Diskurs

Katja Markert,Einige Folien von Bonnie Webber

January 22, 2020

Bis jetzt und heute

Bis jetzt:

- Wörter: Tokenisierung, Bedeutung
- N-grams
- Sätze: syntaktisches und semantisches Parsing
- Was ist mit größeren Texteinheiten?

Nun: **Was macht einen Text kohärent?**

1. Was ist Diskurs?
2. Referentielle Struktur: Basisdefinitionen
3. Referentielle Struktur: Anwendungen
4. Informationsstatus: Genauere Untersuchung von Form und Funktion (OPTIONAL)
5. Pronomenresolution: Restriktionen und Präferenzen

Übersicht

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

Katja fall. Katja aua.

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

Katja versteckte Michaels Autoschlüssel. Er war betrunken.

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

Katja versteckte Michaels Autoschlüssel. Er mag Spinat.

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

A: Das Telefon klingelt.
B: Ich bin unter der Dusche.

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

Diskurs und Kohärenz

Diskurs: kohärente Folge von Sätzen, von einem Sprecher/Schreiber in einer bestimmten Situation produziert

Kohärenz: Was macht einen Diskurs kohärent?

- Kohärenz $\neq$ Grammatikalität einzelner Sätze
- Sätze verbunden durch Diskursrelationen
- gut verständlich in bestimmtem **Kontext**
- kohärenter Text inkludiert oft Kohäsionselemente, aber diese sind weder notwendig noch ausreichend

Kontext

Kontext (= bisheriger Text, Situation, sowie Weltwissen) erlaubt:

- Inferenz von Diskursrelationen zwischen Sätzen (Sequenzen, Gründe, Kontraste)
- Die Auflösung von **diskursgebundenen/discourse-bound** Elementen. Diese sind sonst implizit oder unterspezifiziert.

  Three rabbits came into my garden. They ate the carrots.

Kohäsion

Kohäsionselemente:

- Teile des Sprachsystems, die oft Kohärenz stärken und Lesbarkeit erhöhen
- innerhalb oder zwischen Sätzen
- **Beispiele:** Wortwiederholungen, Benutzung relationierter Worte, Diskurskonnektiva, diskursgebundene Elemente ...
- Kohäsion $\neq$ Kohärenz, aber Kohäsion ist in kohärenten Texten meistens da

Übersicht

Referenz

Referenz

Benutzung von linguistischen Ausdrücken, um sich auf eine Entität zu beziehen. Referentielle Ausdrücke beziehen sich auf Referenten.

- Es gibt eine große Anzahl von **referentiellen Ausdrücken**, die man benutzen kann, um sich auf **Referenten** z u beziehen. *she, John, a man, the man, two rabbits, some cars . . .*
- Auch sind diese Ausdrücke nicht in jedem Kontext referentiell
  - Pleonastisches "Es": *Es regnet*
  - Prädikativer Gebrauch: *Er ist ein Idiot* (nicht referentiell, $\neq$ *Er traf einen Idi oten*)

Koreferenz und Anaphern

Koreferenz

zwei oder mehr referentielle Ausdrücke referenzieren die gleiche Entität

John left the house. He went to the supermarket.

Anaphern

Linguistischer Ausdruck, dessen Referenten man nur vollständig durch die Anwesendheiten eines anderen, normalerweise vorher benutzten Ausdrucks verstehen kann. Der vorherige Ausdruck heisst Antezedent.

John left the house. He went to the supermarket.

Unterschied zwischen Anapher und Koreferenz

Anaphorisch aber nicht koreferent

- Komparative Anaphern: *John left the class. The other students stayed behind.*
- Bridging: *John left the house. He forgot to lock the door.*
- One-anaphora: *John bought a brown shirt. Mark bought a red one.*

Koreferent aber nicht anaphorisch

- *John left the house and went to school. After school, John went home.*

Fragen für die automatische Behandlung

- Wie erkenne ich Anaphern?
- Wie finde ich den Antezedenten?
- Wie baue ich Koreferenzketten?

Übersicht

Textzusammenfassung

Extraktive Zusammenfassungssysteme hinterlassen meist ein Chaos in der referentiellen Struktur:

[i] More than 130 bodies are reported to have been recovered after a Gulf Air jet carrying 143 people crashed into the Gulf off Bahrain on Wednesday. [ii] Distraught relatives also gathered at Cairo airport, demanding . [iii] He also declared three days of national mourning. [iv] He said the jet fell "sharply, like an arrow."
[*Otterbacher et al, 2002*]

Maschinelle Übersetzung?

Übersetzung von Anaphern kann evtl. SMT verbessern: [Le Nagard & Koehn 2010; Hardmeier et al, EMNLP 2013].

Ich kaufte eine Kaffeemaschine. Sie ist schon kaputt

Generierung

Szenario: Man will auf eine vorhergehende Entität zurückreferieren. Benutze ich ein Pronomen, wiederhole den Eigennamen oder nehme ich eine definite NP?

