句 法 | 语 义 | 语 篇 | … 双语句子对齐 | 词对齐 | …
	加工方式	自动 | 人机互助 | 人工
C. 语料应用	应用领域	通用 | 词典编纂 | 机器翻译 | …
	辅助软件	检索工具 | 人机界面 | 数据接口 | …

语料的选取

- 精品原则
- 有影响力原则
- 随机挑选原则
- 高流通度原则
- 典型性原则
- 易于获得原则
- 具有统计样本意义原则
- 符合语言规范原则

语料库的编码体系

- SGML（标准置标语言）

  http://www.w3.org/MarkUp/SGML/
- XML（可扩展的置标语言）

  http://www.w3.org/TR/REC-xml
- TEI（文档编码计划）

  http://www.tei-c.org/
- CES（语料库编码标准）

  http://www.tei-c.org/Applications/index-co02.html

范围缩小，针对性加强

冯志伟，1998，《标准通用置标语言SGML及其在自然语言处理中的应用》，载《当代语言学》1998年第4期。

CES标准（Corpus Encoding Standard）

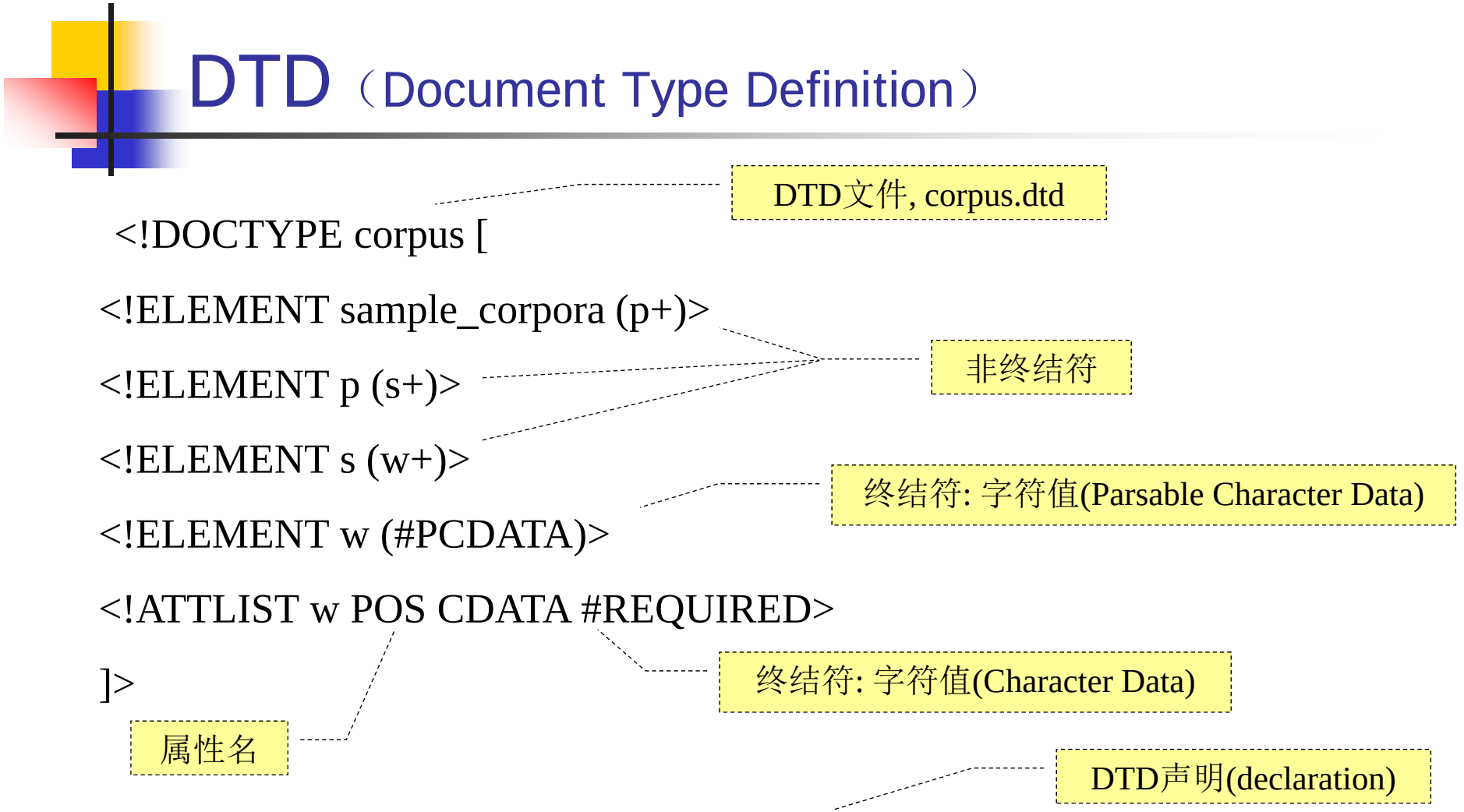
语料库/n 标记/n 应该/v 有/v 规范/n

语料库A: 不符合CES

```
<sample_corpora>
…
<p>
  <s>
    <w POS="n">语料库</w>
    <w POS="n">标记</w>
    <w POS="v">应该</w>
    <w POS="v">有</w>
    <w POS="n">规范</w>
  </s>
</p>
…
</sample_corpora>
```

