Natural Language Processing
SoSe 2016

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Discourse Analysis

Dr. Mariana Neves

June 6th, 2016

Outline

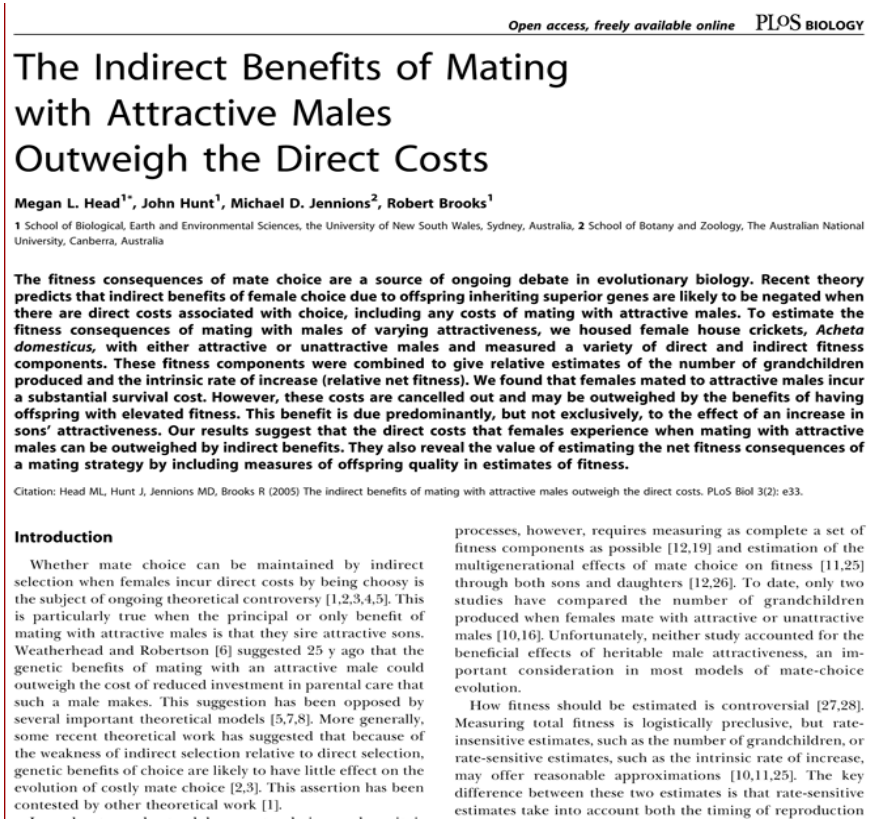
- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

Outline

- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

Discourse

- Discourse is a **coherent structured** group of sentence.



Open access, freely available online PLoS BIOLOGY

The Indirect Benefits of Mating with Attractive Males Outweigh the Direct Costs

Megan L. Head[1*], John Hunt[1], Michael D. Jennions[2], Robert Brooks[1]

1 School of Biological, Earth and Environmental Sciences, the University of New South Wales, Sydney, Australia, 2 School of Botany and Zoology, The Australian National University, Canberra, Australia

The fitness consequences of mate choice are a source of ongoing debate in evolutionary biology. Recent theory predicts that indirect benefits of female choice due to offspring inheriting superior genes are likely to be negated when there are direct costs associated with choice, including any costs of mating with attractive males. To estimate the fitness consequences of mating with males of varying attractiveness, we housed female house crickets, *Acheta domesticus,* with either attractive or unattractive males and measured a variety of direct and indirect fitness components. These fitness components were combined to give relative estimates of the number of grandchildren produced and the intrinsic rate of increase (relative net fitness). We found that females mated to attractive males incur a substantial survival cost. However, these costs are cancelled out and may be outweighed by the benefits of having offspring with elevated fitness. This benefit is due predominantly, but not exclusively, to the effect of an increase in sons' attractiveness. Our results suggest that the direct costs that females experience when mating with attractive males can be outweighed by indirect benefits. They also reveal the value of estimating the net fitness consequences of a mating strategy by including measures of offspring quality in estimates of fitness.

Citation: Head ML, Hunt J, Jennions MD, Brooks R (2005) The indirect benefits of mating with attractive males outweigh the direct costs. PLoS Biol 3(2): e33.

Introduction

Whether mate choice can be maintained by indirect selection when females incur direct costs by being choosy is the subject of ongoing theoretical controversy [1,2,3,4,5]. This is particularly true when the principal or only benefit of mating with attractive males is that they sire attractive sons. Weatherhead and Robertson [6] suggested 25 y ago that the genetic benefits of mating with an attractive male could outweigh the cost of reduced investment in parental care that such a male makes. This suggestion has been opposed by several important theoretical models [5,7,8]. More generally, some recent theoretical work has suggested that because of the weakness of indirect selection relative to direct selection, genetic benefits of choice are likely to have little effect on the evolution of costly mate choice [2,3]. This assertion has been contested by other theoretical work [1].

processes, however, requires measuring as complete a set of fitness components as possible [12,19] and estimation of the multigenerational effects of mate choice on fitness [11,25] through both sons and daughters [12,26]. To date, only two studies have compared the number of grandchildren produced when females mate with attractive or unattractive males [10,16]. Unfortunately, neither study accounted for the beneficial effects of heritable male attractiveness, an important consideration in most models of mate-choice evolution.

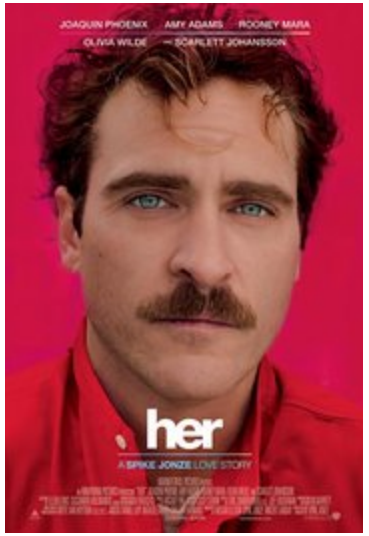
How fitness should be estimated is controversial [27,28]. Measuring total fitness is logistically preclusive, but rate-insensitive estimates, such as the number of grandchildren, or rate-sensitive estimates, such as the intrinsic rate of increase, may offer reasonable approximations [10,11,25]. The key difference between these two estimates is that rate-sensitive estimates take into account both the timing of reproduction

(http://evolution.berkeley.edu/evolibrary/teach/journal/dissectingapaper2.php)

Types of Discourse

- **Monologues**

- Dialogue
  - Human-human
  - Human-computer (conversational agent)



(http://www.imdb.com/title/tt1798709/)

Motivation: Information extraction

- Reference resolution

Angelina Jolie Pitt is an American actress, filmmaker, and humanitarian. **She** has received an Academy Award, two Screen Actors Guild Awards, and three Golden Globe Awards, and has been cited as Hollywood's highest-paid actress.

…

Divorced from actors Jonny Lee Miller and Billy Bob Thornton, **she** has been married to actor Brad Pitt since 2014. **They** have six children together, **three of whom** were adopted internationally.

(https://en.wikipedia.org/wiki/Angelina_Jolie)

Motivation: Summarization

- Text coherence and reference resolution

"To review available studies of empagliflozin, a **sodium glucose co-transporter-2 (SGLT2)** inhibitor approved in 2014 by the European Commission and the United States Food and Drug Administration for the treatment of type 2 diabetes mellitus (T2DM). Inhibitors of the **sodium-glucose co-transporter 2 (SGLT2)** promote the excretion of glucose to reduce glycated hemoglobin (HbA1c) levels."