- Wahl beeinflusst Korrektheit

  Katja kaufte Katja ein neues Auto.

- Wahl beeinflusst Natürlichkeit

  1. *Katja ging in den Laden. Dort kaufte Katja ein neues Auto.*
  2. *Katja ging in den Laden. Dort kaufte sie ein neues Auto.*
  3. *Katja liebt Luise. Sie/Katja kauft ihr ein Geschenk.*

Gordon et al (1993): **repeated name penalty** in Lesezeiten

Zwischenfazit

- Referentielle Struktur ist ein Aspekt von Kohäsion
- (Referentielle) Kohäsion macht einen Text alleine aber nicht unbedingt kohärent
- Es gibt eine Korrelation zwischen Anaphorizität und syntaktischer Form
- Am stärksten ist diese wahrscheinlich für Pronomen: meist anaphorisch
- Referentielle Struktur wichtig für Anwendungen, die größere Textabschnitte betrachten (summarization, generation)

Übersicht

Informationsstatus

Informationsstatus/IS

Das Ausmaß, in dem ein Referent einem Hörer/Leser zugänglich oder schon bekannt ist. Accessibility/Familiarity.

Hauptidee: Informationsstatus korreliert mit syntaktischer Form des referentiellen Ausdrucks

Linguistische Form und IS (Prince 1981,1992)

IS	Disourse-new	Discourse-old
Hearer-new	brand-new	—-
Hearer-old	unused	evoked

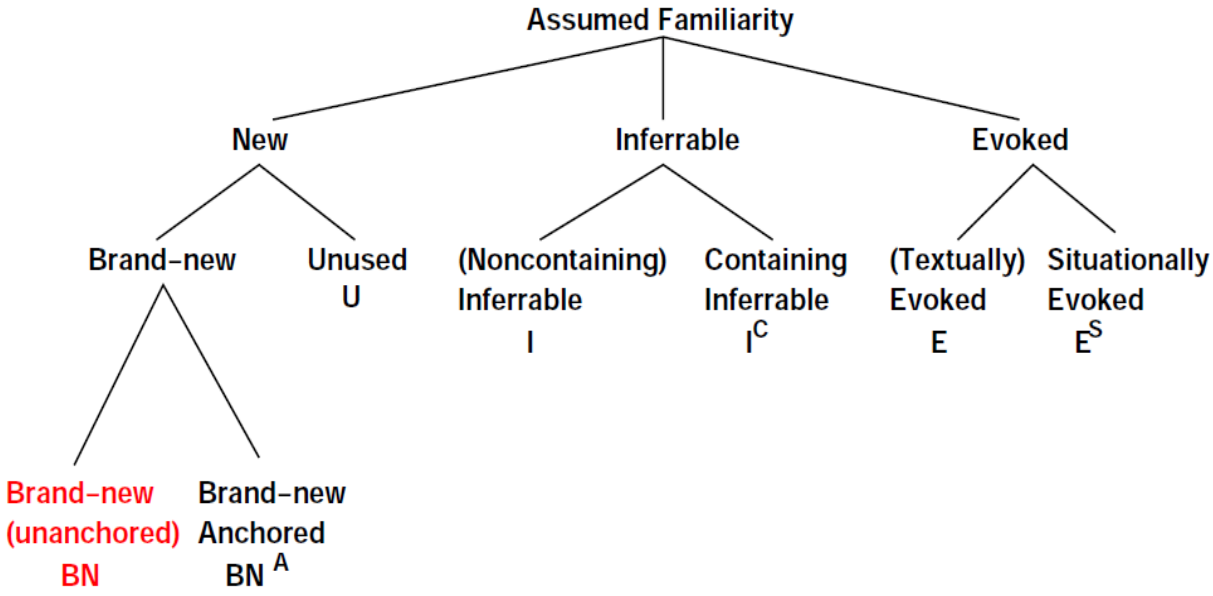
Brand-new: Neu im DM, unbekannte Entität (a man)

Unused: Neu im DM, bekannte Entität (Michael Jackson)

Evoked/old: Schon im DM durch Diskurs (The student) oder Umgebung (You).