语料库B: 符合CES

DTD（Document Type Definition）



4 语料库的加工

语料库标注（Annotation）

1）词性标记（Part-of-speech tagging）

2）句法层次和范畴标记（Grammatical parsing）

3）词义标记（Word sense tagging）

4）篇章指代标记（Anaphoric annotation）

5）韵律标记（Prosodic annotation）

……

http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html

语料库加工工具

分类	工具名称	功能描述
A.文件处理工具	文本过滤器	将不同的文件格式转成为纯文本文件格式
	文本分类器	自动判别文本领域
	语料库辅助校对工具及一致性检查工具	按照语料库加工规范，对语料质量进行管理
B. 语言处理工具	分词与词性标注工具	对语料进行词语识别，词性标记处理
	词义标注工具	对词义进行标注
	浅层分析工具	对语块（chunk）进行标注
	句法分析工具	对句子进行完全句法分析
	双语语料对齐工具	对双语语料进行各个层级（段落、句子、小句、词）的对齐加工

双语语料库（Bilingual Corpora）加工

- 段落对齐
- 句子对齐
- 词对齐
- 短语对齐

双语句子对齐

- 基于长度（length-based）的对齐方法 Gale & Church (1993)

  纯粹基于句子的长度来估计对齐可能性
  资源要求少，算法效率相对较高

- 基于词（word-based）的对齐方法

  一般要依赖词典资源，算法效率相对较低

双语句子对齐示例

中国支持在平等参与、协商一致、求同存异、循序渐进的基础上，开展多层次、多渠道、多形式的地区安全对话与合作。

中国参加了东盟地区论坛、亚洲建立协作与建立信任措施会议、亚太安全合作理事会和东北亚合作对话会等活动，主张通过这些政府和民间讨论安全问题的重要渠道，增进各国的相互了解与信任，促进地区和平与稳定。

……

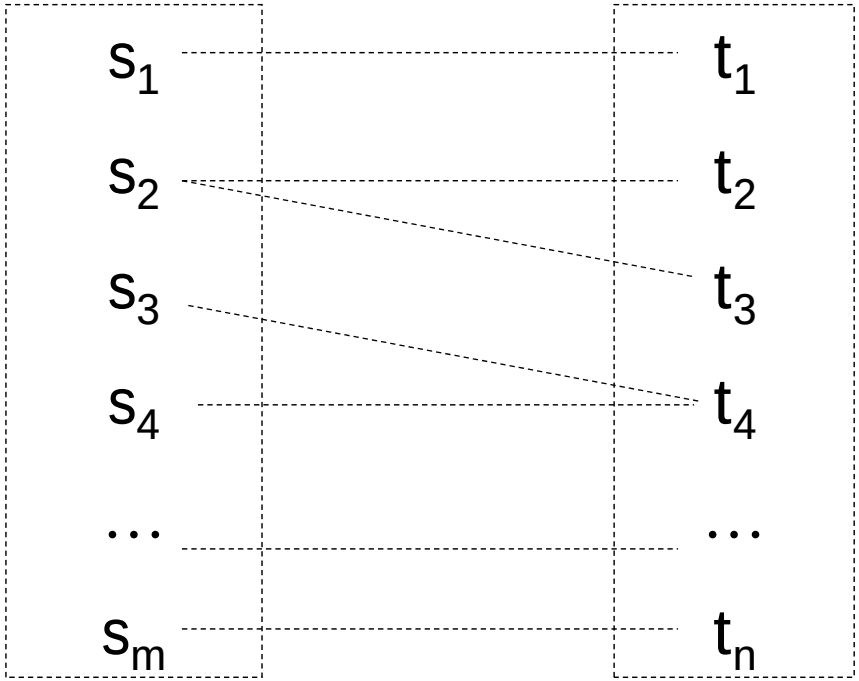
China advocates regional-security dialogue and cooperation at different levels, through various channels and in different forms.

Such dialogue and cooperation should follow these principles: participation on an equal footing, reaching unanmimity through consultation, seeking common ground while reserving differences, and proceeding in an orderly way and step by step.

China has participated in the ASEAN Regional Forum (ARF), Conference on Interaction and Confidence-Building Measures in Asia (CICA), Council on Security Cooperation in Asia and Pacific Regional (CSCAP), Northeast Asia Cooperation Dialogue (NEACD) and other activities, holding that all countries should further mutual understanding and trust by discussions on security issues through these important governmental and non-governmental channels, so as to promote regional peace and stability.