"this protein"

"SGLT2"

Motivation: Conversational agents

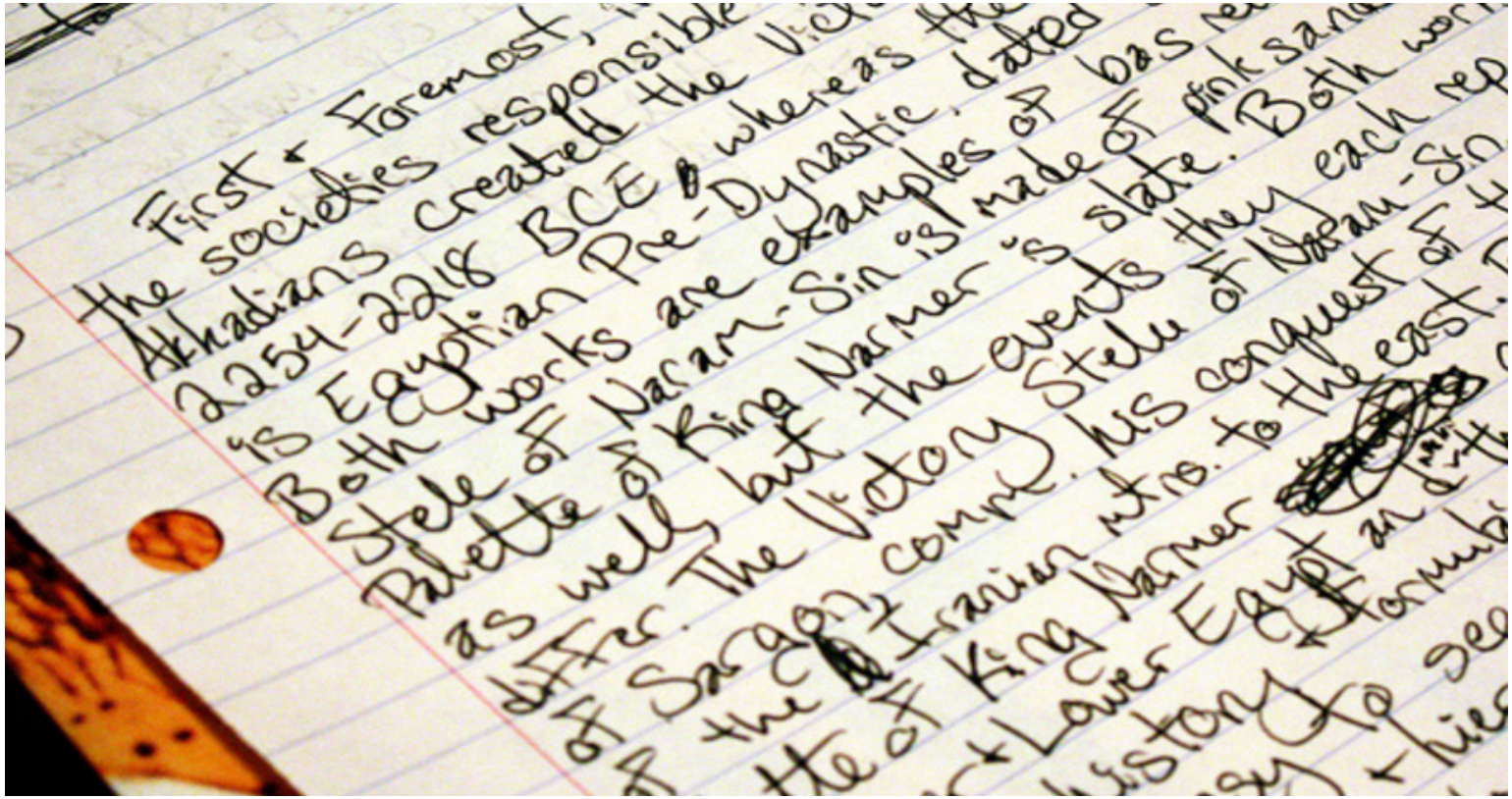
- Reference resolution and text coherence

- Goog morning, I want to fly to Denver.
- When do you want it to leave?
- Next Thursday.
- An early or a late flight?
…..

- Goog morning, I want a flight to Denver.
- Which kind of hotel do you want to book?

Motivation: Automatic Essay Grading

- Text coherence

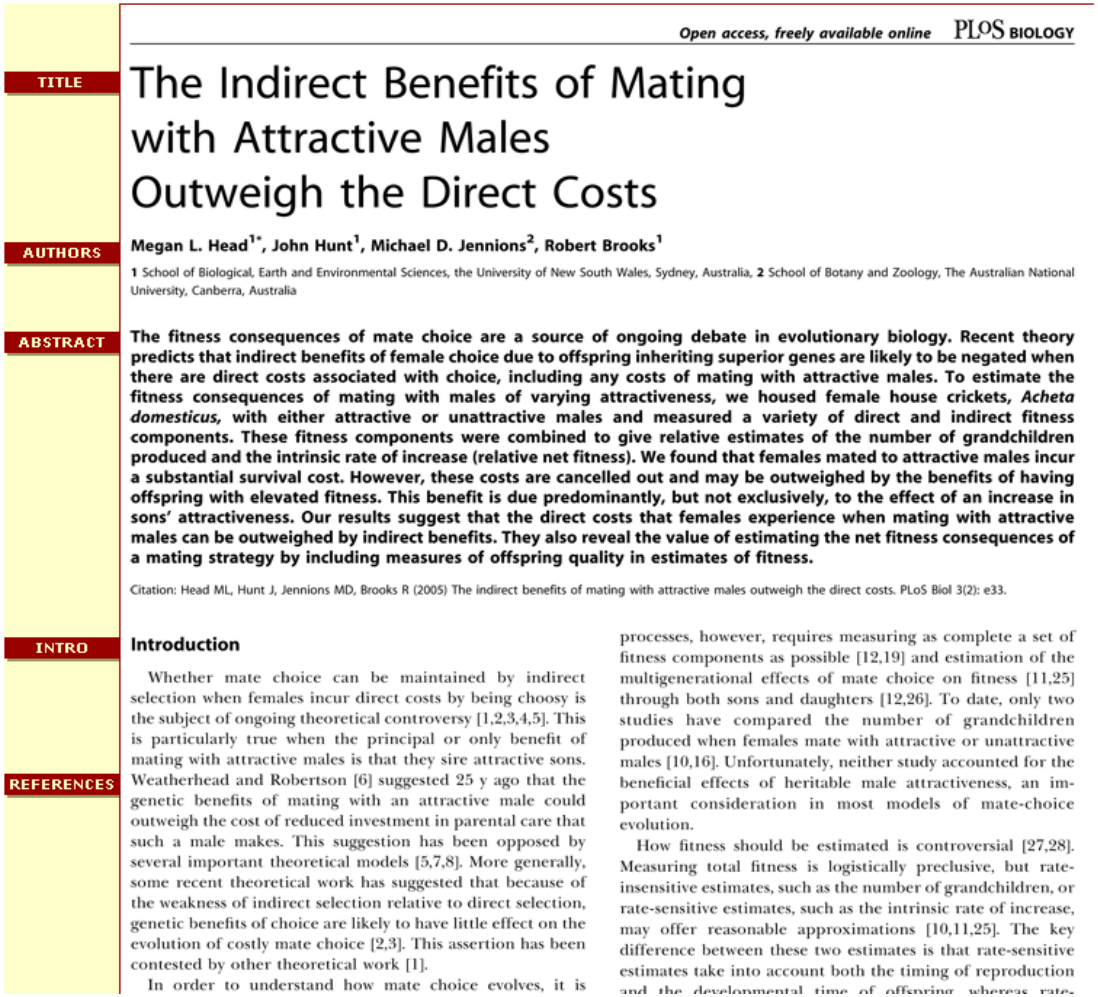


(http://www.educationnews.org/technology/idea-works-using-automated-grading-for-collaborative-learning/)

