

# 第七章 篇章分析

王峰

华东师大计算机系

# 主要内容

- 简介
- 衔接与连贯
- 指代消解
- 指代消解应用

# 语篇(Discourse)

- 前后意义关联的句子序列
- 几种说法：
  - 话语、语篇(篇章)、文本。(英文: **discourse, text**)
- 例子：
  - Ex1: 比尔来自于美国。今天交通非常拥挤。长江贯穿中国的多个省市。因此，计算语言学是计算机科学与语言学的交叉。
    - *4 correct sentences but collectively do not make meaning*
  - Ex2: 这里的交通非常拥挤。张先生早上**6: 40**之前就得出发去上班，常常会提前半小时到单位；如果稍晚一点，他就很可能迟到。
    - *It makes meaning*

# 意义相关性的体现 - 衔接

- 例子：
  - 张三擅长素描。他给家里的每个人都画了一幅[], 挂在房间的[]是自画像。
- 意义上是如何关联的？
  - 通过词汇语义表达关联
  - 围绕着“画”而展开：素描、画像、一幅[ ]
  - 通过“指代”形成关联
    - 人称代词“他”
    - 零型代词[]所表示的对象
  - 以词汇表示的关联，通常称为“衔接(cohesion)”

# 意义相关性的体现 - 连贯

- 例子：
  - [s1]张三把李四的车钥匙藏起来了。[s2]他喝醉了。
  - [s3]张三把李四的车钥匙藏起来了。[s4]他喜欢逗着乐。
  - [s5]张三把李四的车钥匙藏起来了。[s6]他爱看电影。
- 意义上是如何关联的？
  - 通过句子的意义表示关联
    - [s1]和[s2]构成合理的篇章：两个句子表示“因果关系”
    - [s3]和[s4]也构成合理篇章：同样表示“因果关系”
    - [s5]和[s6]构成合理篇章吗？
  - 通过句子意义表示的关联称为连贯(coherence)
    - 如何解释[s5] 和 [s6]
    - 一种推断：“他希望李四请他看电影”（可能需要更大的上下文）

# Cohesion vs. Coherence

- Cohesion(衔接): 强调其构成成分（主要是词或短语）之间的关联性.
  - [s1]张三喜欢**骑单车**上班, [s2]李四通常**步行**去办公室
  - 在词汇层面上相对容易处理
- Coherence(连贯): 强调整体上表达某种意义
  - [s3] A: 我有两张票, 想请你今晚看电影。
  - [s4-1] B:很遗憾, 我今晚不能看电影（衔接+连贯, 简洁易懂）
  - [s4-2] B:我还有一大堆的作业没有完成（连贯, 没有衔接）
  - [s4-3] B:我就不客气了（连贯, 没有衔接）
  - [s4-4] B: 武汉又称江城（不衔接、不连贯）
  - 在处理上相对困难, 不容易切入

# 篇章分析的假设

- 篇章分析：也称为文本分析(Text analysis)，或者文章分析
- 一篇待分析的文章假定为“合理”的，其“合理”性应表现在是否围绕某个话题或“意义”而展开，这就是所谓的连贯性。
- 一篇待分析的文章假定为“简洁易懂”的，其“简洁易懂”不仅表现为连贯，也表现为衔接。

# 衔接的进一步解释

- **Cohesion:** Five cohesive relations (Halliday & Hasan, 1976)
  - Reference (指代)
  - Substitution (替换)
  - Ellipsis (省略)
  - Conjunction (连接)
  - Lexical cohesion (词汇衔接)
- 语篇中为什么会有衔接现象？
  - 追求表达的经济（省略、指代）；
  - 追求表达的变化（指代、替换、词汇衔接）。
- Assumption: One word one sense per discourse
- Word sense(meaning)
  - Reiteration with the same word(s);
  - Reiteration without the same word(s);
  - Hyponymy & meronymy;
  - collocation



# 词汇衔接的例子

- **社交**的吃饭种类虽然复杂，性质极其简单。把饭给自己有饭的人吃，那是请饭；自己有饭可吃而去吃人家的饭，那是赏面子。交际的微妙不外乎此。反过来说，把饭给没饭吃的人吃，那是施食，赏面子就一变而成丢脸。这便是慈善救济，算不上**交际**了。（钱钟书：《吃饭》）
- 起衔接作用的词
  - 饭
  - 交际（社交）
  - 面子（赏面子、丢脸）
  - 施舍（施食、救济）
  - 复杂（简单）
- 应用：通过衔接关系，可以用于提取**文本的关键词**

# 为什么需要指代？

- 假设有这样一组句子：
  - 张三一大早就赶到了学校。张三先到食堂吃早餐，然后张三到张三的宿舍拿张三自己的教材和张三自己的笔记本。当张三匆忙来到教室时，张三发现张三的课本拿错了。
- 设想修改为这样表达：
  - 张三一大早就赶到了学校。他先到食堂吃早餐，然后[X]到[X]宿舍拿自己的教材和[X]笔记本。当[X]匆忙来到教室时，他发现[X]课本拿错了。
- 哪一种表达更符合人们的习惯？
- 语言的表达追求“经济”与“变化”
- 不妨将指代、省略、替换都看称广义“指代”

# 指代

- 指代(anaphora) 的定义(Hirst, 1981)

*Anaphor*

*Antecedent or  
Referent*

- Anaphora is the device of making in discourse an abbreviated **reference** to some **entity** in the expectation that the perceiver will be able to disabbreviate the reference and thereby **determine the identity of the entity**.