……

双语句子对齐问题



影响对齐猜测的两个因素：

- 配对模式
- 句长差距

句子配对模式(记做Match)

- Gale & Church(1993) 定义了六种配对模式，在实际语料[1]中的分布频度为：

句子配对模式 (Match)	出现次数	概率 P(Match)
1-0 或 0-1	13	0.0099
1-1	1167	0.89
1-2 或 2-1	117	0.089
2-2	15	0.011
	1312	1.00

Note1: UBS/Union Bank of Switzerland出版的经济报告，同时使用英、法、德三种语言

句长相关性 Gale & Church(1993)

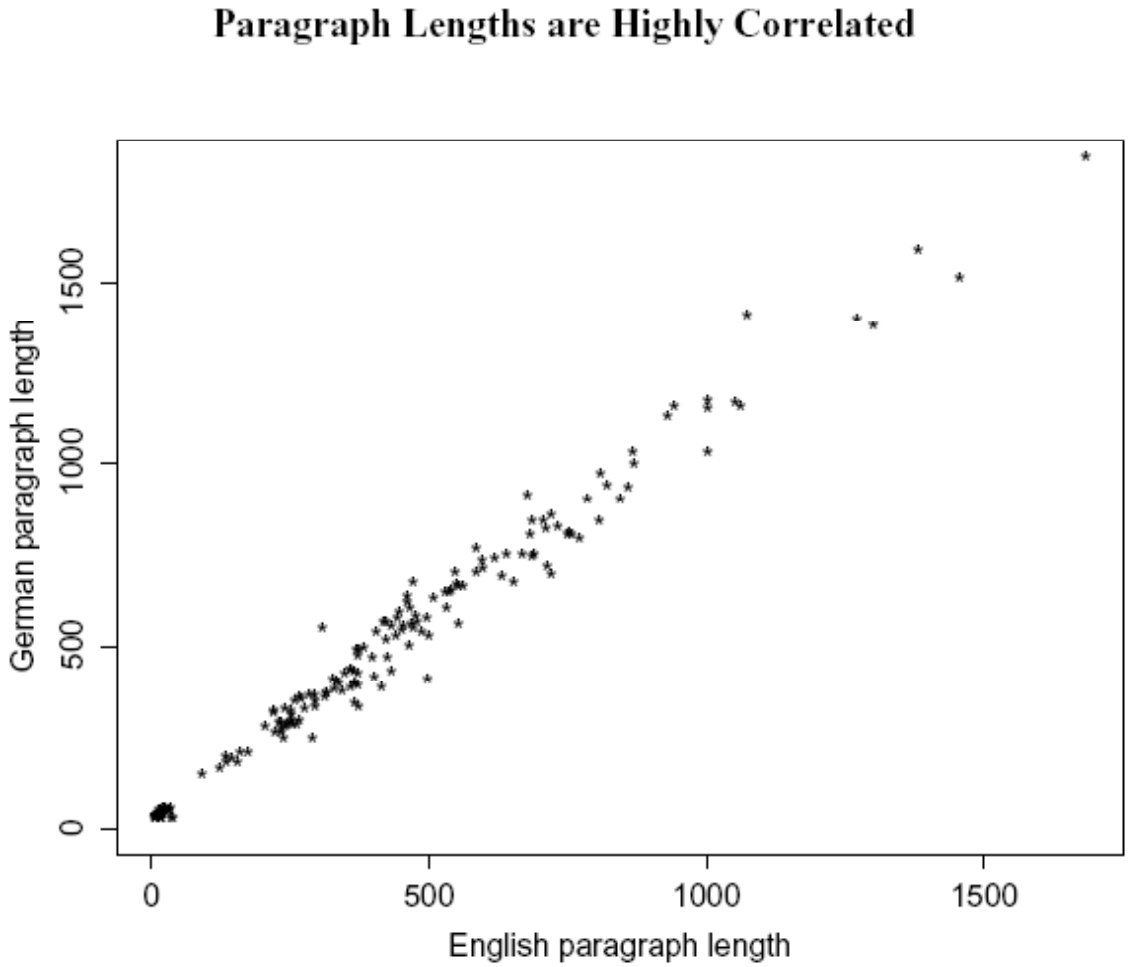


Figure 1. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

句子长度差距关系(记做δ)