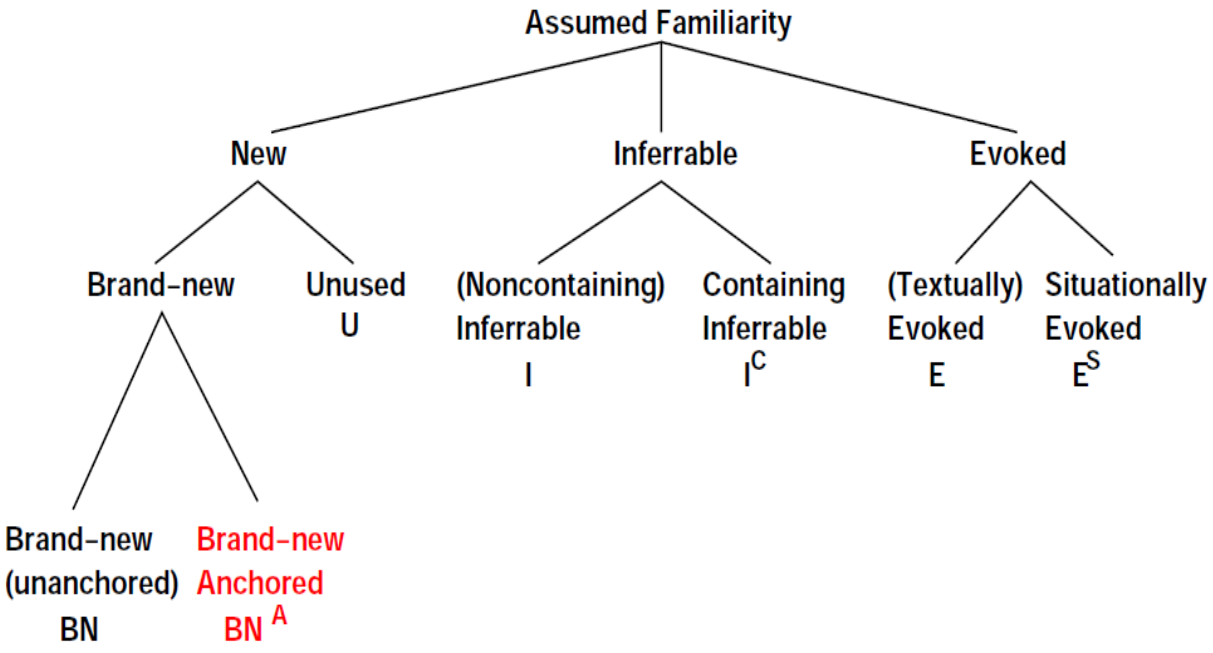
Inferrable: Neu im DM, aber mit schon existierender Entität verbunden (weder discourse-new noch discourse-old). *John left the house. He left the door open.*

Prince Familiarity hierarchy



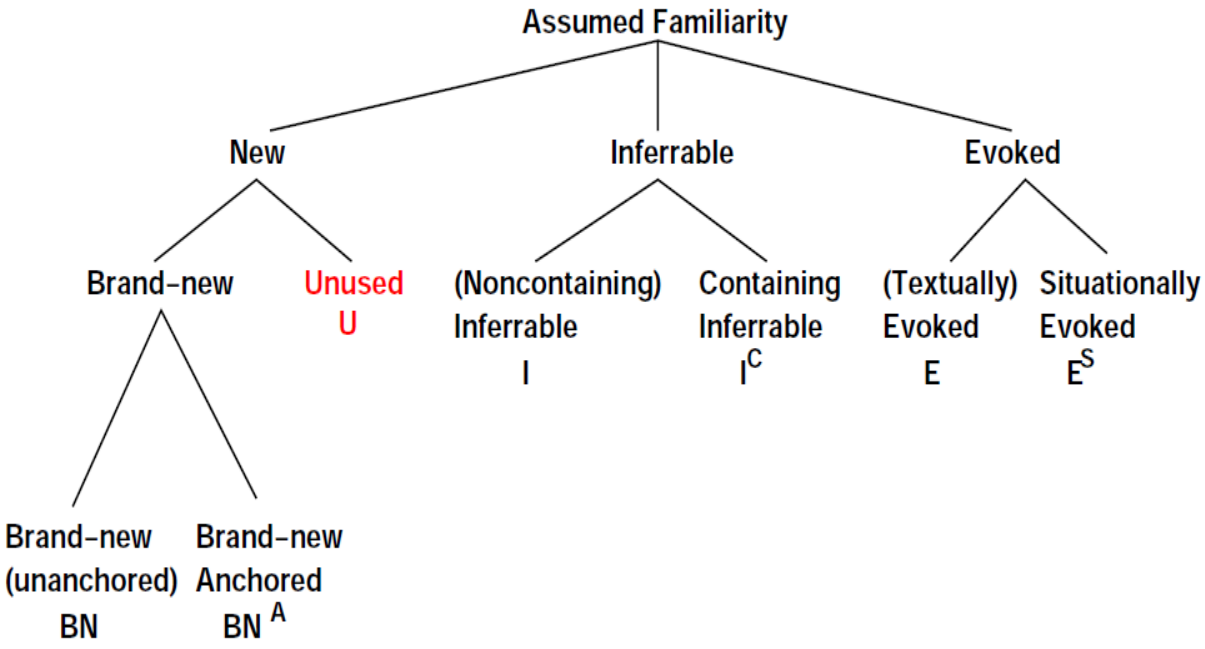
Yesterday I got onto a bus. The driver was drunk.