Outline

- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

Text Segmentation



TITLE

The Indirect Benefits of Mating with Attractive Males Outweigh the Direct Costs

AUTHORS

Megan L. Head[1*], John Hunt[1], Michael D. Jennions[2], Robert Brooks[1]

1 School of Biological, Earth and Environmental Sciences, the University of New South Wales, Sydney, Australia, 2 School of Botany and Zoology, The Australian National University, Canberra, Australia

ABSTRACT

The fitness consequences of mate choice are a source of ongoing debate in evolutionary biology. Recent theory predicts that indirect benefits of female choice due to offspring inheriting superior genes are likely to be negated when there are direct costs associated with choice, including any costs of mating with attractive males. To estimate the fitness consequences of mating with males of varying attractiveness, we housed female house crickets, *Acheta domesticus*, with either attractive or unattractive males and measured a variety of direct and indirect fitness components. These fitness components were combined to give relative estimates of the number of grandchildren produced and the intrinsic rate of increase (relative net fitness). We found that females mated to attractive males incur a substantial survival cost. However, these costs are cancelled out and may be outweighed by the benefits of having offspring with elevated fitness. This benefit is due predominantly, but not exclusively, to the effect of an increase in sons' attractiveness. Our results suggest that the direct costs that females experience when mating with attractive males can be outweighed by indirect benefits. They also reveal the value of estimating the net fitness consequences of a mating strategy by including measures of offspring quality in estimates of fitness.

Citation: Head ML, Hunt J, Jennions MD, Brooks R (2005) The indirect benefits of mating with attractive males outweigh the direct costs. PLoS Biol 3(2): e33.

Open access, freely available online PLoS BIOLOGY

INTRO

Introduction

Whether mate choice can be maintained by indirect selection when females incur direct costs by being choosy is the subject of ongoing theoretical controversy [1,2,3,4,5]. This is particularly true when the principal or only benefit of mating with attractive males is that they sire attractive sons. Weatherhead and Robertson [6] suggested 25 y ago that the genetic benefits of mating with an attractive male could outweigh the cost of reduced investment in parental care that such a male makes. This suggestion has been opposed by several important theoretical models [5,7,8]. More generally, some recent theoretical work has suggested that because of the weakness of indirect selection relative to direct selection, genetic benefits of choice are likely to have little effect on the evolution of costly mate choice [2,3]. This assertion has been contested by other theoretical work [1].

In order to understand how mate choice evolves, it is processes, however, requires measuring as complete a set of fitness components as possible [12,19] and estimation of the multigenerational effects of mate choice on fitness [11,25] through both sons and daughters [12,26]. To date, only two studies have compared the number of grandchildren produced when females mate with attractive or unattractive males [10,16]. Unfortunately, neither study accounted for the beneficial effects of heritable male attractiveness, an important consideration in most models of mate-choice evolution.

How fitness should be estimated is controversial [27,28]. Measuring total fitness is logistically preclusive, but rate-insensitive estimates, such as the number of grandchildren, or rate-sensitive estimates, such as the intrinsic rate of increase, may offer reasonable approximations [10,11,25]. The key difference between these two estimates is that rate-sensitive estimates take into account both the timing of reproduction and the developmental time of offspring, whereas rate-

REFERENCES

(http://evolution.berkeley.edu/evolibrary/teach/journal/dissectingapaper2.php)

Motivation for Text Segmentation

- Information extraction
  - Some sections are more informative than others
- Summarization
  - Include information from all sections
- Information retrieval
  - Retrieve particular information from the correct section

Text Segmentation

- Linear segmentation (no hierarchy)

- Approaches
  - Unsupervised discourse segmentation
  - Supervised discourse segmentation

Unsupervised Discourse Segmentation

- Based on **cohesion**:linguistic devices to link textual units
- Lexical cohesion: given by relations between words
  - e.g., identical words, synonyms or hypernyms.
- **Cohesion chain**: sequence of related words.

Peel, core and slice the **pears** and the **apples**.
Add the **fruits** to the skillet.
When **they** are soft, ….

Unsupervised Discourse Segmentation

- TextTiling algorithm (Hearst 1997)
  - Tokenization
    - Lowercase conversion, stoplist removal and stemming
    - Create pseudo-sentences (e.g., length 20)
      - No real sentences!
  - Lexical score determination
  - Boundary identification

