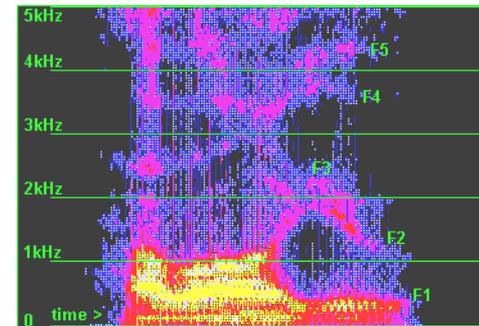
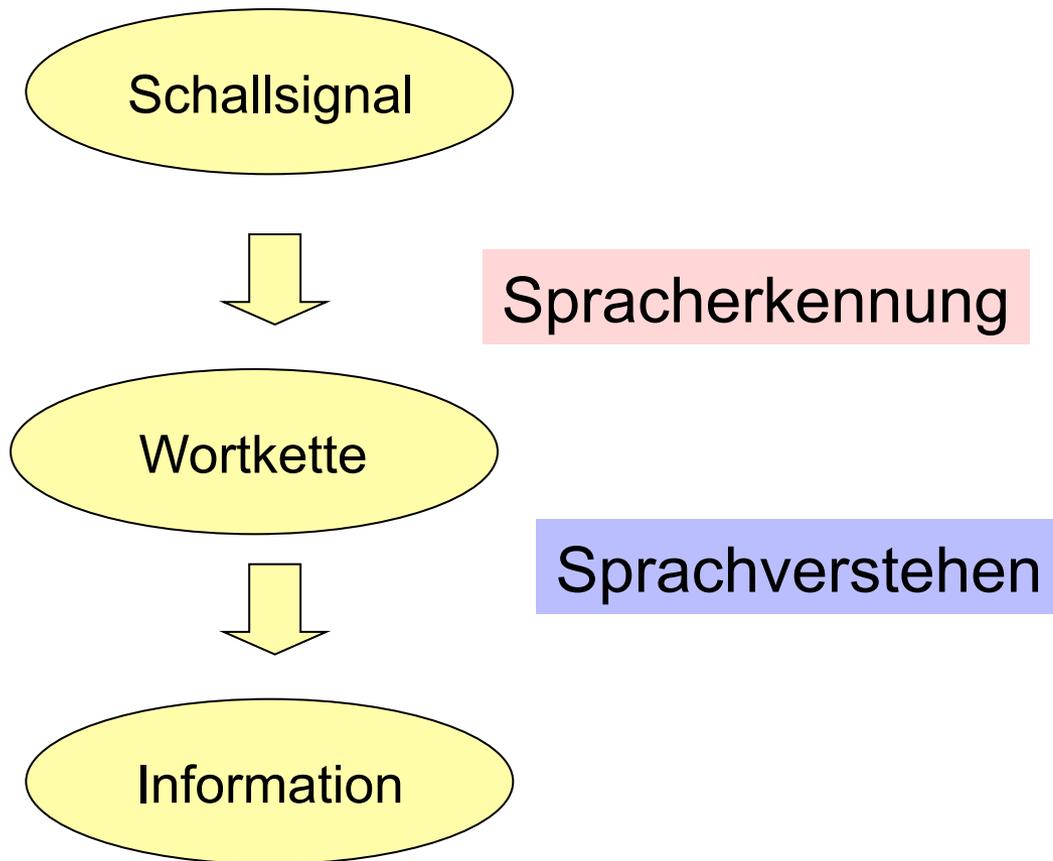


Einführung in die Computerlinguistik

Semantik

WS 2014/15

Vera Demberg



A yellow arrow pointing downwards from the spectrogram to the text below.