*Resolution*

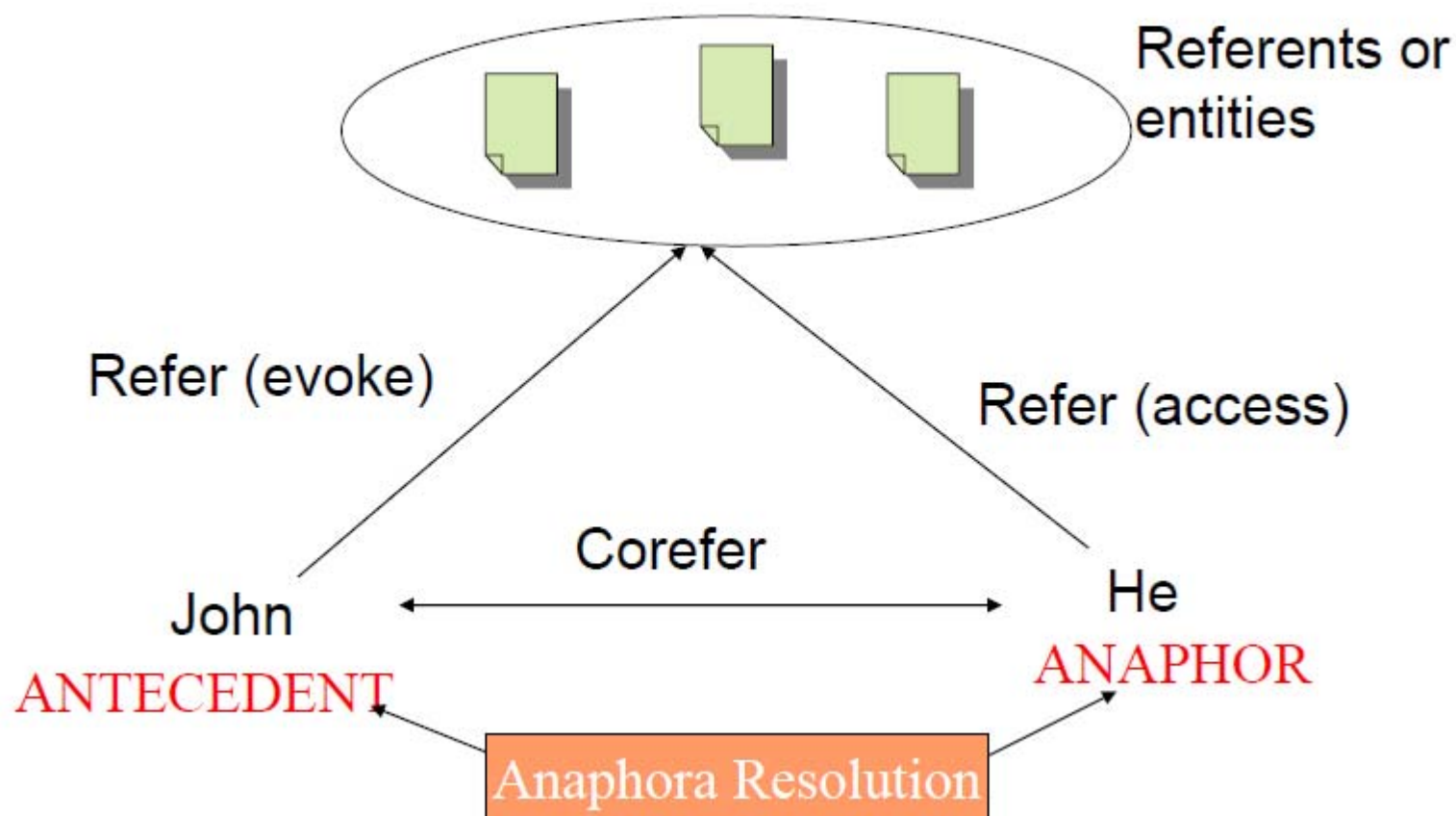
# 五个概念

- **Anaphor:** 指代语。当语篇中提到某个实体后，再一次提及时，常用一种简洁的形式表示（如代词“他”），这一简洁的形式称为指代语
- **Entity (referent):** 实体(指称对象)。实际存在或传说存在（如，孙悟空）的对象，主要包括，人、机构、地方等
- **Reference:** 指称。用于指称实体的语言表示
- **Antecedent:** 先行语。语篇中引入的一个相对明确的指称意义的表述（如张三）
- **Coreference:** 共指（同指）。当两种表述均指称相同对象（实体）时，这两种表述具有共指关系。

# 一个例子

- 我/rr 和/cz 黄/nrf 若噉/nrg 两/mx 个/qe 小青年/nap 病/vt 卧/vi 小龙坎/ns 的/ud 库房/nas, /wd 恩来/nr 同志/nap 亲自/d 把/p2 殷殷/z 亲情/nh 给予/vx 我们/rr, /wd 他/rr 的/ud 探视/vn、/wu 他/rr 的/ud 微笑/vn、/wu 他/rr 的/ud 火热/z、/wu 他/rr 的/ud 革命/aa 领袖/nap 的/ud 恩情/ne, /wd 永远/dt 珍藏/vt 在/ps 我/rr 的/ud 心中/smh。/wj

# 三角关系图



# 指代与共指

- 指代(Anaphora)关系：强调指代语与另一个表述之间的关系。指代语的指称对象通常不明确，需要确定其与先行语之间的关系来解释指代语的语义
  - 张先生走过来，给大家看他<sub>他</sub>的新作品
- 共指(coreference)：强调一个表述与另一个表述是否指向相同的 实体
  - 现任美国总统与奥巴马
- 两者的目标
  - 指代消解：寻找指代语对应的先行语
  - 共指消解：发现指向相同实体的语言表示单元（包括多语篇）