- 任一个双语句对 (S_i, T_j) S表示原文，T表示译文

$$S_i = s_1 s_2 \ldots s_m \qquad l_i = L(S_i)$$

$$T_j = t_1 t_2 \ldots t_n \qquad l_j = L(T_j)$$

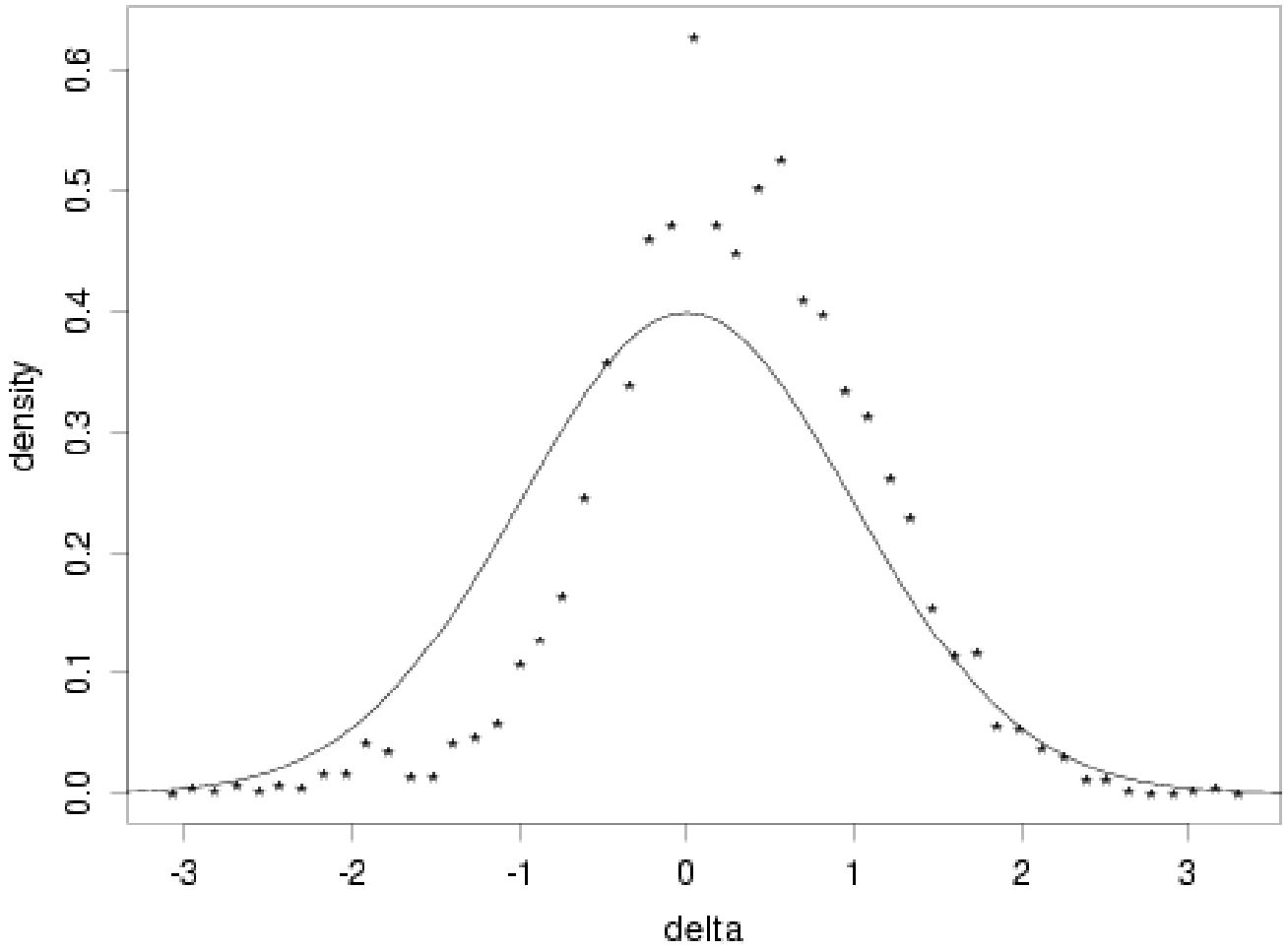
S中任意一个字符在T中所对应的字符数是个随机变量，记做X

X呈正态分布，X的期望记做c，X的方差记做V^2

由此则可定义随机变量 δ 来度量两个句子之间的长度差距关系

$$\delta(l_i, l_j) = \frac{l_j - c \times l_i}{\sqrt{l_i \times V^2}}$$

δ 呈正态分布 Gale & Church(1993)



句子长度差距关系（续）

- 随机变量X的期望c和方差V^2可以从已经对齐好的双语平行语料库中估算得到

  比如：英语→法语　$c \approx 72302/68450 \approx 1.06$

  $V^2 \approx 5.6$

  Gale & Church (1993)

  汉语→英语　$c \approx 1.46$　　$V^2 \approx 2.9$

  刘昕　等(1995)

基于长度的双语句子对齐方法

- 任意句子S_i与T_j对齐的可能性就可以表示为一个条件概率：

$$P(Match(s_i,t_j)\mid\delta(l_i,l_j))=\frac{P(\delta\mid Match)\times P(Match)}{P(\delta)}$$

$$\approx P(\delta\mid Match)\times P(Match)\quad \text{公式1}$$

$$P(\delta\mid Match)\approx 2(1-P(|\delta|))\quad \text{公式2}$$

δ 服从标准正态分布，P(| δ |) 可通过查标准正态函数分布表得到

$$P(\delta)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\delta}e^{\frac{-z^2}{2}}dz\quad \text{公式3}$$