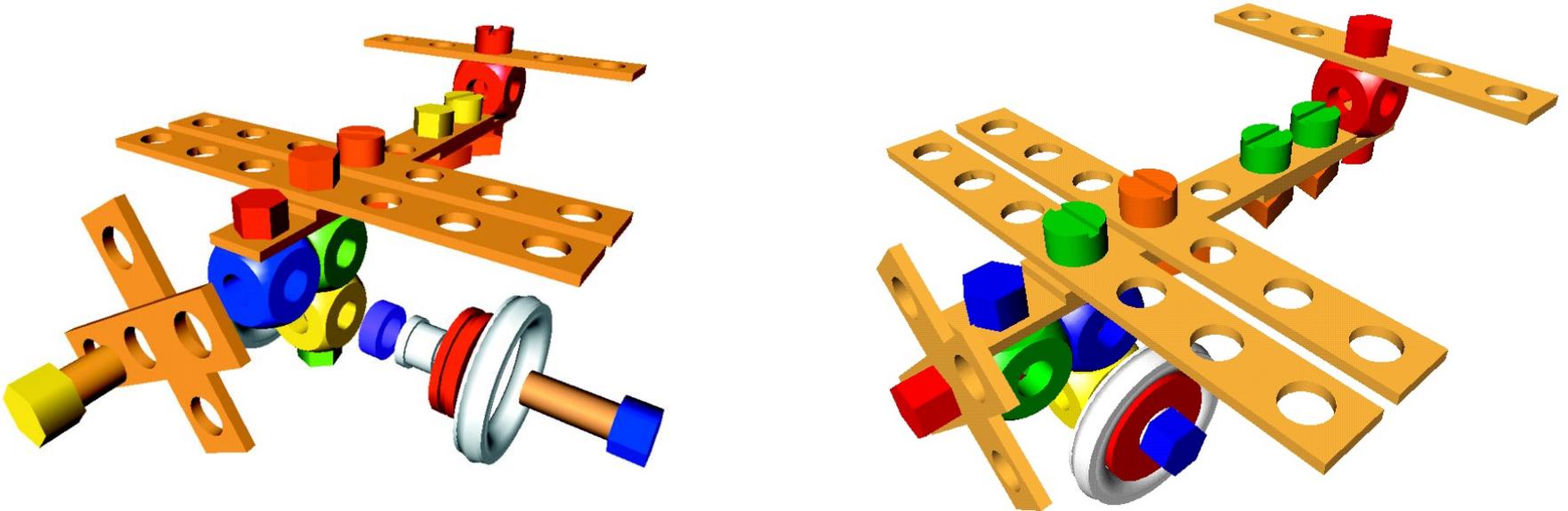
Laura schlaft



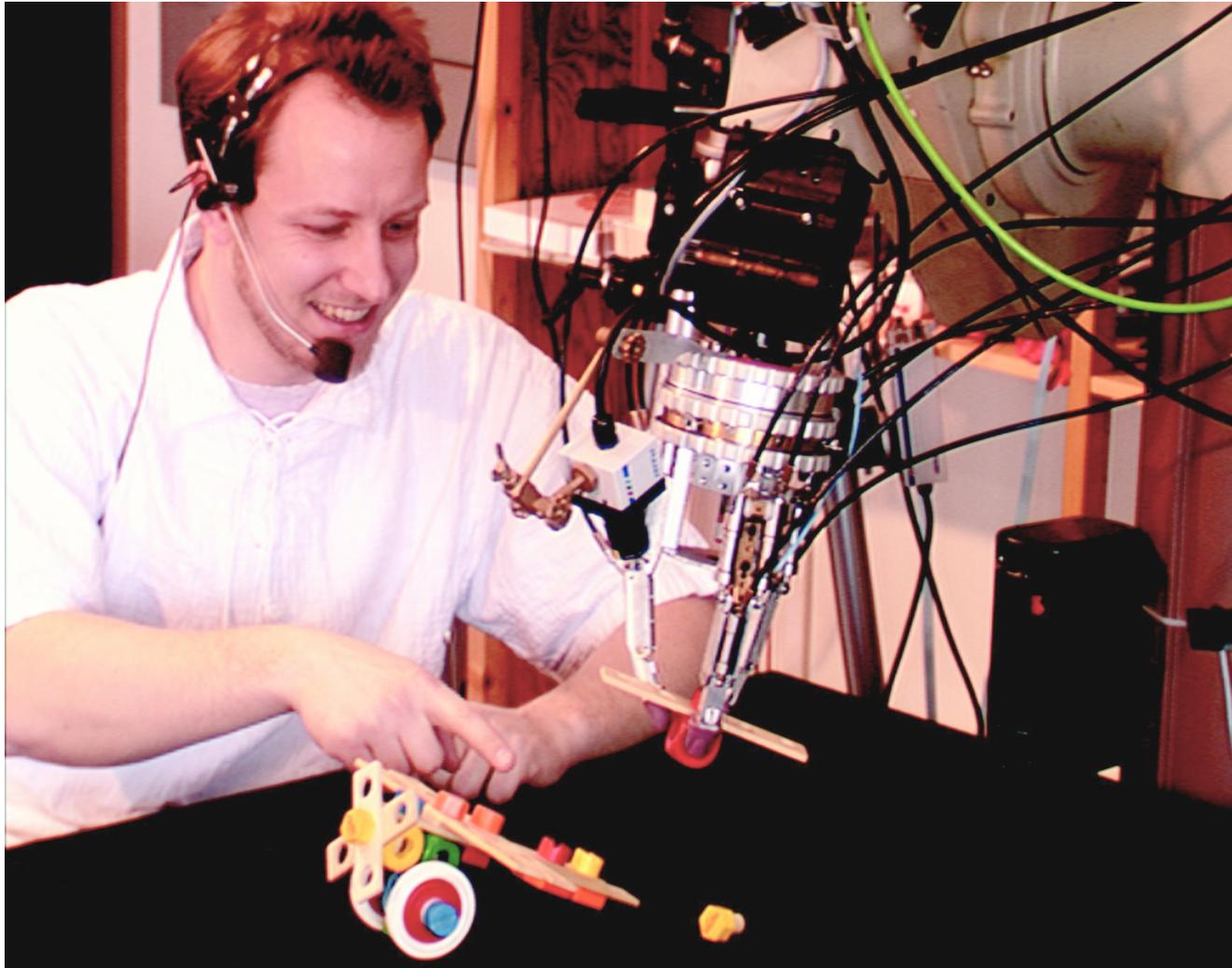
Semantik

- Semantik ist der Teilbereich der (Computer-)Linguistik, der sich mit **sprachlicher Bedeutung** befasst.
- Semantik hat einen Doppelcharakter: Sie ist **Teil der Grammatik**, und sie ist **Schnittstelle zwischen Sprache und Welt**:
 - Wie berechnen wir, auf der Grundlage von Lexikon und Satzstruktur, die Bedeutungsrepräsentationen von Sätzen und Texten?
 - Wie kodieren, verarbeiten und kommunizieren wir Wissen (über die Welt) mit Sprache?

Eine Robotik-Anwendung



früherer Sonderforschungsbereich
„Situierete Künstliche Kommunikatoren“
Bielefeld



Bedeutungsrepräsentation

- Ziel heute: wir wollen Bedeutung repräsentieren

Anforderungen an Bedeutungsrepräsentation

- **Verifizierbarkeit:** es muss möglich sein, eine Aussage mit dem Zustand der Welt zu verbinden, sodass wir testen können, ob ein Satz wahr ist.
- **Beispiel**

Das Flugzeug hat einen Propellor.

part-of (Propellor, Flugzeug)

Das Restaurant Almaz hat auch veganes Essen.

serves(Almaz, VeganFood)

Anforderungen an Bedeutungsrepräsentation

- **Eindeutigkeit:** Sprache ist ambig, aber wir wollen eine eindeutige Bedeutungsrepräsentation.
- **Beispiel**

Befestige den Propellor mit der Schraube.

Der Propellor der eine Schraub dran hat?

Benutze die Schraube um den Propellor zu befestigen?

Anforderungen an Bedeutungsrepräsentation

- **Kanonische Form:** alle Sätze, die das gleiche bedeuten, sollten gleich repräsentiert werden.
- **Beispiel**

Gibt es beim Restaurant Almaz vegane Speisen?

Bietet Almaz auch veganes Essen an?

Kriegt man bei Almaz vegane Gerichte?

Kann man bei Almaz vegan essen?

→ egal wie die Frage gestellt wird, wir wollen eine Repräsentation für unsere Datenbankabfrage.

Wie repräsentieren wir Satzbedeutungen?