Prince Familiarity hierarchy



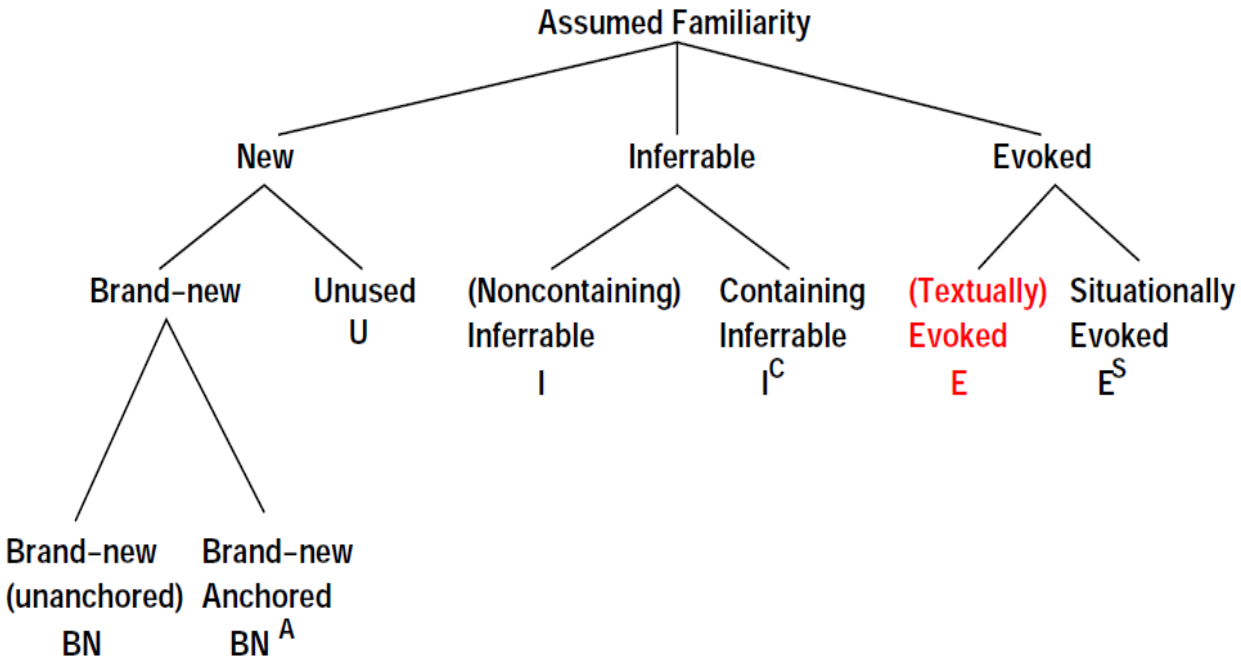
Yesterday I got onto a bus that went to Berlin. The driver was drunk.

Prince Familiarity hierarchy



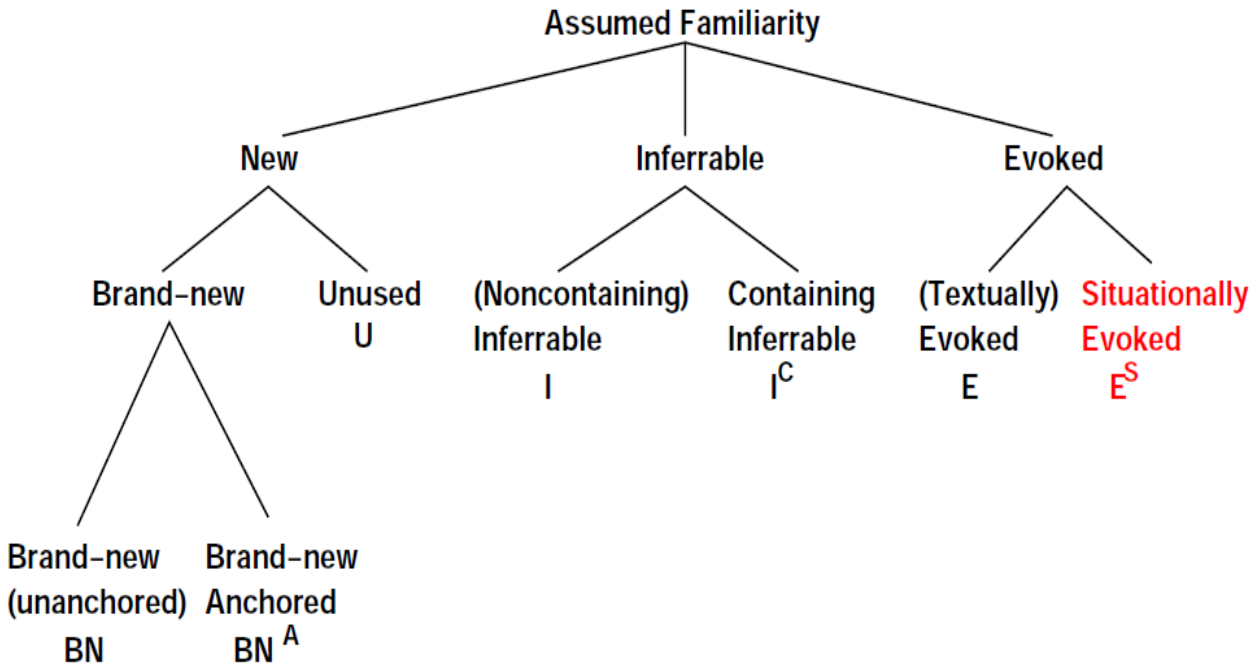
Yesterday I went to Berlin by bus. The driver was drunk.