## 6类指称表示

- Indefinite NPs（无定名词）：一辆汽车
- Definite NPs（有定名词）：那个人
- Pronouns（人称代词）：它， 他
- Demonstratives（指示代词）：这， 那
- One-anaphora（one指代）：one (in English)
- Zero anaphora（0型指代）：省略



# 指称表示

- Indefinite NPs (无定名词)
  - 为读者引入一个新的实体时常用无定形式
  - 引入的实体，可能的确存在（明确的），也可能不明确
  - 张先生娶了一位法国太太 (Specific)
  - 史密斯想娶一位中国姑娘 (non-specific)
- Definite NPs (有定名词)
  - 无论读者知道否，一定存在
    - 首位进入太空的宇航员（即，前苏联宇航员尤里.加加林；通过某些知识可以知道）
    - Look, how beautiful the girl is! (实际存在)
    - 为了消除小兵兵对生人的陌生感，两位女记者带着这个小男孩逛街...(在上下文中)
  - 最后一种情况需要指代消解。
    - 特点：定冠词（这/那）引导的名词短语

# 指称表示

- Demonstratives (指示代词)
  - 典型的指示代词包括：那, 这,...
  - 当指示代词与后面的名词(短语)连用时，此时变为了定冠词，形成有定表示。
  - 例：刘博士刚买了一套房子，**那**是一套性价比相当好的房子。
- One-anaphora (替换)
  - 出现在英语中
  - 表示某集合中的一个元素
  - He had a BMW before. Now he got another **one**.
  - John has two BMWs, but I have only **one**.

# 替换

- The man who gave **his paycheck** to his wife was wiser than the man who gave **it** to his mistress.
- That's a **rhinoceros**.
- A what? Spell **it** for me.
- 刘博士**买的**是新房，张博士**买的**是二手的。
- 朋友陈把手一拍，我们便看见一只大鸟飞过去，接着又看见**第二只**，**第三只**。我们继续拍掌。很快这个树林变得热闹了。到处都是鸟声，到处都是鸟影。**大的，小的，花的，黑的，有的**站在树上叫，**有的**飞起来，**有的**在扑翅膀。  
(巴金：《鸟的天堂》)

# 省略 — 零指代(Zero anaphora)

例：张三一大早就赶到了学校。他先到食堂吃早餐，然后[X]到宿舍拿自己的教材和[X]笔记本。当[X]匆忙来到教室时，他发现[X]课本拿错了。

- 英语中的零指代很少见，但汉语中十分常见：
  - They said **they** were coming to help us with **our** house repair today.
  - 他们说[X]今天来帮我们修[X]房子
  - 他们说**他们**今天来帮我们修**我们的**房子（很少这样说）
  - 美国宣布（X）部分取消（X）对朝鲜长达近半个世纪的经济制裁。
  - 李向阳机智地组织游击队攻城并烧毁了敌人的粮库，（？）迫使松井撤出了李庄。
  - 我自来是如此，（X）从会吃饮食时便吃药，（X）到今未断。（X）请了多少名医，（X）修方配药，（X）皆不见效。
- 0形式的判断和消解
  - 需要在**句子层面**上判断哪些必须的成分省略了

# 以衔接为基础的篇章分析

- 分析单元：
  - 通常情况下是词
  - 有时也可以是短语
- 建立词汇之间的关系：
  - 形成词汇链（或词汇集合）
- 词汇链的形式定义：
  - 设文本T可以表示为词的集合  $T = \{w_1, w_2, \dots, w_n\}$
  - 设衔接关系为 R，则 R 将 T 划分为：
  - $CL_1 = \{w_{11}, w_{12}, \dots, w_{1m_1}\}, CL_2 = \{w_{21}, w_{22}, \dots, w_{2m_2}\}, \dots$  其中，对任意的  $w_{kp}, w_{kq} \in CL_k$ ，都有  $(w_{kp}, w_{kq}) \in R$ 。
  - R 可以看成为广义“等价”关系。

# 如何建立衔接关系

- 分析五种关系
  - Reference（指代）
  - Substitution（替换：发现替换关系）
  - Ellipsis（省略：找回省略部分）
  - Conjunction（连接，主要在连贯分析中使用）
  - Lexical cohesion（词汇衔接）
    - 重复（词的形式判断）
    - 近义+反义（借助于词典）
    - 上下位义+整体部分义（借助于词典）
    - 搭配（词典+统计方法）

# 连贯 (coherence)

- Longman: a reasonable connection or relation between **ideas, arguments, statements**, etc. An overall theme will help to give your essay coherence.

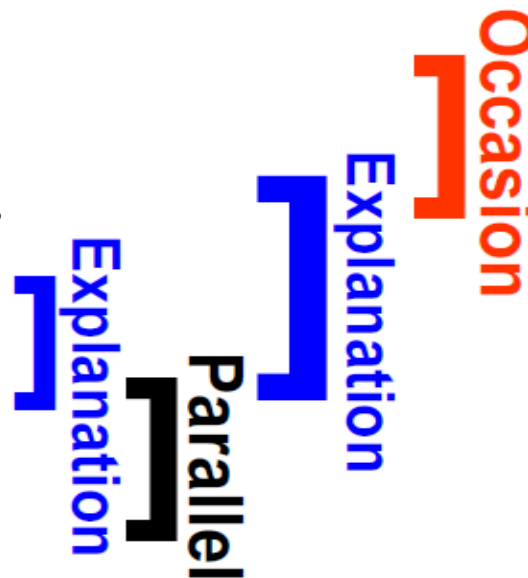
S0: 张三去银行办理支票.