基于长度的双语句子对齐方法（续）

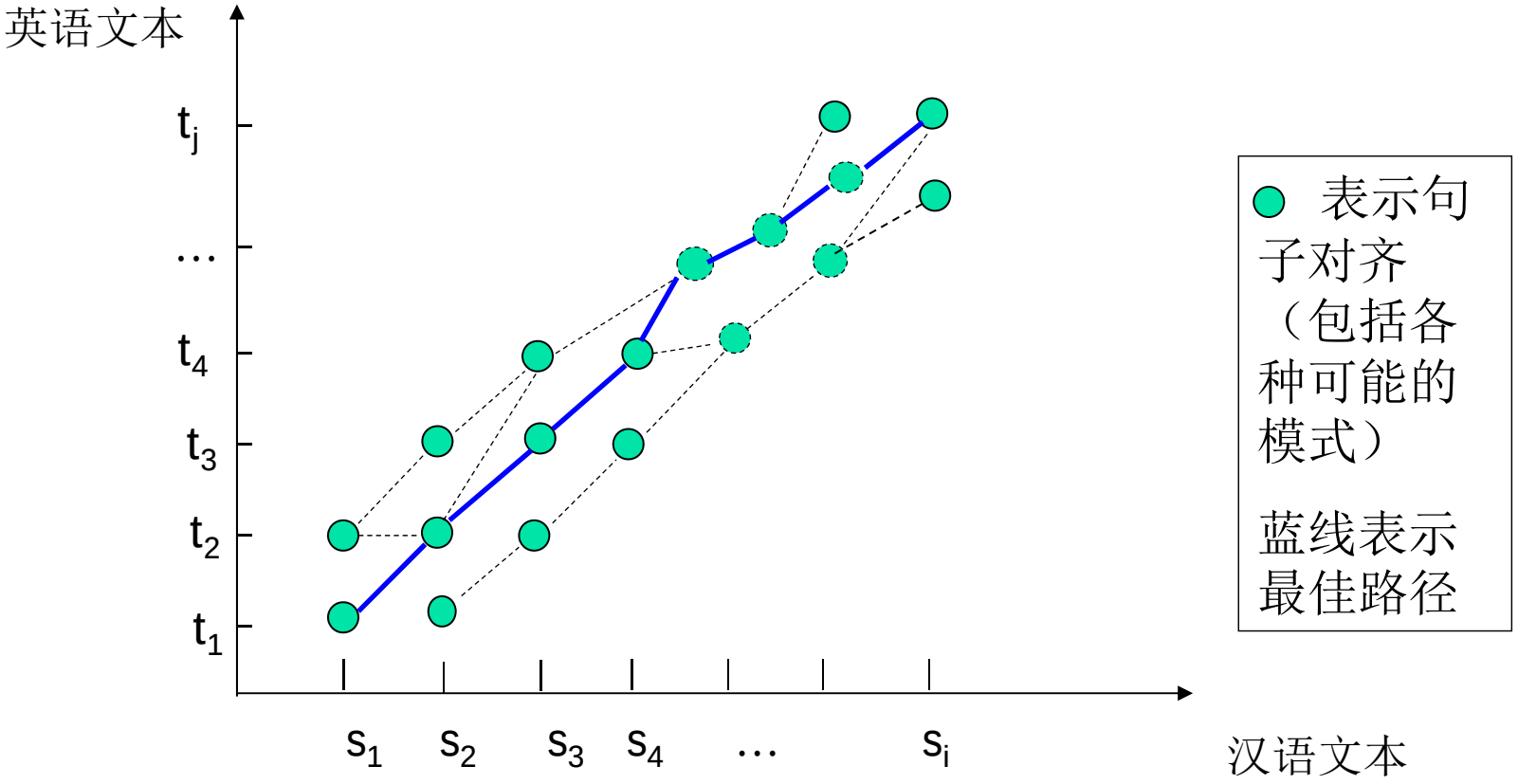
- 对公式1取对数，将乘法运算变为加法运算

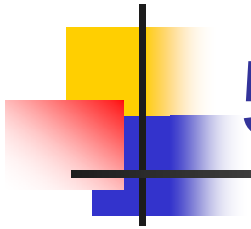
$$Score(s_i, t_j) = -(\log_2(P(\delta \mid Match)) + \log_2(P(Match)))$$

公式4

Score是对两个句子配对可能性的一个评估，可以形象地理解为两个句子之间的距离。得分越低，表示两个句子之间距离越近，因而配对的可能性越高

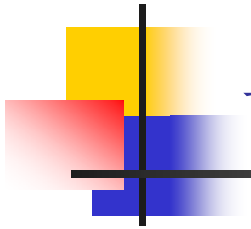
求解双语句子对齐示意图





5 语料库应用

- 支持自然语言处理应用系统开发
- 支持语言学研究和语言教学研究



语料库对NLP的支持

- 基于大规模语料库的语音识别；
- 基于大规模语料库的音字转换技术（中文输入）；
- 基于大规模语料库的自动文本校对技术；
- 利用语料库训练HMM模型进行分词，词性标注，词义标注，等等；
- 基于语料库的句法分析；
- 基于语料库的机器翻译；
- 基于机器学习技术，通过语料库获取语言知识，包括搭配特征，句法规则，等等；
- 基于语料库的语言模型训练和语法模型评价；
  支持NLP自动评测；

中文音字转换

拼音串（无声调）	xue xi dian nao ji shu

中文音字转换（续）

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	

中文音字转换（续）

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	
候选词串	学习 电脑 级数	共有2×1×7＝14 种可能性
	血洗 电脑 奇数	
	血洗 电脑 基数	
	……	