Prince Familiarity hierarchy



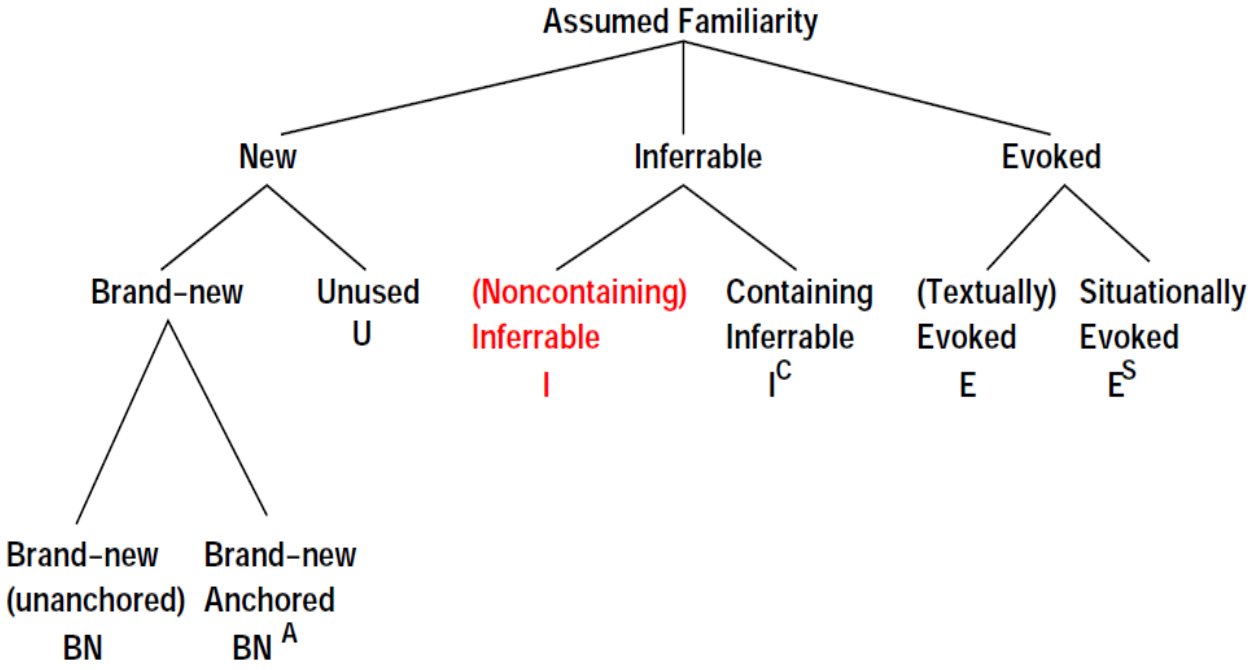
Yesterday I got onto a bus. It was very dirty.

Prince Familiarity hierarchy



Yesterday I got onto a bus. The driver was drunk.