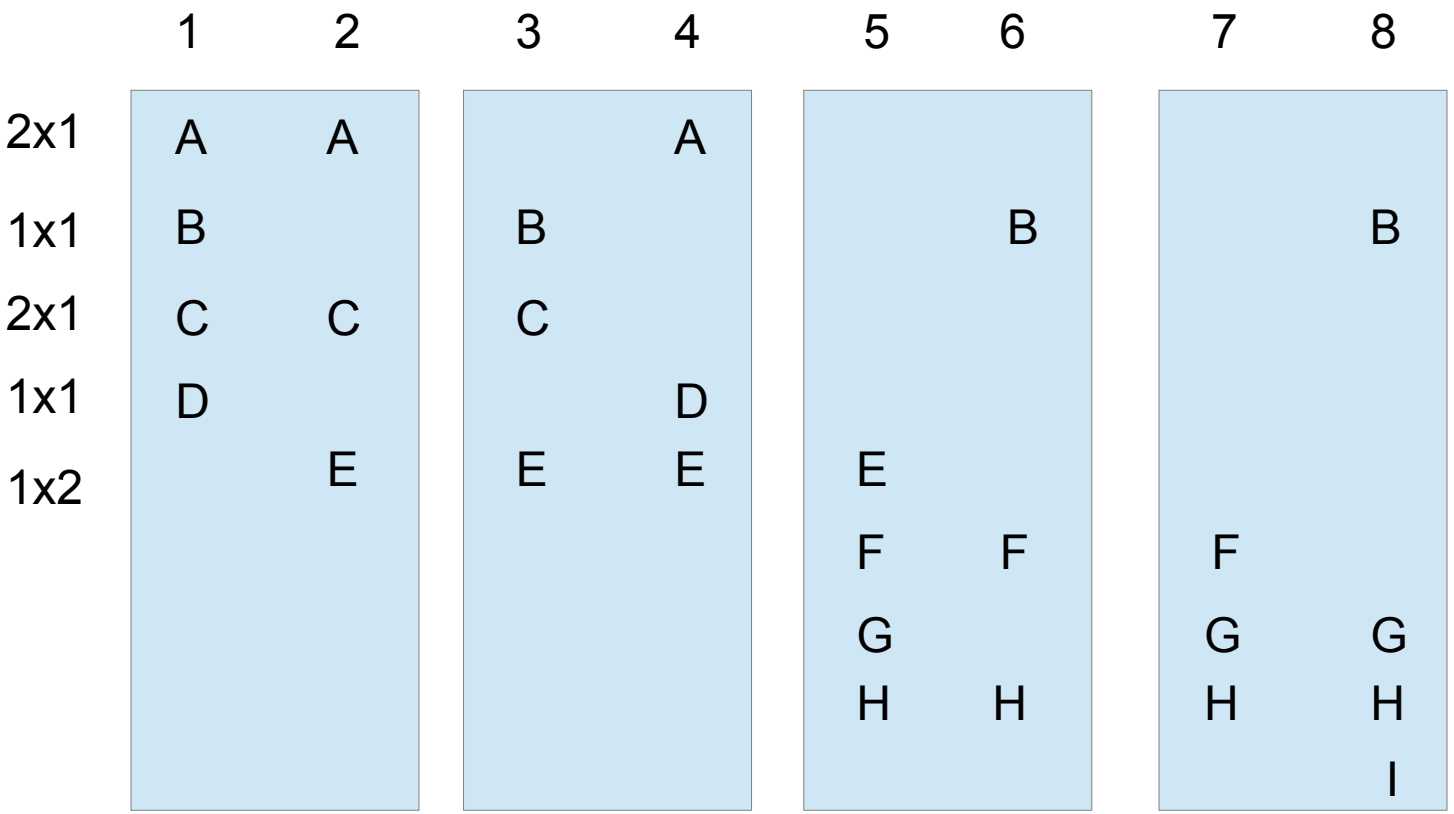
Unsupervised Discourse Segmentation

- TextTiling algorithm (Hearst 1997)
  - Tokenization
  - Lexical score determination
    - Average similarity of words in the pseudo-sentences
    - Create two vectors
      - Blocks of k pseudo-sentences before and after each gap
      - Calculate cosine similarity between the vectors
  - Boundary identification

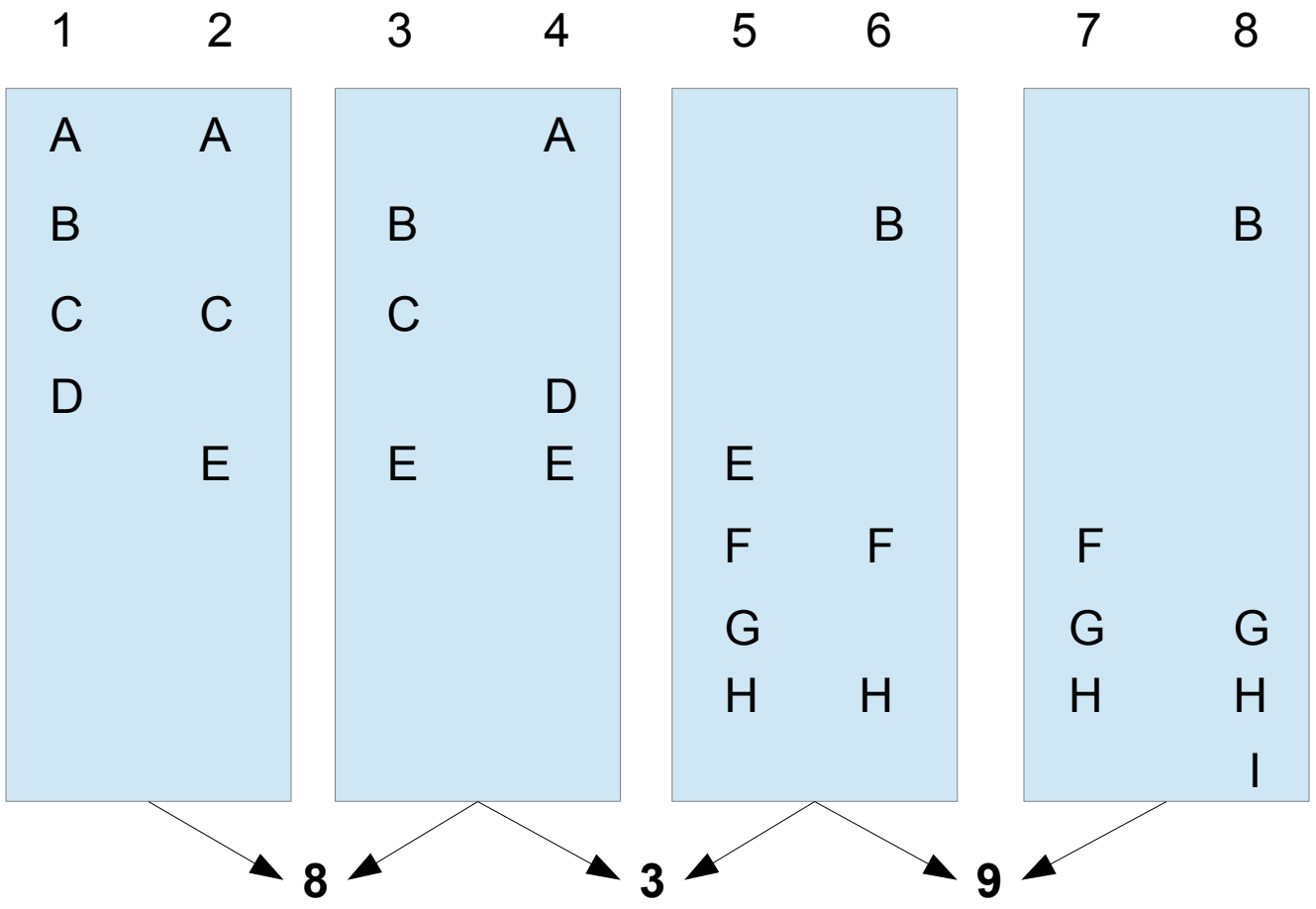
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

(https://en.wikipedia.org/wiki/Cosine_similarity)

Unsupervised Discourse Segmentation



Unsupervised Discourse Segmentation



Unsupervised Discourse Segmentation

- TextTiling algorithm (Hearst 1997)
  - Tokenization
  - Lexical score determination
  - Boundary identification
    - Compute depth score: distance from the peaks in both sides of the valley
      - $(y_{a1}-y_{a2})+(y_{a3}-y_{a2}) = (8-3)+(9-3)$
    - Define boundaries for valleys deeper than a cutoff threshold

Supervised Discourse Segmentation

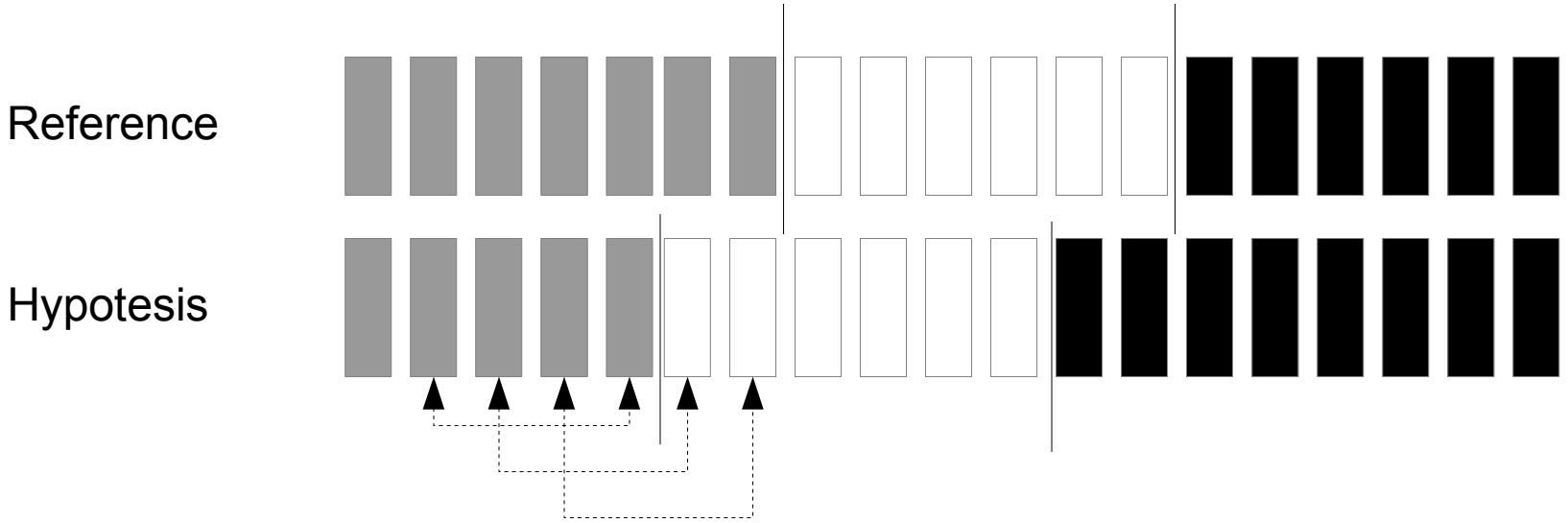
- There are labeled data available:
  - Paragraph segmentation (e.g., Web pages, <p> tag)
- Methods
  - Binaries classifiers (e.g., SVM, Naïve Bayes)
  - Sequential classifiers (HMM, CRF)
- Features
  - Word overlap, word cosine, Latent Semantic Analysis (LSA), lexical chains, coreference, etc
  - Discourse markers or cue words