中文音字转换（续）

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树 学 洗 电 闸 给 述 学 西 颠 挠 记 书 ……	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
候选词串	学习 电脑 级数 血洗 电脑 奇数 血洗 电脑 基数 ……	共有2×1×7＝14 种可能性
正确文字串	学习 电脑 技术	

基于语料库的语言研究

- Concordance （索引 —— 相关集列）
- Collocation （搭配的定量研究）

- Cobuild Concordance and Collocations Sampler
  http://titania.cobuild.collins.co.uk/form.html

- 台湾“中研院”现代汉语平衡语料库
  http://www.sinica.edu.tw/ftms-bin/kiwi.sh

- 孙茂松 等，1997，《汉语搭配定量分析初探》，载《中国语文》1997年第1期。pp29-38。

关于搭配的定义

- 搭配是重复出现的；
  - “大房子” —— “大手笔” —— “大文科” / “大历史”
- 搭配是不可类推的；（自由组合 —— 受限组合）
  - “ “吃白菜” —— 吃豆腐” —— “喝西北风”
- 搭配一般具有正常的句法结构；
  - “戴高帽” —— ？“戴高” —— “风马牛不相及”
- 搭配通常与领域相关；
  - “语言习得” —— “学说话” “风险投资”

搭配的量化分析（一）

- 语料库：90-91年新华社新闻语料库，1000万字，710万词
- 搭配强度： 重复出现越多，搭配强度越大

$$MI(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

$$S(w_i, w_j) = \log_2 \frac{N \sum_{k=-5}^{+5} Count_k(w_i, w_j)}{Count(w_i)Count(w_j)}$$

k表示w_j相对于w_i的位置
－ 表示在左，+表示在右
（+号一般略去）

K= -5, -4, -3, -2, -1, 1, 2, 3, 4, 5

N 表示语料库规模 $N = 7.1 \times 10^6$

搭配强度的量化分析示例

- 候选搭配：(能力，弱)　　(能力，大)
- 通过语料库统计得到：

$Count_{-3}(能力,弱)=1$　$Count_{1}(能力,弱)=3$　$Count_{2}(能力,弱)=5$

$Count_{-5}(能力,大)=6$　$Count_{-4}(能力,大)=4$　$Count_{-3}(能力,大)=8$

……

$Count_{1}(能力,大)=9$　　……　　$Count_{5}(能力,大)=5$

$Count(能力)=2241$　　$Count(弱)=177$　　$Count(大)=19913$

搭配强度的量化分析示例（续）

$$S(\text{能力},\text{弱}) = \log_2 \frac{7.1\times10^6(1+3+5)}{2241\times177} = 7.33$$

$$S(\text{能力},\text{大}) = \log_2 \frac{7.1\times10^6(6+4+8+4+2+9+6+4+6+5)}{2241\times19913} = 3.10$$

同理可得：

$S(\text{能力},\text{强}) = 7.45$　$S(\text{能力},\text{差}) = 6.63$　$S(\text{能力},\text{小}) = 0.74$

与“能力”的搭配能力：　强 ≻ 弱 ≻ 差 ≻ 大 ≻ 小

搭配的量化分析（二）

- 搭配的离散度

方差公式

$$u(w_i, w_j) = \frac{\sum_{k=-n}^{n}((Count_k(w_i, w_j) - \overline{Count}(w_i, w_j))^2}{2n}$$