S1: 然后他乘车到了李四的汽车销售店.

S2: 他想买一部车.

S3: 他的工作单位距公交站较远.

S4: 他也不想同李四讨论一下他们的垒球协会的事情.



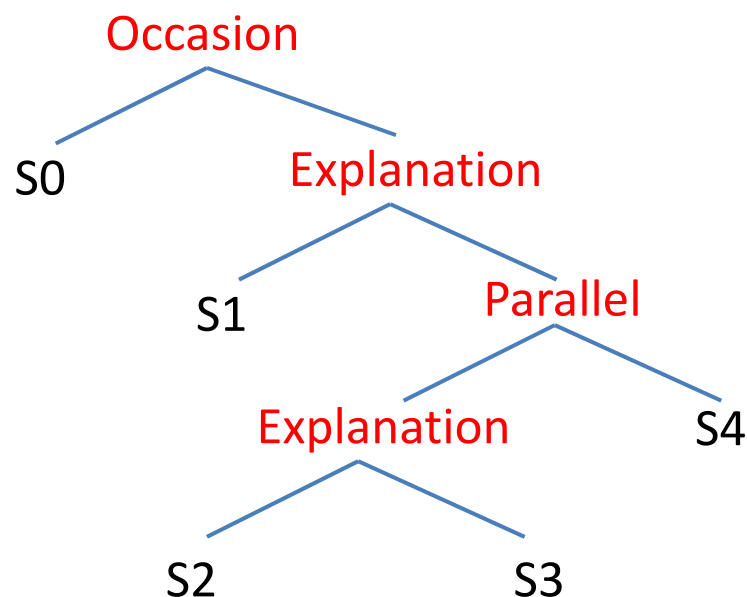
# 连贯关系(coherence relation)

- 语段（如句子）之间可能的语义连接关系称为连贯关系。
- Hobbs(1979)提出的连贯关系（设S0和S1为两个相关的句子的意义）：
  - 结果关系(result): 推测S0所声明的状态或事件（可能）导致S1所声明的状态或事件；
  - 解释关系(explanation): 推测S1所声明的状态或事件（可能）导致S0所声明的状态或事件；
  - 平行关系(parallel): 推测S0所声明的P(a1,a2,...)与S1所声明的P(b1, b2,...) 是类似的；
  - 细化关系(Elaboration): 推测S1和S0所声明的是同一命题P；
  - 时机关系(Occasion): 推测由S0所声明的状态到S1最终状态的变化，或者由S1所声明的状态到S0的最初状态的变化。



# 以连贯为基础的篇章结构分析

- 建立句间语义关系



# RST(Rhetorical structure theory)

- 修饰结构理论：认为语篇的构成具有层次结构关系（树形图），通过修饰结构表示语篇结构
- 理论的建立者为： William Mann and Sandra Thompson, 1987（南加州大学）
- 层次结构关系由修饰关系刻画
- 修饰关系是对前面Hobbs连贯关系的细化
  - 共23种关系；
  - 关系的双方： **Nucleus** 与 **Satellite**
    - 具有支配作用： **Nucleus + Satellite**
    - 平等关系： **Nucleus + Nucleus**

# RST中的关系

## Subject matter (informational)

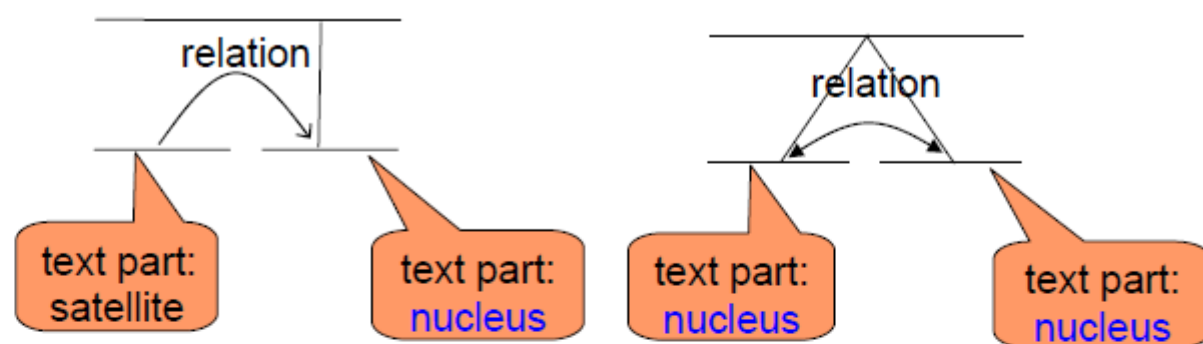
- Elaboration
- Circumstance
- Solutionhood
- Volitional Cause
- Volitional Result
- Non-Volitional Cause
- Non-Volitional Result
- Purpose
- Condition
- Otherwise
- Interpretation
- Evaluation
- Restatement
- Summary
- Sequence
- Contrast