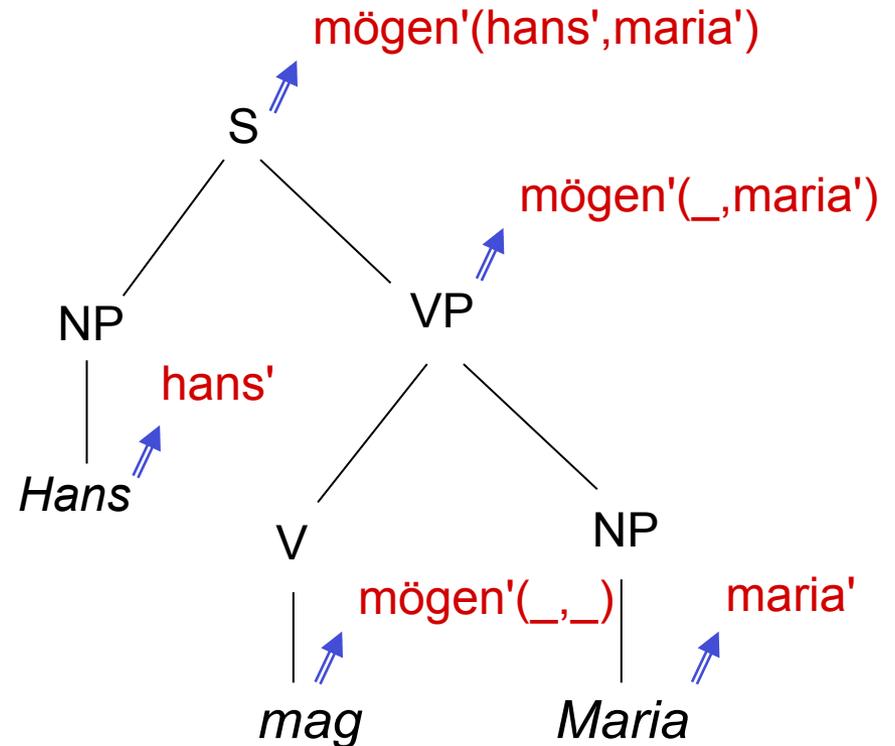
- Der Standardformalismus für die Repräsentation satzsemantischer Information ist die **Prädikatenlogik**.
- Inhaltswörter werden als Konstanten repräsentiert.
- Beispiele:
 - **laura'**, **hans'**, **maria'** sind "Individuenkonstanten"
 - **schlafen'** ist einstelliges, **mögen'** zweistelliges Prädikat
- Sätze werden als einfache Prädikat-Argument-Strukturen oder als komplexe logische Formeln präsentiert
- Beispiele
 - Laura schläft* \Rightarrow **schlafen'(laura')**
 - Hans mag Maria* \Rightarrow **mögen'(hans', maria')**
 - Jeder Student arbeitet* \Rightarrow $\forall x(\text{student}'(x) \rightarrow \text{arbeiten}'(x))$
 - Jede Teilnehmerin präsentierte ein Papier*
 \Rightarrow $\forall t (\text{teilnehmerin}'(t) \rightarrow \exists p (\text{papier}'(p) \wedge \text{präsentieren}'(t,p)))$

Wie berechnen wir die Satzbedeutung aus der syntaktischen Struktur?

„Bedeutungskomposition“ oder
„Semantikkonstruktion“

Semantik-Konstruktion: Der einfache Fall

- **Semantik-Konstruktion:** Wir bauen an der Konstituentenstruktur entlang komplexe semantische Ausdrücke "**kompositionell**" aus einfacheren Ausdrücken zusammen.



Für jeden Lexikoneintrag und jede syntaktische Regel fügen wir noch eine semantische Komponente hinzu

Semantikkonstruktion: Eine Herausforderung

Jede Teilnehmerin präsentierte ein Papier

$\forall t (\text{teilnehmerin}'(t) \rightarrow \exists p (\text{papier}'(p) \wedge \text{präsentieren}'(t,p)))$

Semantikkonstruktion: Eine Herausforderung

Jede Teilnehmerin präsentierte ein Papier

$\forall t (\text{teilnehmerin}'(t) \rightarrow \exists p (\text{papier}'(p) \wedge \text{präsentieren}'(t,p)))$

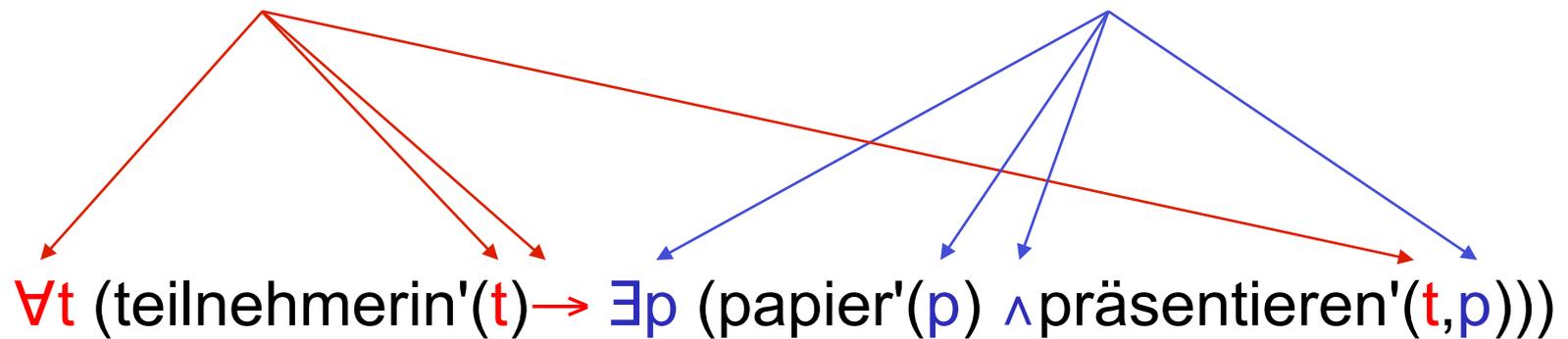
Semantikkonstruktion: Eine Herausforderung