Supervised Discourse Segmentation

- Discourse markers are domain-specific:
  - Broadcasting news: „Good evening, I'm ...“, „coming now...“, etc.
  - Scientific articles: „Introduction“, „Background“, „Methods“, „results“, etc.
  - Business: „XYZ incorporated“ then only „XYZ“

Evaluation of Text Segmentation

- WindowDiff [Pevzner and Hearst 2002]
  - Moving window of size „k“
  - „k“ is half the average segment in reference text
  - # boundaries in the probe: r_i (reference) and h_i (hypothesis)



Outline

- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

Text Coherence

- Meaning relation between two units.
  - The meaning of different units can combine to build a discouse meaning for the larger unit.

John hid Bill's car keys. He was drunk. ✓

John hid Bill's car keys. He likes spinach. ✗

Text coherence

- Better if the focus is on one entity

a. **John** went to **his** favorite music store to buy a piano.
b. **He** had frequented the store for many years.
c. **He** was excited that he could finally buy a piano. ✓
d. **He** arrived just as the store was closing for the day.

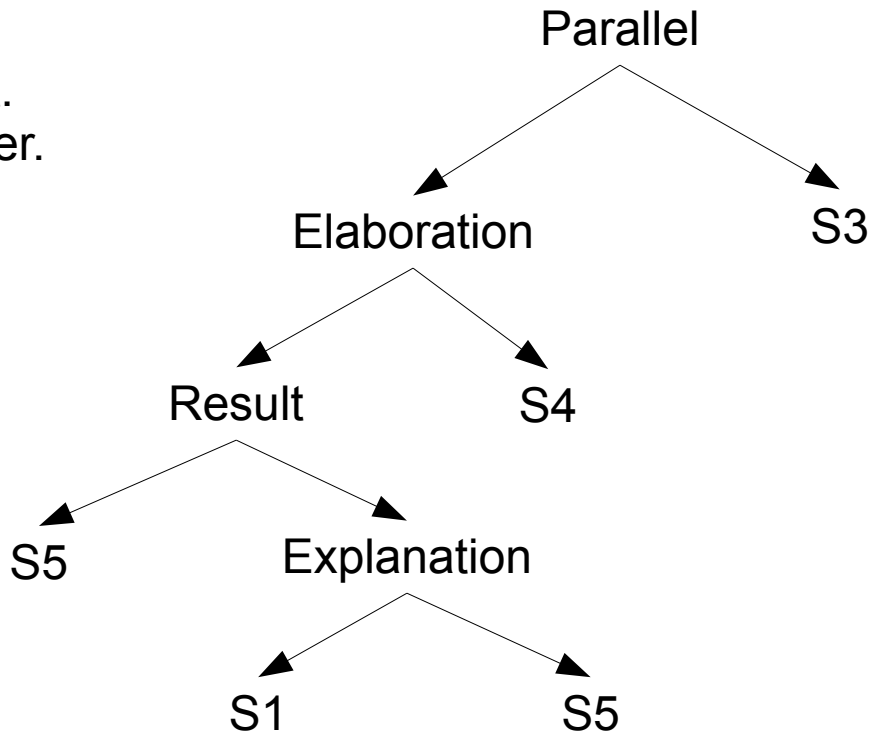
a. **John** went to his favorite music store to buy a piano.
b. **It** was a store John had frequented for many years.
c. **He** was excited that he could finally buy a piano. ✗
d. **It** was closing for the day just as John arrived.

Coherence Relations

- **Result**: one event can cause a following event
  - „John hid Bill's car keys. He was very upset about it.“
- **Explanation**: a previous event causing one event.
  - „John hid Bill's car keys. He was drunk.“
- **Parallel**: both events happening at the same time
  - „John hid Bill's car keys. Bill was sleeping.“
- **Elaboration**: Detailed elaboration of an event.
  - „John hid Bill's car keys. He put it in his bag.“
- **Occasion**: change of state:
  - „John hid Bill's car keys. He found it ten minutes later.“

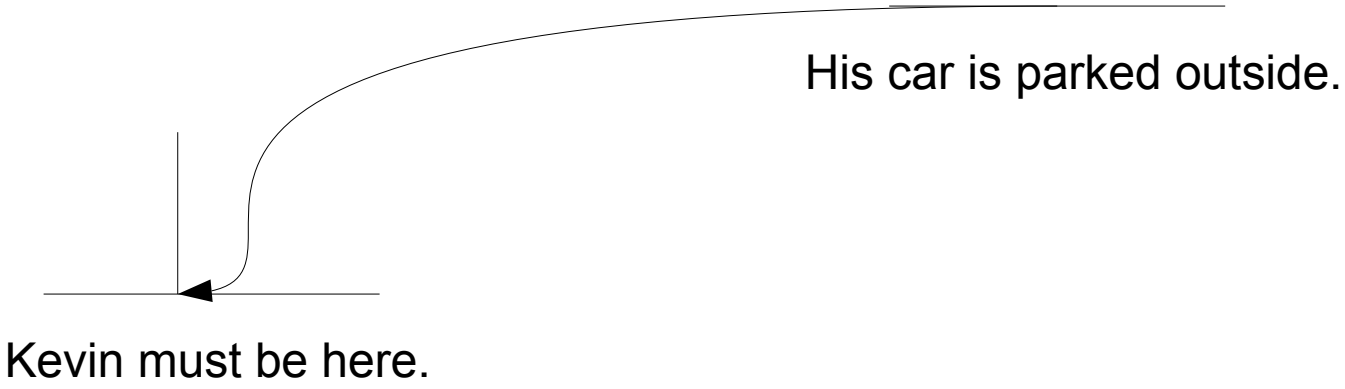
Discourse structure

(S1) Bill was drunk.
(S2) John hid Bill's car keys.
(S3) (While) Bill was sleeping.
(S4) He put it in his bag.
(S5) Bill was very upset about it.
(S6) Bill found it ten minutes later.