## Presentational (intentional)

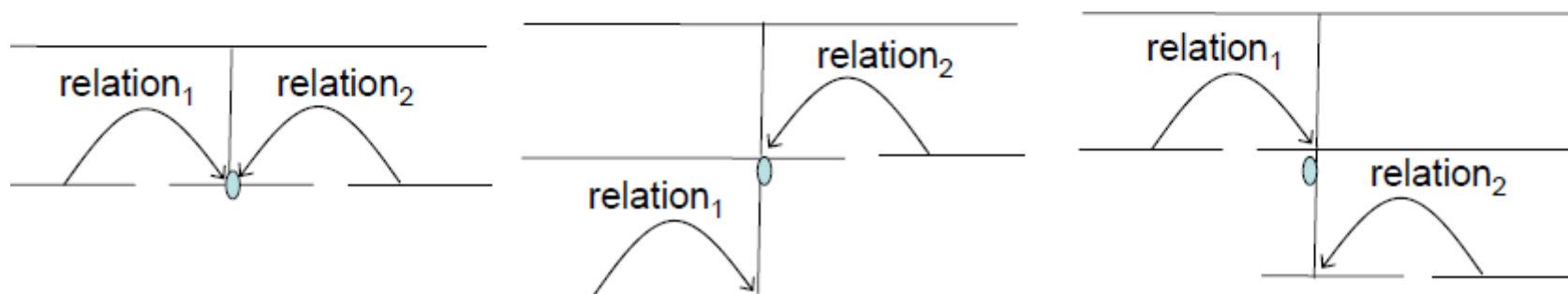
- Motivation
- Antithesis
- Background
- Enablement
- Evidence
- Justify
- Concession

# 基本关系模式

- 二元关系



- 多元素关系



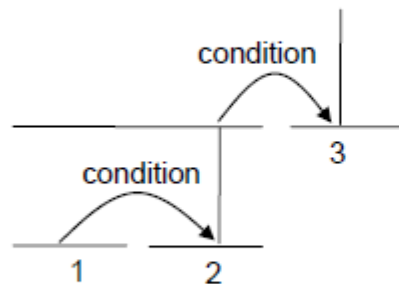
# 基于RST的分析：问题

- 语篇中究竟需要多少关系以及需要什么样的关系？
  - 没有统一的标准
- 两个片段（句子）之间可能存在多种解释
  - 不同的解释都能接受
- 如何确认两个片段之间的关系？
  - 并不是一件容易的事（很多情况下没有形式标记，需要靠意义确定关系）
- 构造树结构的复杂性高

# 多种解释

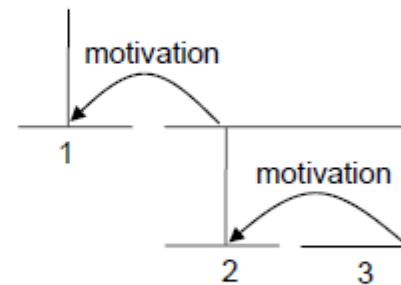
- 1. *Come back at 5:00*
- 2. *Then we can go to the hardware store before it closes.*
- 3. *This way we can finish the bookshelves tonight.*

## Informational level



**Condition:** The satellite presents a situation which is necessary for the nucleus to obtain.

## Intentional level



**Motivation:** Satellite presents information which should make the reader want to perform the action in the nucleus

# 中心理论 CT(Centering Theory)

- 提出者： **Grosz Barbara**
- 特点：
  - 是一种局部化的语篇连贯性理论；
  - 解释了为什么某个语篇比另一个语篇在处理（理解）上更困难；
  - 解释了为什么会以这种方式使用代词而不是用其他方式；
  - 给出了指代消解的一种实用化方法。

# 语篇比较—哪一段更容易理解？

- a. *Jeff<sub>1</sub> helped Dick<sub>2</sub> wash the car.*
- b. ***He<sub>1</sub> washed the windows as Dick<sub>2</sub> waxed(擦亮) the car.***
- c. ***He<sub>1</sub> soaped a pane (玻璃) .***

- a. *Jeff<sub>1</sub> helped Dick<sub>2</sub> wash the car.*
- b. ***He<sub>1</sub> washed the windows as Dick<sub>2</sub> waxed the car.***
- c. ***He<sub>2</sub> buffed (擦亮) the hood (发动机罩) .***

- 从句子的关系和意义上看，哪一段更连贯呢？
- 原因：第一段的中心(Center) 没有变，一直是Jeff； 在第2段中，C的中心 He 变为了Dick



# 中心可以帮助消解代词歧义

- 代词消解（指代消解）：确定代词的所指过程
- 一个例子：
  1. Susan<sub>1</sub> is a fine friend.
  2. She<sub>1</sub> gives people the most wonderful presents.
  3. She<sub>1</sub> just gave Betsy<sub>2</sub> a wonderful bottle of wine.
  4. She<sub>1</sub> told her<sub>2</sub> it was quite rare.
  5. She<sub>1</sub> knows a lot about wine.
- 为什么后四个句子中的 she 都表示 Susan？为什么 her 表示 Betsy？
- 中心理论可以给出合理的解释！

## 另一个例子详解

- a. *Terry really goofs sometimes.*
- b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
- c. *He wanted Tony to join on a sailing expedition.*
- d. *He called him at 6 A.M.*

- 语篇的正常表述！
- 其中，**He** 是谁？ **him** 是谁？ 为什么？

## To continue...

- 后面再增加一个句子，得到语篇：
  - a. Terry really goofs sometimes.*
  - b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - c. He wanted Tony to join on a sailing expedition.*
  - d. He called him at 6 A.M.*
  - e. He was sick and furious at being woken up so early.*
- 语篇的表述似乎不太正常！
- 其中，最后一个 **He** 是谁？

## To continue...

- 假设后面增加句子变成为：
  - a. Terry really goofs sometimes.*
  - b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - c. He wanted Tony to join on a sailing expedition.*
  - d. He called him at 6 A.M.*
  - e. Tony was sick and furious at being woken up so early.*
- 语篇的表述又可以接受！ 其中的变化是：最后一个 **He** 改成了 **Tony**

# To continue...

- 假设后面再增加一个句子:
  - a. *Terry really goofs sometimes.*
  - b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - c. *He wanted Tony to join on a sailing expedition.*
  - d. *He called him at 6 A.M.*
  - e. *Tony was sick and furious at being woken up so early.*
  - f. *He told Terry to get lost and hung up.*