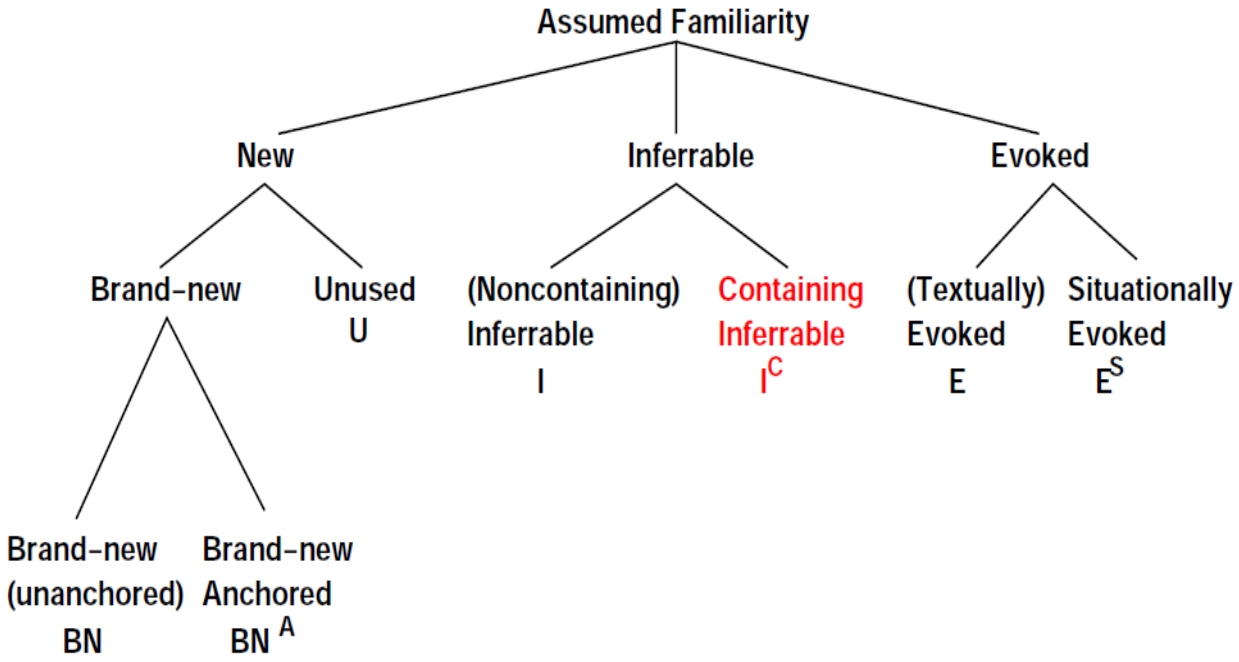
Prince Familiarity hierarchy



Yesterday I got onto a bus. The driver was drunk.

Auch Bridging Anaphern genannt.

Prince Familiarity hierarchy



Yesterday I got onto a bus. Its driver was drunk.

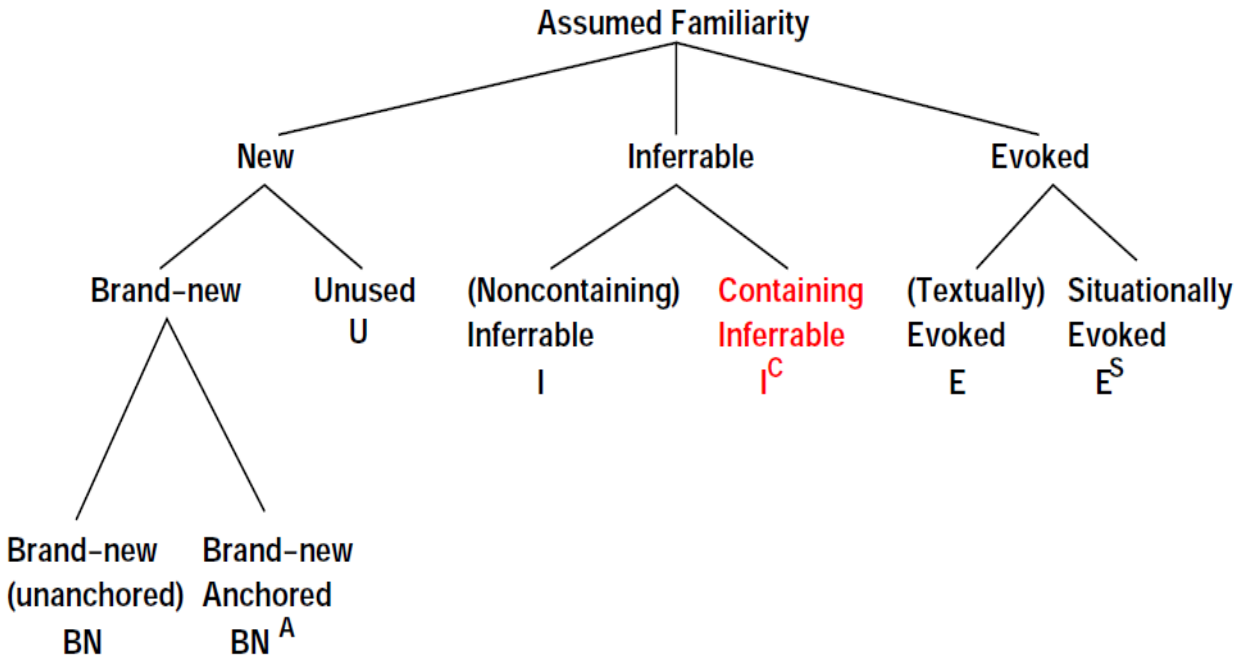
Eine Übung

Beispiel

The Bakersfield Supermarket went bankrupt last May. The business located in northern Manhattan closed when its owner was robbed and murdered. Friends expressed outrage at the murder. Unfortunately, such crimes are not unusual.

Markieren Sie alle referentiellen Ausdrücke und weisen Sie die IS Kategorie zu.