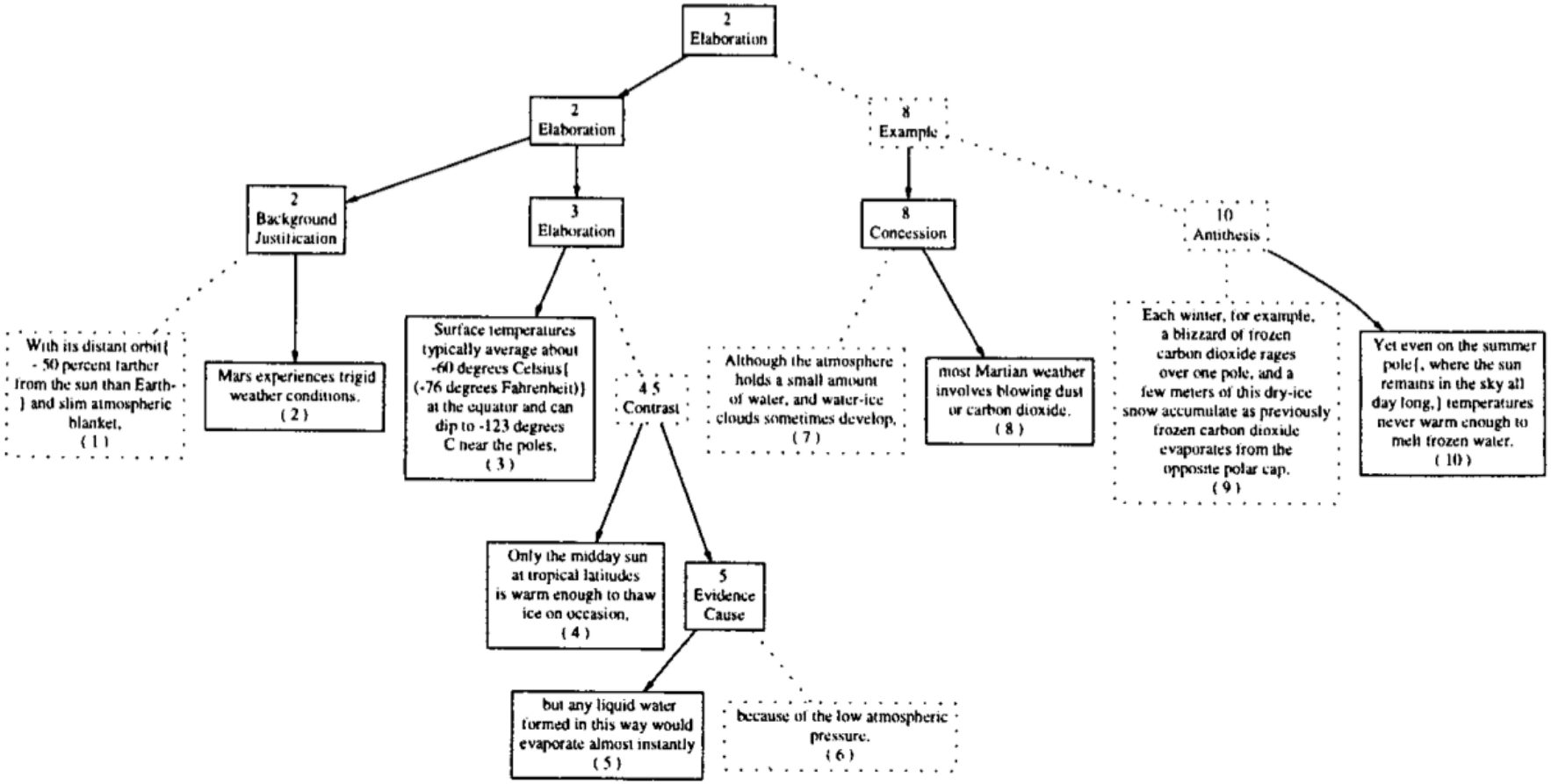


Rhetorical Structure Theory (RST)

- Model of text organization originally for text generation (Mann and Thompson 1987)
- Uses 23 rethorical relations hold between a **nucleus** and a **satellite** clauses or sentences



Rhetorical Structure Theory



(figure taken from Marcu 2000)

Automatic Coherence Assignment

- Also called discourse parsing
  - It is still a hard task
  - Open research question

Cue-phrase-based algorithm

1. Identify the cue phrases in text
    - Search for connectives, e.g., „because“, „but“, „for example“, „with“, „and“, etc.
    - „John hid Bill's car keys **because** he was drunk.“ (EXPLANATION)
2. Segment the text into discourse segments
3. Classify the relation between each consecutive segment

Cue-phrase-based algorithm

1. Identify the cue phrases in text
2. Segment the text into discourse segments
   - Segments can be clauses or sentences
     - „[John hid Bill's car keys] [**because** he was drunk].“
   - Segmentation can be carried out rule-based and can rely on the output of syntactic parsing
3. Classify the relation between each consecutive segment

Cue-phrase-based algorithm

1. Identify the cue phrases in text
2. Segment the text into discourse segments
3. Classify the relation between each consecutive segment
    - Usually also rule-based and according to the cue phrases
    - However, one cue phrase can be related to various RST relations
        - e.g., „because“ can be CAUSE or EVIDENCE

When no cue phrases are available

- Explore lexical semantics
  - „I don't want a truck; I'd prefer a convertible“ [CONTRAST]
    - negative vs. affirmative
    - truck vs. convertible

- Use of bootstrapping:
  - Use strong cue markers (e.g., „because“, „but“) to acquire text;
  - Remove the cue markers from the text;
  - Use the labeled data as training data