$$\overline{Count}(w_i, w_j) = \frac{\sum_{k=-n}^{n} Count_k(w_i, w_j)}{2n}$$

均值公式

$$n = 5$$

离散度反映了两个成分共现的分布情况

离散度越高，越可能是搭配

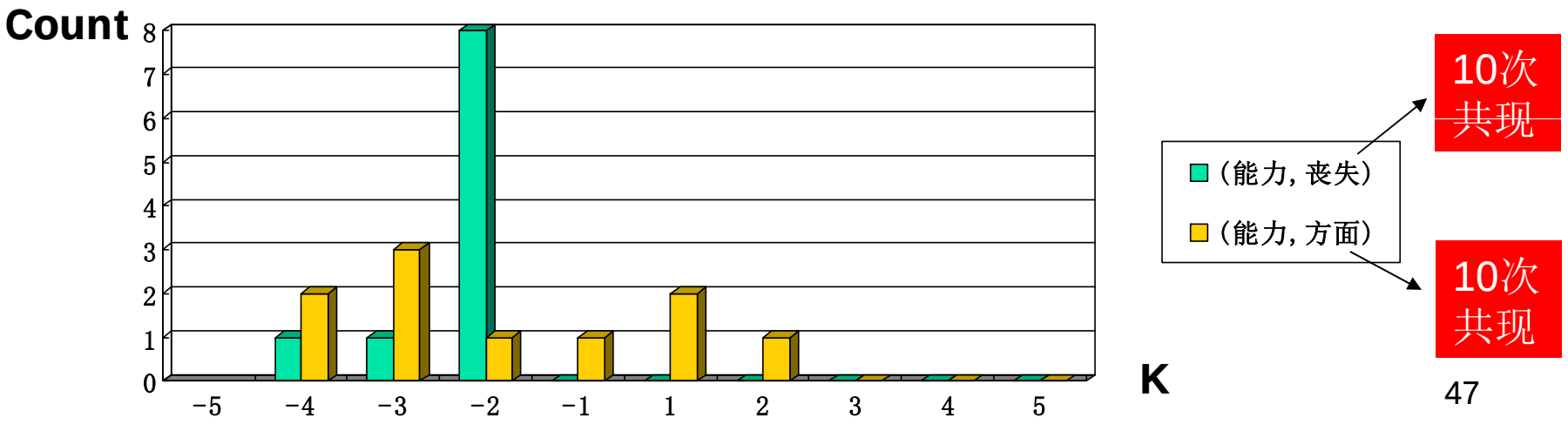
搭配离散度的量化分析示例

- 候选搭配： (能力,丧失)　　(能力,方面)
- 通过语料库统计得到：

$$Count_{-4}(\text{能力,丧失}) = Count_{-3}(\text{能力,丧失}) = 1 \quad Count_{-2}(\text{能力,丧失}) = 8$$

$$Count_{-4}(\text{能力,方面}) = Count_{1}(\text{能力,方面}) = 2 \quad Count_{-3}(\text{能力,方面}) = 3$$

$$Count_{-2}(\text{能力,方面}) = Count_{-1}(\text{能力,方面}) = Count_{2}(\text{能力,方面}) = 1$$



搭配离散度的量化分析示例（续）

$$\overline{Count}(\text{能力},\text{丧失}) = \frac{1+1+8}{2\times 5} = 1 \qquad \overline{Count}(\text{能力},\text{方面}) = \frac{2+3+1+1+2+1}{2\times 5} = 1$$

$$u(\text{能力},\text{丧失}) = \frac{(1-1)^2+(1-1)^2+(8-1)^2}{10} = 5.60$$

$$u(\text{能力},\text{方面}) = \frac{2\times(2-1)^2+(3-1)^2+3\times(1-1)^2}{10} = 1.00$$

△ “丧失”与“能力”构成搭配关系，而“方面”跟“能力”不构成搭配关系

搭配的量化分析（三）

- 搭配的尖峰位置度量

$$Z_k(w_i, w_j) = \frac{(Count_k(w_i, w_j) - \overline{Count}(w_i, w_j)}{\sqrt{u(w_i, w_j)}}$$

$$Z_{-2}(\text{能力}, \text{丧失}) = \frac{Count_{-2}(\text{能力}, \text{丧失}) - \overline{Count}(\text{能力}, \text{丧失})}{\sqrt{u(\text{能力}, \text{丧失})}}$$