Prince Familiarity Hierarchy



$evoked > Unused > inferrable > Containing - inferrable > brandnew - anchored > brandnew - unanchored$

Prince Familiarity Hierarchy

- Wenn wir einen in der Hierarchie schwächeren Ausdruck benutzen, dann nur weil die stärkere nicht möglich war
- *I saw John with a woman yesterday* nur wenn *a woman* nicht seine Frau ist.
- Korrelation zwischen IS und syntaktischer Form
  - Pronomen meist hearer-old (oft discourse-old)
  - Definite NPs selten brand-new
  - Indefinite NPs oft brand-new
  - Es gibt aber auch immer noch viel Variation!

Zwischenfazit

- Referentielle Struktur ist ein Aspekt von Kohäsion
- IS modelliert die familiarity/accessibility von Diskursentitäten
- Es gibt eine Korrelation zwischen IS und syntaktischer Form
- Am stärksten ist diese wahrscheinlich für Pronomen:
  hearer-old, bei geschriebenen Texten meist auch discourse-old

Übersicht

Pronomen

Eines der ältesten und meistbeachteten NLP Probleme. (Warum?)

- Anaphorische Pronomen meist koreferent mit Antezedent
- Antezedent zumeist im gleichen Satz oder in relativ nahestehendem vorherigem Satz
  1. *Katja sagte sie mag Karotten.*
  2. *Katja ging in den Garten. Sie sitzt gern in der Sonne.*

Pronomen

Eines der ältesten und meistbeachteten NLP Probleme. (Warum?)

- Anaphorische Pronomen meist koreferent mit Antezedent
- Antezedent zumeist im gleichen Satz oder in relativ nahestehendem vorherigem Satz
  1. *Katja sagte sie mag Karotten.*
  2. *Katja ging in den Garten. Sie sitzt gern in der Sonne.*
- Oft mehrere mögliche Antezedenten
  1. *Maria schlug eine Frau vor dem Supermarkt. Sie war wütend.*

Pronomen

Eines der ältesten und meistbeachteten NLP Probleme. (Warum?)

- Anaphorische Pronomen meist koreferent mit Antezedent
- Antezedent zumeist im gleichen Satz oder in relativ nahestehendem vorherigem Satz
  1. *Katja sagte sie mag Karotten.*
  2. *Katja ging in den Garten. Sie sitzt gern in der Sonne.*
- Oft mehrere mögliche Antezedenten
  1. *Maria schlug eine Frau vor dem Supermarkt. Sie war wütend.*

Wie wählen wir den besten Antezedenten aus?

Restriktionen: Agreement

Numerus:

**John bought a new car. They are red.*

Person:

- *You and I own cars. We are rich.*
- * *You and I own cars. They are rich.*

Gender

- *Mary hit Bill. She hates Bill.*
- * *Mary hit Bill. He hates Bill.*

Aber...

Sprachabhängig:

** Maria hat eine Katze. Es frisst Mäuse*

Maria has a cat. It . . .

Diskontinuierliche Mengen:

John has a car and Mary has a car. They drive them all the time.

Implizite Mengen

John kauft jedes Jahr ein neues Auto. Sie sind immer rot.

Inferrables?

I brought the car to the garage. They said they will fix it in one day.

Syntaktische Restriktionen

Für satzinterne Pronominalanaphern

Reflexiva: himself, herself, sich

Syntaktische Restriktionen

Für satzinterne Pronominalanaphern

Reflexiva: himself, herself, sich

- *John bought himself a new car.* $\longrightarrow$ [himself=John]

Syntaktische Restriktionen

Für satzinterne Pronominalanaphern

Reflexiva: himself, herself, sich

- *John bought himself a new car.* $\longrightarrow$ [himself=John]
- *John bought him a new car.* $\longrightarrow$ [him $\neq$ John]

Syntaktische Restriktionen

Für satzinterne Pronominalanaphern

Reflexiva: himself, herself, sich

- *John bought himself a new car.* $\longrightarrow$ [himself=John]
- *John bought him a new car.* $\longrightarrow$ [him $\neq$ John]
- *He said that he bought John a new car.* $\longrightarrow$ [He $\neq$ John, he $\neq$ John]

Syntaktische Restriktionen

Für satzinterne Pronominalanaphern

Reflexiva: himself, herself, sich

- *John bought himself a new car.* $\longrightarrow$ [himself=John]
- *John bought him a new car.* $\longrightarrow$ [him $\neq$ John]
- *He said that he bought John a new car.* $\longrightarrow$ [He $\neq$ John, he $\neq$ John]

Binding theory: Syntaktische Restriktionen für Koreferenz

Präferenzen

Selbst nach allen Restriktionen bleiben meist mehr als ein möglicher Antezedent übrig (**referentielle Ambiguität**)

1. *Maria schlug eine Frau vor dem Supermarkt. Sie war wütend.*
2. *A: Welches Gebäude beobachtest Du? B: Ich beobachte das Gebäude mit dem beschädigten Dach vor dem Hochhaus. Es ist rot.*

Präferenzen

Selbst nach allen Restriktionen bleiben meist mehr als ein möglicher Antezedent übrig (**referentielle Ambiguität**)

1. *Maria schlug eine Frau vor dem Supermarkt. Sie war wütend.*
2. *A: Welches Gebäude beobachtest Du? B: Ich beobachte das Gebäude mit dem beschädigten Dach vor dem Hochhaus. Es ist rot.*

Salienz: Antezedenten für Pronomina brauchen einen hohen Grad and Aktivierung im DM

Welche Faktoren beeinflussen Salienz? Wir schauen uns Minimalpaare an sowie inkohärente Beispiele

Übung

Was ist der präferierte Antezedent und warum?