Outline

- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

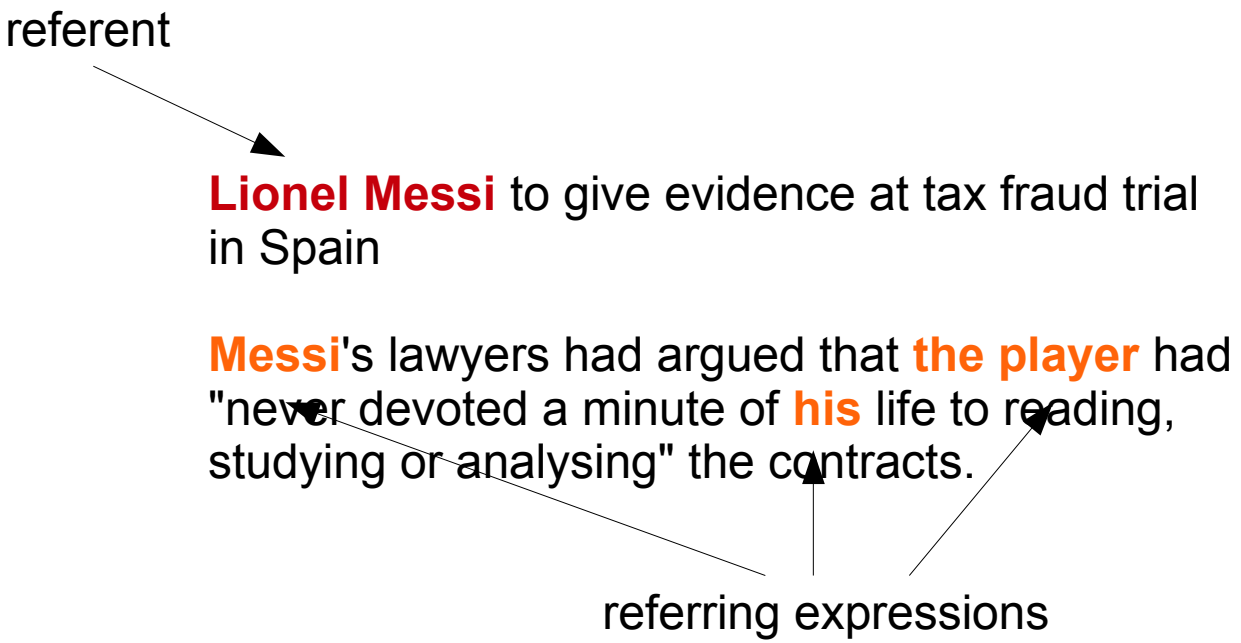
Reference resolution

Lionel Messi to give evidence at tax fraud trial in Spain
The Argentina and Barcelona footballer Lionel Messi is due to give evidence in a Spanish court on tax fraud charges.
Messi and **his** father Jorge, who manages **his** financial affairs, are accused of defrauding Spain of more than €4m (£3m; $4.5m) between 2007 and 2009.
The authorities allege that **the two** used tax havens in Belize and Uruguay to conceal earnings from image rights.
Spain's tax agency is demanding heavy fines and prison sentences. **Both men** deny any wrongdoing.
...
Messi's lawyers had argued that **the player** had "never devoted a minute of **his** life to reading, studying or analysing" the contracts.
But the high court in Barcelona ruled in June 2015 that **the football star** should not be granted immunity for not knowing what was happening with **his** finances, which were being managed in part by **his** father.
..
The footballer is the five-time World Player of the Year and one of the richest athletes in the world.

(http://www.bbc.com/news/world-europe-36433061)

Reference resolution

- The task of determining what entities are referred to by which linguistic expressions.
- Anaphora: reference to an entity (antecedent) previously introduced in the discourse



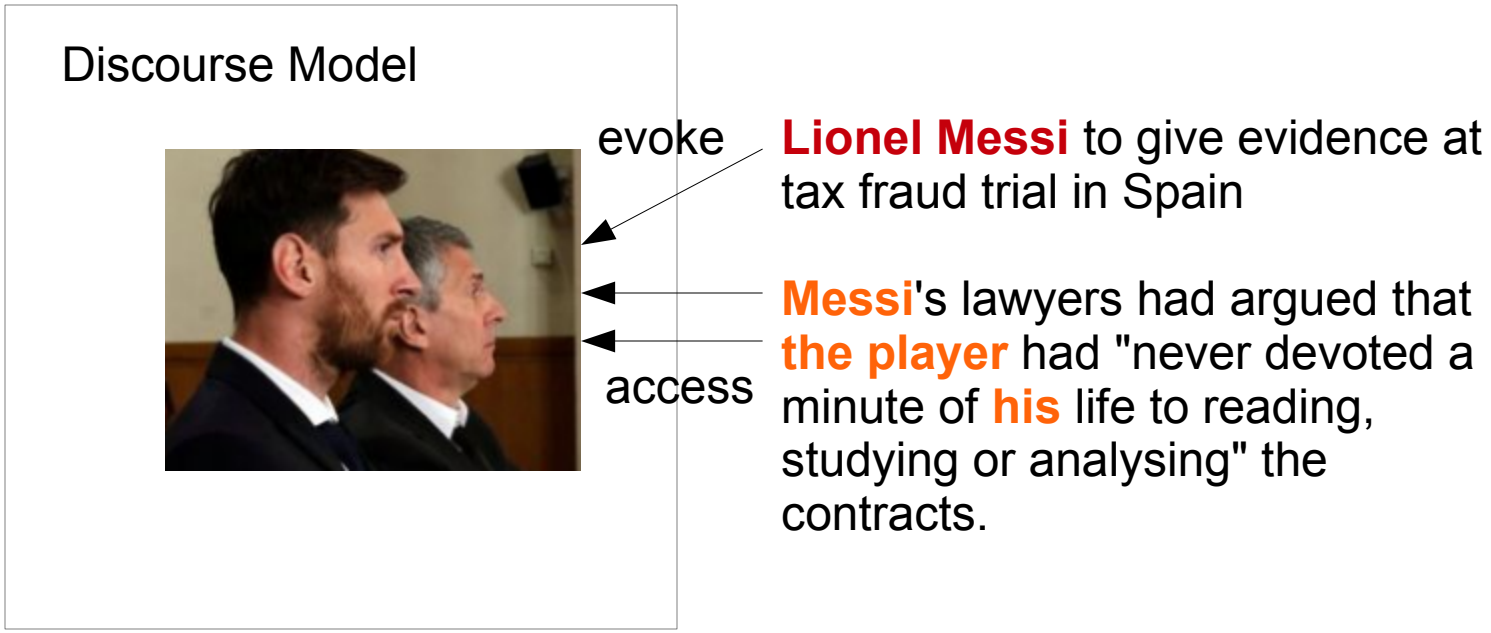
Reference resolution

- Discourse model:
  - Representation of the entities that have been referred to in the discourse and the relationships in which they participate.

- Two componnets are required
  - A discourse model should evolve with the dinamically changing discourse
  - A method for mapping between the signals that various referring expressions encode and the hearer's set of beliefs

Reference resolution

- evoke (first mention) vs. access (subsequent mentions)



Reference resolution

- Coreference resolution
  - Find referring expressions in the text that refer to the same entity, i.e., that **corefer**.
  - e.g., „Messi“, „the player“, „both men“, „his father“

- Pronomial anaphora resolution
  - Find the antecedent for a single pronoun
    - e.g., „he“, „they“
  - It is subtask of coreference resolution.

Types of referring expressions

- Indefinite noun phrases:
  - e.g., „a“, „an“, „this“, „some“
  - An entity new to the hearer
  - Evoke a new entity into the discourse model
  - Can be specific or non-specific

“The Argentina and Barcelona footballer Lionel Messi is due to give evidence in **a Spanish court** on tax fraud charges.”

“**A verdict** is not expected until next week.”

Types of referring expressions

- Definite noun phrases:
  - An entity identifiable to the hearer
  - Evoke a representation of the referrent into the discourse model

"The authorities allege that **the two** used tax havens in Belize and Uruguay to conceal earnings from image rights."

"Because of **the trial**, Messi has missed part of his national team's preparations for the Copa America, which starts on Friday in the US. Argentina's first game is on Monday June 6."

Types of referring expressions

- Pronouns (Pronominalization):
  - e.g., „he“, „she“, „they“
  - The entity was usually referred one or two sentence back.
  - **Cataphora**, mentioned before the referent.

“Because of the trial, **he** has missed part of his national team's preparations for the Copa America, which starts on Friday in the US. Argentina's first game is on Monday June 6.”

“**His** lawyers had argued that the player had "never devoted a minute of his life to reading, studying or analysing" the contracts.”

Types of referring expressions

- Demonstratives:
  - e.g., „this“, „that“
  - Can appear either alone or as determiners.

“The trial began on Tuesday, and Thursday is expected to be the final day. A verdict is not expected until the end of **this week**.”

“The authorities allege that the two used tax havens in Belize and Uruguay to conceal earnings from image rights. Both men deny **that**.”

Types of referring expressions

- Names:
  - Names of people, organizations, locations, etc.
  - Can be both new and old entities.

"**Messi** and his father Jorge, who manages his financial affairs, are accused of defrauding Spain of more than €4m (£3m; $4.5m) between 2007 and 2009."

"The Argentina and **Barcelona** footballer Lionel Messi is due to give evidence in a Spanish court on tax fraud charges."

Information Status (or Information Structure)

- The way that different referential forms are used to provide new or old information.