$$= \frac{8-1}{\sqrt{5.6}}$$

$$= 2.96$$

“丧失”在 -2 位置形成尖锋

“尖锋位置”的语言学含义

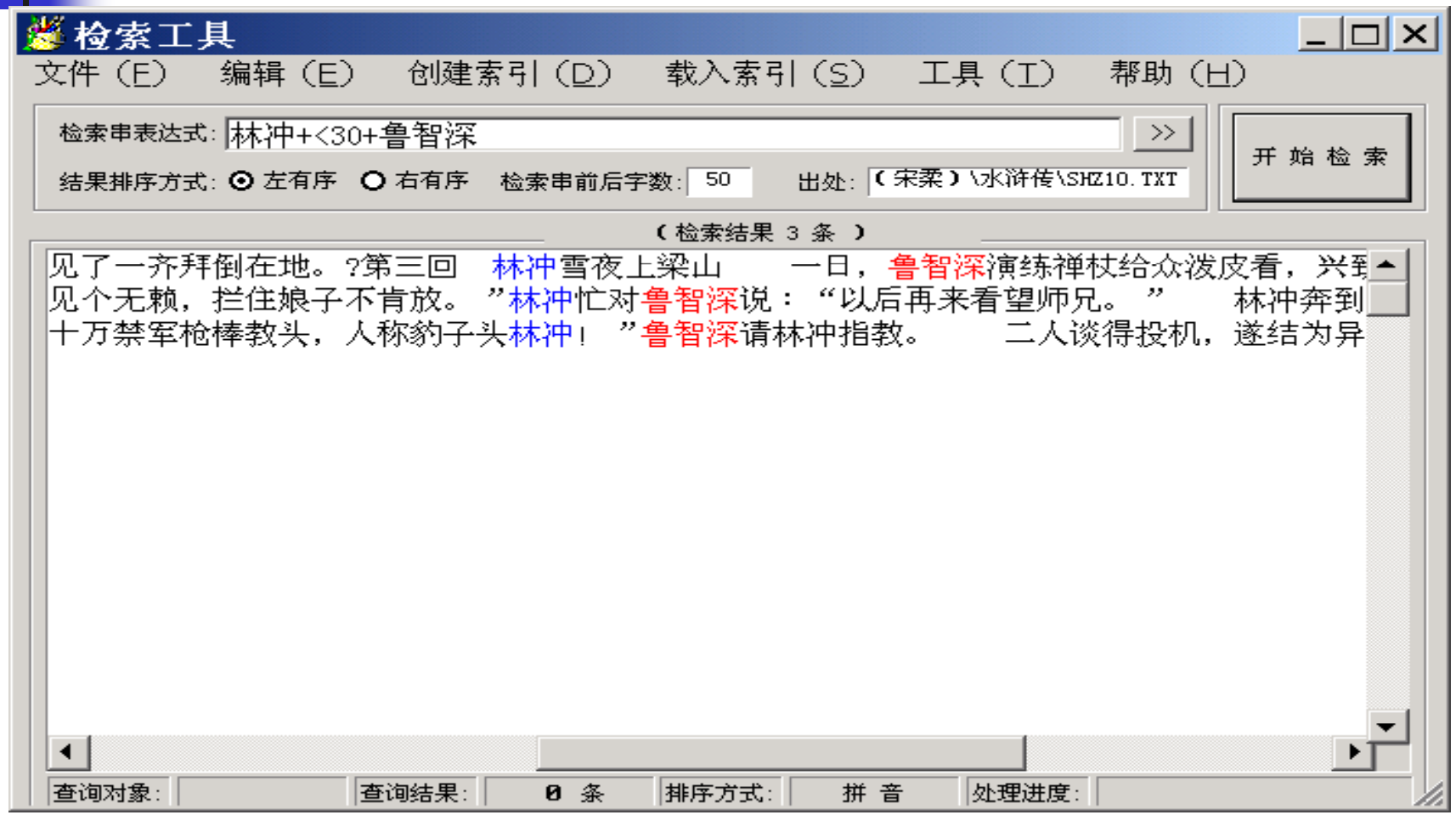
- 反映W_j与W_i可能形成的句法结构

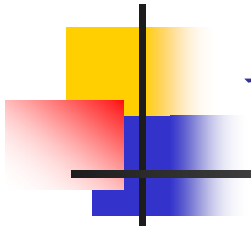
  - 能力 — 具有 ： 尖峰位置 -3, -2 述宾结构
  - 能力 — 差 ： 尖锋位置 1 主谓结构
  - 能力 — 提高 ： 尖峰位置 -4,-3; 1,3 述宾/主谓

$$Z_{-3}(\text{能力},\text{提高}) > Z_{1}(\text{能力},\text{提高})$$

  - 能力 — 吞吐 ： 尖峰位置 -1 定中结构

语料库检索


检索工具
文件（F）
编辑（E）
创建索引（D）
载入索引（S）
工具（T）
帮助（H）
检索串表达式：林冲+<30+鲁智深
>>
开 始 检 索
结果排序方式：左有序
右有序
检索串前后字数：50
出处：（宋柔）\水浒传\SHZ10.TXT
（检索结果 3 条）
见了一齐拜倒在地。?第三回 林冲雪夜上梁山 一日，鲁智深演练禅杖给众泼皮看，兴到
见个无赖，拦住娘子不肯放。”林冲忙对鲁智深说：“以后再来看望师兄。” 林冲奔到
十万禁军枪棒教头，人称豹子头林冲！”鲁智深请林冲指教。 二人谈得投机，遂结为异
查询对象：
查询结果：
0 条
排序方式：
拼 音
处理进度：



语料库检索（续）

- 双语Concordance
- 树库短语结构检索

其他研究

- 风格学研究
- 抽取词表（单语，双语）
- 统计字频、词频，编写语言教材
- 词典编纂
- 句法结构研究
- 句型研究
- ……

进一步阅读文献

- Gale, W. & Church,K., *A program for aligning sentence in bilingual corpora*, In *Computational linguistics*, Vol.19, No.1, 1993.
- 刘昕，周明，黄昌宁，1995，《基于长度算法的中英双语文本对齐的试验》，载 陈力为等主编《计算语言学进展与应用》，清华大学出版社1995年版。
- 孙宏林，1997，《从标注语料库中归纳语法规则：“V+N”序列试验分析》，载陈力为、袁琦主编《语言工程》，清华大学出版社1997年版，pp157-163。
- 《当代语言学》1998年第1期，语料库语言学专刊。
- Graeme Kennedy, 1998, *An Introduction to Corpus Linguistics*, Addison Wesley Longman Limited. （外语教育与研究出版社2000年原版引进）
- Anthony Woods, Paul Fletcher, Arthur Hughes,1986, *Statistics in Language Studies*, Cambridge University Press. （外语教育与研究出版社2000年原版引进）
- 黄昌宁 李涓子，2002，《语料库语言学》，商务印书馆

复习思考题

1. 访问台湾中研院“现代汉语平衡语料库”网站，查询“能力”这个词在语料中的使用情况，撰写分析报告；

2. 访问[语言学光标]网站，阅读有关语料库语言学的文献，撰写一篇有关“语料库语言学”的评述报告。

   语言学光标 之 语料库语言学板块

3. 访问网上语料库资源

   LDC（Linguistic Data Consortium）

   http://www.ldc.upenn.edu/Catalog/index.html

   Upenn 中文树库

   http://www.ldc.upenn.edu/Catalog/LDC2000T48.html

   Livac共时语料库

   http://www.rcl.cityu.edu.hk/livac/search.php?lang=sc

   BNC语料库

   http://corpus.byu.edu/bnc/