1. *John met Sally. He likes her.*
2. *John met Sally. Then he met Mary. He likes her.*

Übung

Was ist der präferierte Antezedent und warum?

1. *John met Sally. He likes her.*
2. *John met Sally. Then he met Mary. He likes her.*

Recency preference: Entitäten aus näheren Sätzen haben höhere Salienz als Entitäten aus weiter wegstehenden (siehe auch Clark und Sengal 1979)

Theorie: priviligierter Platz im Arbeitsgedächtnis

Präferenzen

Was ist der präferierte Antezendent und warum?

1. *John went to the car dealership with Bill. He bought a sports car.*
2. *Bill went to the car dealership with John. He bought a sports car.*

Präferenzen

Was ist der präferierte Antezendent und warum?

1. *John went to the car dealership with Bill. He bought a sports car.*
2. *Bill went to the car dealership with John. He bought a sports car.*

Syntaktische Rolle beeinflusst Salienz

subjekt > direktes Objekt > andere

Präferenzen

John needed a car top get to his new job. He decided that he wanted sth sporty. Bill went to the car dealership with him. He bought a sportscar.

Präferenzen

John needed a car top get to his new job. He decided that he wanted sth sporty. Bill went to the car dealership with him. He bought a sportscar.

Repeated Mention: Eine Entität, die schon oft wiederholt wurde, bleibt weiterhin salient.

Präferenzen

- *Mary likes Sue's brother, and Harry her sister.*

Präferenzen

- *Mary likes Sue's brother, and Harry her sister.*
- *Melania Trump admires Hillary Clinton, and Donald Trump adores her.*

Präferenzen

- *Mary likes Sue's brother, and Harry her sister.*
- *Melania Trump admires Hillary Clinton, and Donald Trump adores her.*

Parallelismus: Wenn man Sätze parallel interpretieren kann, indem man ein Pronomen auf eine bestimmte Art und Weise auflöst, dann gibt es eine starke Tendenz dies zu tun.

Weltwissen und semantische Präferenzen

All dies ist oft immer noch nicht genug: Man braucht oft großes semantisches oder Weltwissen

Weltwissen und semantische Präferenzen

All dies ist oft immer noch nicht genug: Man braucht oft großes semantisches oder Weltwissen

- *John parked his car in the garage. It was messy, with coffee cups and maps everywhere.*

Weltwissen und semantische Präferenzen

All dies ist oft immer noch nicht genug: Man braucht oft großes semantisches oder Weltwissen

- *John parked his car in the garage. It was messy, with coffee cups and maps everywhere.*
- *John parked his car in the garage. It was messy, with old car parts and bikes everywhere.*

Fazit

- Diskurse sind nicht eine willkürliche Ansammlung von Sätzen. Sie müssen kohärent sein.
- Referentielle Struktur ist eine From der Kohäsion.
- Viele verschiedene Arten von referentiellen Ausdrücken mit gewisser Korrelation zu Familiarität der Entitäten
- Pronomen: meist anaphorisch sowie koreferent zu vorherigem Ausdruck (aber nicht immer: pleonastisches “es”, Sonderfälle).
- Es gibt viele Restriktionen und Präferenzen für Pronomenresolution.
- Diese interagieren $\longrightarrow$ schweres Problem

Nächstes Mal: Algorithmen zur Pronomen- sowie Koreferenzresolution

Referenzen

- **J&M (Edition2, book) : Kapitel 18.1 und 18.2
- **J&M (Edition3, online) : Kapitel 22.1
- Gordon et al. (1993): Pronouns, names and the centering of attention in discourse. In *Cognitive Science*, 17(3), p. 311 -347
- Clark und Sengal (1979): In search of referents for nouns and pronouns. In *Memory and Cognition*, 7, pp.35-41
- Otterbacher et al (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4. Association for Computational Linguistics, 2002.*
- Hardmeier et al (2013) Latent anaphora resolution for cross-lingual pronoun prediction. *EMNLP 2013; Conference on Empirical Methods in Natural Language Processing*

Referenzen

- Le Nagard und Koehn (2010). Aiding pronoun translation with co-reference resolution. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Association for Computational Linguistics*
- Prince, Ellen F. (1981). Towards a taxonomy of given-new information. *Radical pragmatics*.
- Prince, Ellen (1992). The ZPG letter: Subjects, definiteness, and information-status. *Discourse description: diverse analyses of a fund raising text*, 295-325.