- 语篇的表述可以接受！ 最后一个 **He** 指代谁？

# To continue...

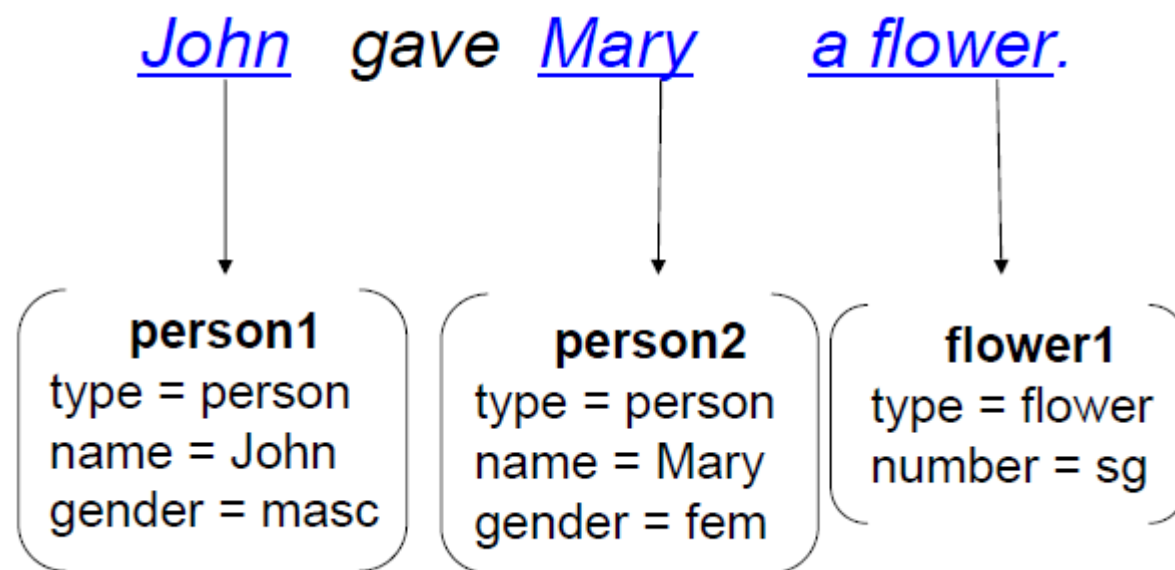
- 在后面进一步增加句子：
  - a. *Terry really goofs sometimes.*
  - b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - c. *He wanted Tony to join on a sailing expedition.*
  - d. *He called him at 6 A.M.*
  - e. *Tony was sick and furious at being woken up so early.*
  - f. *He told Terry to get lost and hung up.*
  - g. *Of course **he** hadn't intended to upset Tony.*
- 表述似乎又有问题！ 问题出在最后一个 **he** 上？

## To continue...

- 最后面句子再修改:
  - a. *Terry really goofs sometimes.*
  - b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - c. *He wanted Tony to join on a sailing expedition.*
  - d. *He called him at 6 A.M.*
  - e. *Tony was sick and furious at being woken up so early.*
  - f. *He told Terry to get lost and hung up.*
  - g. *Of course **Terry** hadn't intended to upset Tony.*
- 此次的语篇看起来怎么样？

# 中心理论的进一步解释

- 中心：语篇中的实体之一
- 什么是实体？





# 如何界定中心？

CT 理论提出了三类中心

- 前瞻中心表(a list of forward-looking centers)
  - 句子 $u$ 的前瞻中心表是句中实体有序集  $C_f(u) = \langle e_1, e_2, \dots, e_k \rangle$
  - 其中，实体的排序规律为：
    - subject > direct-object > indirect-object > others
- 回看中心(a backward-looking center)
  - 句子 $u$ 中的回看中心 $C_b(u)$ 是出现在 $u$ 中，且在前面句子中排顺最靠前的实体；
- 优先中心(a preferred center)
  - 句子 $u$ 中的优先中心 $C_p(u)$ 是 $C_f(u)$ 中排序最靠前的实体。

# 中心转换关系

	$C_b(u) = C_b(u-1)$	$C_b(u) \neq C_b(u-1)$
$C_b(u) = C_p(u)$	<b>CONTINUING</b>	<b>SMOOTH SHIFT</b>
$C_b(u) \neq C_p(u)$	<b>RETAINING</b>	<b>ABRUPT SHIFT</b>

- 连贯性比较

***CON > RET > SSH > ASH***

# 基于中心的连贯性比较

- U1. John went to his favorite music store to buy a piano.
  - U2. He had frequented the store for many years.
  - U3. He was excited that he could finally buy a piano.
  - U4. He arrived just as the store was closing for the day.
- 
- U1. John went to his favorite music store to buy a piano.
  - U2. It was a store John had frequented for many years.
  - U3. He was excited that he could finally buy a piano.
  - U4. It was closing just as John arrived.
- 
- 由中心理论可以推断，第一段比第二段连贯

# 第一段

- U1. John went to his favorite music store to buy a piano.
  - $C_f(U1) = \langle \text{John, store, piano} \rangle$ .
- U2. He had frequented the store for many years.
  - $C_b(U2) = \text{John}$ .  $C_f(U2) = \langle \text{John, store} \rangle$ .
  - **CONTINUATION.**
- U3. He was excited that he could finally buy a piano.
  - $C_b(U3) = \text{John}$ .  $C_f(U3) = \langle \text{John, piano} \rangle$ .
  - **CONTINUATION.**
- U4. He arrived just as the store was closing for the day.
  - $C_b(U4) = \text{John}$ .  $C_f(U4) = \langle \text{John, store} \rangle$ .
  - **CONTINUATION.**