- Givenness hierarchy (Gundel et al. 1993):

in focus > activated > familiar > uniquely identifiable > referential > type identifiable

{it}	{that this this N}	{that N}	{the N}	{indef. this N}	{a N}

- Accessibility scale (Ariel 2001):

full name > long definitive description > short definite description > last name > first name > distal demonstrative > NP > stressed pronoun > unstressed pronoun

Information Status

- Complicating factors for relations between referring expressions and information status:
  - **Inferrable**: not explicitly evoked in text, but related to an evoked entity
    - „Because of the trial, Messi has missed part of his national team's preparations for the Copa America, which starts on Friday in the US. Argentina's **first game** is on Monday.“

Information Status

- Complicating factors for relations between referring expressions and information status:
  - **Generics**: not explicitly evoked in text, but a generic reference
    - „I only worried about playing football," he told the judge. But **they** did not believe a word of it.“

Information Status

- Complicating factors for relations between referring expressions and information status:
  - **Non-referential uses**:
    - „**It** was the judge who asked Messi the questions.“

Features Pronomial Anaphora Resolution

- Number agreement:
  - „I only worried about playing football," Messi told the judge. But **she** did not believe a word of it.“

    instead of
  - „I only worried about playing football," he told the judge. But **they** did not believe a word of it.“

    but semantically plural entities can use *it* or *they*:
  - „Argentina's first game is on Monday. **They** cannot count with Messi for the first game.“

Features Pronomial Anaphora Resolution

- Person agreement:
  - „I only worried about playing football," Messi told the judge. But **he** should not be granted immunity for not knowing what was happening with his finances.“
- Gender agreement:
  - „Messi's lawyers had argued that the player had never devoted a minute of his life to reading, studying or analysing the contracts. **He** should be more careful about **them**.“
- Binding Theory Constraints:
  - „Messi said that the father **himself** manages his finances.“

Preferences in Pronoun Interpretation

- Recency:
  - „Messi played in the match against Brazil last week. Argentina will play against Chile this week. They hope to win **it**."
- Grammatical rules: subject position more salient than object position
  - „Angentina played a match against Brazil last week. **They** won."
- Repeated mention: focused entities are likely to continue to be focused
  - „Angentina played a match against Brazil last week. They also played matches against Chile and Peru. But Uruguay didn't want to play against **them**."

Preferences in Pronoun Interpretation

- Parallelism: preferences can be induced by parallelism effects.
  - „Argentina's first match was against Brazil in the last America Cup. Chile will also play against **them** this year.“
- Verb semantics: interpretation can be biased by semantically oriented emphasys.
  - „Argentina beat Brazil. They won.“
  - „Argentina lost to Brazil. They won.“
- Selectional restrictions: semantic role can play a role in referent preferences.
  - „Messi said he signed the documents related to his finances without reading **them**.“

Algorithms for Anaphora resolution

- Hobbs
- Log-linear

- Centering (check the book)

Hobbs algorithm

- Relies on
  - Syntactic parser
  - Morphological gender and number checker
- Input: the pronoun to be resolved
- Output: an noun phrase

Hobbs algorithm (Breadth-first, left-to-right search)

I only worried about playing football, Messi told the judge.

```
(ROOT
  (S
    (S
      (NP (PRP I))
      (ADVP (RB only))
      (VP (VBN worried)
        (PP (IN about)
          (S
            (VP (VBG playing)
              (NP (NN football)))))))
    (, ,)
    (NP (NNP Messi))
    (VP (VBD told)
      (NP (DT the) (NN judge)))
    (. .)))
```

But **she** did not believe a word of it.

```
(ROOT
  (S (CC But)
    (NP (PRP she))
    (VP (VBD did) (RB not)
      (VP (VB believe)
        (NP
          (NP (DT a) (NN word))
          (PP (IN of)
            (NP (PRP it))))))
    (. .)))
```

Hobbs algorithm (Breadth-first, left-to-right search)

I only worried about playing football, Messi told the judge.

```
(ROOT
 (S
  (S
   (NP (PRP I))
   (ADVP (RB only))
   (VP (VBN worried)
    (PP (IN about)
     (S
      (VP (VBG playing)
       (NP (NN football)))))))
  (, ,)
  (NP (NNP Messi))
  (VP (VBD told)
   (NP (DT the) (NN judge)))
  (. .)))
```

No NPs found

But **she** did not believe a word of it.

```
(ROOT
 (S (CC But)
  (NP (PRP she))
  (VP (VBD did) (RB not)
   (VP (VB believe)
    (NP
     (NP (DT a) (NN word))
     (PP (IN of)
      (NP (PRP it))))))
  (. .)))
```

Hobbs algorithm (Breadth-first, left-to-right search)

Check all NPs

I only worried about playing football, Messi told the judge.

```
(ROOT
 (S
  (S
   (NP (PRP I))
   (ADVP (RB only))
   (VP (VBN worried)
    (PP (IN about)
     (S
      (VP (VBG playing)
       (NP (NN football)))))))
  (, ,)
  (NP (NNP Messi))
  (VP (VBD told)
   (NP (DT the) (NN judge)))
  (. .)))
```

But **she** did not believe a word of it.

```
(ROOT
 (S (CC But)
  (NP (PRP she))
  (VP (VBD did) (RB not)
   (VP (VB believe)
    (NP
     (NP (DT a) (NN word))
     (PP (IN of)
      (NP (PRP it))))))
  (. .)))
```

Hobbs algorithm

- Advantages:
  - The search order gives considers binding theory, recency and grammatical role
  - A final check accounts for gender, person and number constraints
- Disadvantages:
  - Do not have a explicit discourse model

Log-Linear Algorithm

- Supervised machine learning approach
- Log-linear, but also any other ML algorithm
- Must rely on
  - annotated data with positive and negative examples
  - full parser or chunker
- Usually, pleonasms are removed, e.g., „It is raining.“

- Input: pair of NP and pronoun
- Output: 0 or 1 (binary classification)

Features for ML

- Strict number (true/false)
- Strict gender (true/false)
- Regarding pronoun and NP
  - Compatible number (true/false)
  - Compatible gender (true/false)
  - Sentence distance [0,1,2,3,...]: # of sentences
  - Hobbs distance [0,1,2,3,...]: # of skipped NPs
  - Grammatical role [subject,object,PP] of the potential antecedent
  - Linguistic form [proper, definite, indefinite, pronoun] of the potential antecedent

Coreference Resolution

- Deal with definite noun phrases and names
  - „Messi“, „the player“, „the football star“, „Barcelona, „Argentina“
- We can use a similar log-linear/ML classifier

Coreference Resolution

- We can rely on the same features for anaphora, plus:
  - Anaphor edit distance: minimum edit distance from potential antecedent to anaphor
  - Antecedent edit distance: minimum edit distance from anaphor to antecedent
  - Alias (true/false) based on named-entity recognition
  - Appositive (true/false): e.g., „Lionel Messi, the Barcelona football star, ….“
  - Linguistic form [proper, definite, indefinite, pronoun]

Outline

- Introduction
- Text Segmentation
- Text Coherence
- Reference Resolution
- Evaluation

Evaluation of Coreference Resolution

- B-CUBED
  - Reference chain (or true chain)
  - Hypothesis chain
  - Computation of precision and recall of entities in the hypothesis against the reference chains

$$Precision = \sum_{i=1}^{N} w_i \frac{\#\ of\ correct\ elements\ in\ hypothesis\ chain\ containing\ entity}{\#\ of\ elements\ in\ hypothesis\ chain\ containing\ entity}$$

$$Recall = \sum_{i=1}^{N} w_i \frac{\#\ of\ correct\ elements\ in\ hypothesis\ chain\ containing\ entity}{\#\ of\ elements\ in\ reference\ chain\ containing\ entity}$$

Further Reading

- Book „Speech and Language Processing“
  - Chapter 21