## 第二段

- U1. John went to his favorite music store to buy a piano.
  - $C_f(U1) = (\text{John}, \text{store}, \text{piano})$ .
- U2. It was a store John had frequented for many years.
  - $C_b(U2) = \text{John}$ .  $C_f(U2) = (\text{store}, \text{John})$ .
  - **RETAINING.**
- U3. He was excited that he could finally buy a piano.
  - $C_b(U3) = \text{John}$ .  $C_f(U3) = (\text{John}, \text{piano})$ .
  - **CONTINUATION.**
- U4. It was closing just as John arrived.
  - $C_b(U4) = \text{John}$ .  $C_f(U4) = (\text{store}, \text{John})$ .
  - **RETAINING.**

# 尝试比较下面两段

- a. *Jeff<sub>1</sub> helped Dick<sub>2</sub> wash the car.*
  - b. ***He<sub>1</sub> washed the windows as Dick<sub>2</sub> waxed(擦亮) the car.***
  - c. ***He<sub>1</sub> soaped a pane (玻璃) .***
- 
- a. *Jeff<sub>1</sub> helped Dick<sub>2</sub> wash the car.*
  - b. ***He<sub>1</sub> washed the windows as Dick<sub>2</sub> waxed the car.***
  - c. ***He<sub>2</sub> buffed (擦亮) the hood (发动机罩) .***

# 基于CT的指代消解算法

规则：

- 如果  $C_f(u_{i-1})$  的某元素以代词形式出现在  $u_i$ , 那么, 这个元素就是  $C_b(u_i)$
- 规则给出了凸显性的直观解释, 即被代词表示的实体 具有显著性 (一目了然)
- 如果有多个代词, 那么其中之一是  $C_b(u_i)$
- 如果只有一个代词, 那么一定是  $C_b(u_i)$

解释:  $C_b(u_i)$  的确定依赖于两个条件:

- (1) 一定是在  $u_i$  中出现的语义实体;
- (2) 该实体也一定在  $C_f(u_{i-1})$  中出现过, 如果  $u_i$  有多个实体也在  $u_{i-1}$  中出现, 那么, 作为  $C_b(u_i)$  的实体在  $C_f(u_{i-1})$  中应有更高的排位。

# BFP算法

- BFP(Brennan, Friedman and Pollard,1987)

– 步骤:

- **Step1.** 如果在 $u_i$ 中出现人称代词，则自左至右顺序检验 $C_f(u_{i-1})$ 中的元素，直至同时满足词汇句法（Morphosyntactic）、约束（Binding）和类型标准（Sortal criteria）；这样的元素作为先行语；
- **Step2.** 完全读取表述 $u_i$ ，计算 $C_b(u_i)$ 并生成 $C_f(u_{i-1})$ ，对 $C_f(u_{i-1})$ 进行排序。



# 例子解释

- a. *Terry really goofs sometimes.*
  - $C_f = ([\mathbf{Terry}])$
- b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*
  - $C_f = (he=his=[\mathbf{Terry}], [\mathbf{the\ sailboat}])$
  - $C_b = [\mathbf{Terry}]$
- c. *He wanted Tony to join on a sailing expedition (划艇队) .*
  - $C_f = (he=[\mathbf{Terry}], [\mathbf{Tony}], [\mathbf{the\ expedition}])$
  - $C_b = [\mathbf{Terry}]$
- d. *He called him at 6 A.M.*
  - $C_f = (he=[\mathbf{Terry}], him=[\mathbf{Tony}])$
  - $C_b = [\mathbf{Terry}]$

# 再看例子变形

a. *Terry really goofs sometimes.*

–  $C_f = ([\text{Terry}])$

b. *Yesterday was a beautiful day and he was excited about trying out his new sailboat.*

–  $C_f = (\text{he}=\text{his}=[\text{Terry}], [\text{the sailboat}])$

–  $C_b = [\text{Terry}]$

c. *He wanted Tony to join on a sailing expedition.*

–  $C_f = (\text{he}=[\text{Terry}], [\text{Tony}], [\text{the expedition}])$

–  $C_b = [\text{Terry}]$

d. *Terry called him at 6 A.M.*

–  $C_f = ([\text{Terry}], \text{him}=[\text{Tony}])$  Disobeyed (violation)

–  $C_b = [\text{Terry}] \neq$

不如前面连贯，实际上很少使用这种表达

# 中心理论的问题

- 属于局部连贯性（通过相邻句子的中心变化表征），跨越多个句子的指代消解如何处理？
- 中心理论要求单位是utterance(没有明确界定为句子-sentence/clause)，什么是utterance，特别是在汉语中如何界定？
- Cf中的排列顺序目前只用到了表层信息，是否还有深层信息（如语义）可用？
- 汉语中大量存在0-指代，如何处理

# 指代消解的其他方法

## 基于语言知识

- 过滤原则：
  - 性别、单复数和人称的一致性规则；
- 优选原则：
  - 距离近优先
  - 句法、语义平行优先
  - 例子：
    - **A** 喜欢与 **B** 闲聊，**他** 也喜欢与 **C** 闲聊
    - **A** 喜欢与 **B** 闲聊，**C** 也喜欢与 **他** 闲聊
  - 算法：Lappin & Leass 提出的算法RAP（处理单数三人称代词，略）

# 基于分类（ML）的方法

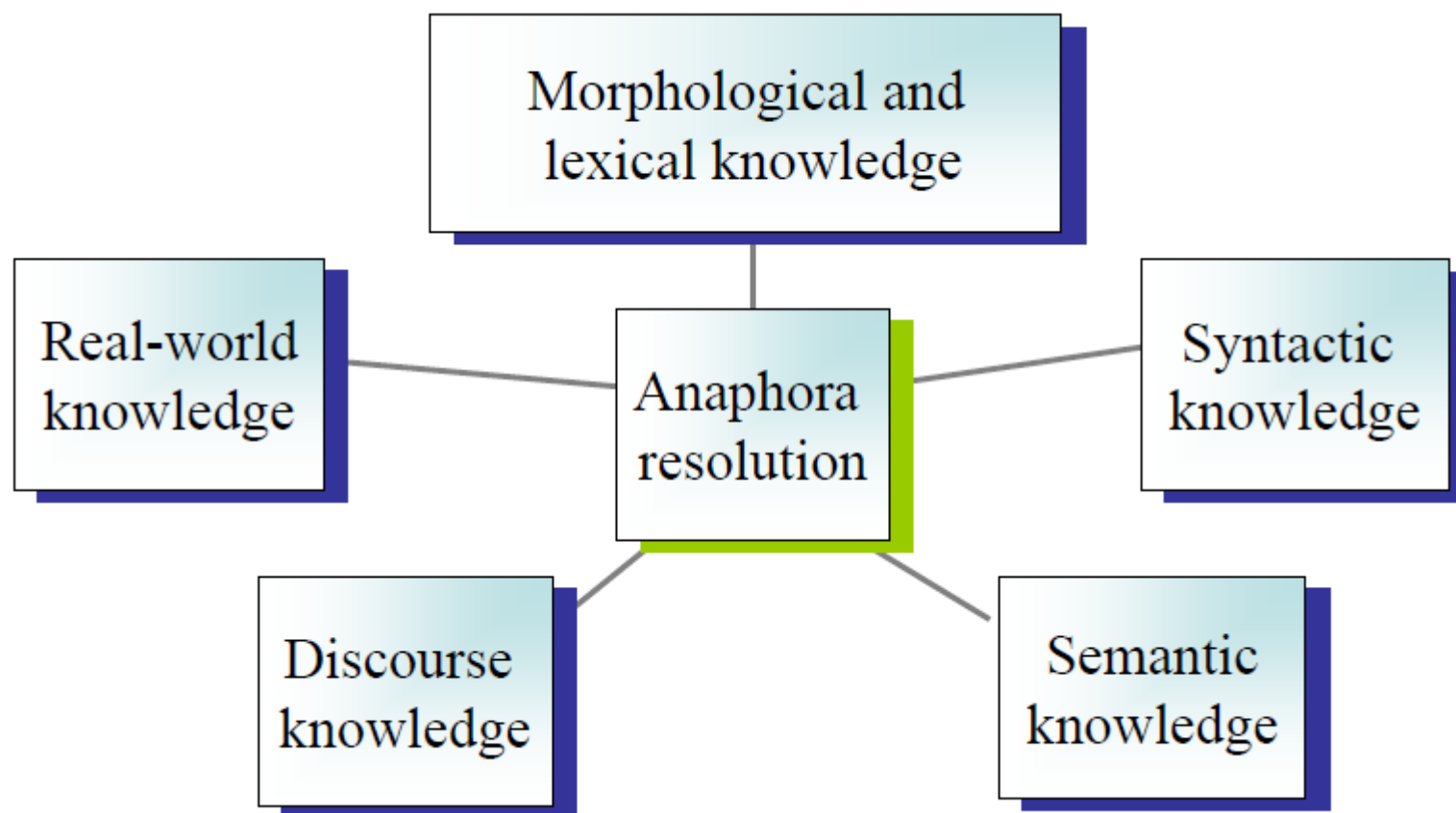
- 利用机器学习方法建立分类器：
- 方法：
  - 选取对共指消解产生影响的特征，主要包括：
  - 两者的距离, 字符的匹配程度, 单复数一致性, 性别一致性, 语义类的一致性, 是否是别称....

[聂/nr 卫平/nr] 今天/t 取胜/v 不易/a 。/w 布局/vn 阶段/n 便/d 与/p  
[实力派/n 人物/n] [刘/nr 小光/nr] 九/m 段/q 展开/v 激战/vn ，/w 棋  
局/n 跌宕起伏/l ，/w 互/d 有/v 优劣/n 。/w 直到/v 官子/vn 阶段  
/n ，/w [聂/nr 卫平/nr] 才/d 因/c [对手/n] 的/u 缓/a 手/n 而/c  
最终/d 取胜/v 。/w

[对手/n] => [聂/nr 卫平/nr]      属于一类吗 ？

[对手/n] => [刘/nr 小光/nr]      属于一类吗 ？

# 指代消解的困难



# 汉语中的指代消解困难

几类典型的问题：

- 0-指代（省略）如何识别？
  - 例：张三对[]弟弟保护得很好，[]每次出去，[]都是牵着[]弟弟的手。
- 如何识别可能的指称语（除了人名、代词之外，还有其他吗？如，【【美国】总统】即将访华）
- 抽象指代问题 美国的一些学者认为中国强大后必然走上一条对外扩张的道路，其实这完全是一种误解

# 指代消解的应用

- 机器翻译
  - They 是翻译成“他们”，“她们”， 还是“它们”？
- 文本摘要：
  - 理解原文本时，需要理解代词的对应关系
  - 生成摘要时，避免名字（同一个词）的反复使用，用代词（或O-形式）表示，以便符合习惯
- 信息抽取：
  - 识别文本中的实体，建立实体之间的关系
  - 实体常常用代词表示，关系的建立需要明确代词的指向
- .....