Jede Teilnehmerin präsentierte ein Papier

$\forall t (\text{teilnehmerin}'(t) \rightarrow \exists p (\text{papier}'(p) \wedge \text{präsentieren}'(t,p)))$

Semantikkonstruktion: Eine Herausforderung

Jede Teilnehmerin präsentierte ein Papier



Semantikkonstruktion: Amiguität

Peter sieht den Mann mit dem Fernrohr

$\exists m \text{ mann}'(m) \wedge \text{sehen}'(\text{peter}', m) \wedge \exists f \text{ fernrohr}'(f) \wedge \text{tool}'(\text{sehen}', f)$
 $\exists m \text{ mann}'(m) \wedge \text{sehen}'(\text{peter}', m) \wedge \exists f \text{ fernrohr}'(f) \wedge \text{haben}'(m, f)$

Jedes Kind klettert auf einen Baum

$\forall k (\text{kind}'(k) \rightarrow \exists b (\text{baum}'(b) \wedge \text{klettert-auf}'(k, b)))$

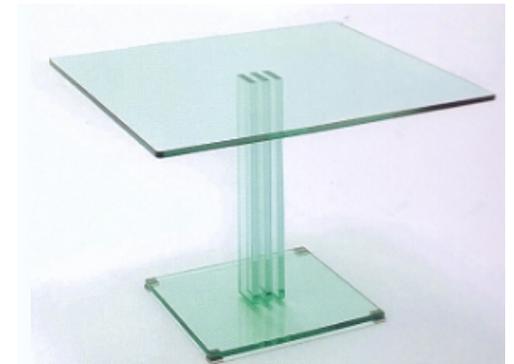
$\exists b \text{ baum}'(b) \wedge \forall k (\text{kind}'(k) \rightarrow \text{klettert-auf}'(k, b))$

Wortbedeutung

What's the meaning of life?

life'

Wortbedeutung



Tisch



Die Wort-Bedeutungs-Relation

- Tisch-Beispiel:
ein Konzept muss verschiedenen Instanzen zugeordnete werden.
- Der Wortschatz natürlicher Sprachen ist extrem mehrdeutig.
Ein Wort ist in der Regel mit verschiedenen Konzepten assoziiert:
Lexikalische Ambiguität

Die Wort-Bedeutungs-Relation

Lexikalische Ambiguität

- **Homonymie**: Ambiguität zwischen **nicht-verwandten** Konzepten
 - *Absatz, versetzen, Bank* (→ *Finanzinstitut vs. Sitzgelegenheit*)
- **Polysemie**: Ambiguität zwischen **semantisch verwandten** Konzepten
 - *Baum, leicht, Bank* (→ *Blutbank*)
- **Metonymie**: Ein Aspekt eines Konzepts wird benutzt, um auf ein anderen Aspekt oder die Entität selbst zu referieren.
 - *Bank* (→ *Gebäude in dem das Geldinstitut untergebracht ist*)
- **Homophonie**: Wort klingt gleich wird aber unterschiedlich geschrieben
 - *Leere vs. Lehre*

Die Wort-Bedeutungs-Relation

Lexikalische Ambiguität

- Genau genommen bestehen deshalb semantische Relationen nicht zwischen **Wörtern**, sondern zwischen **Konzepten** bzw. **Wortbedeutungen** (“word senses”).
- Konzepte können durch unterschiedliche Wörter realisiert werden:
Synonymie: Auto – Wagen

WordNet

- WordNet ist eine große lexikalisch-semantische Ressource: Ein Netzwerk aus semantischen Relationen zwischen Konzepten, mit der **Hyponymie**-Relation als Kern.
- Konzepte werden als „**synsets**“ repräsentiert: Mengen von synonymen Wörtern, die sich gegenseitig disambiguieren.
- Außerdem enthält WordNet für alle Synsets „Glossen“ (Umschreibungen) und Anwendungsbeispiele.

Hyponym= semantisch untergeordneter Begriff

WordNet-Synsets, Beispiel

- **Absatz**

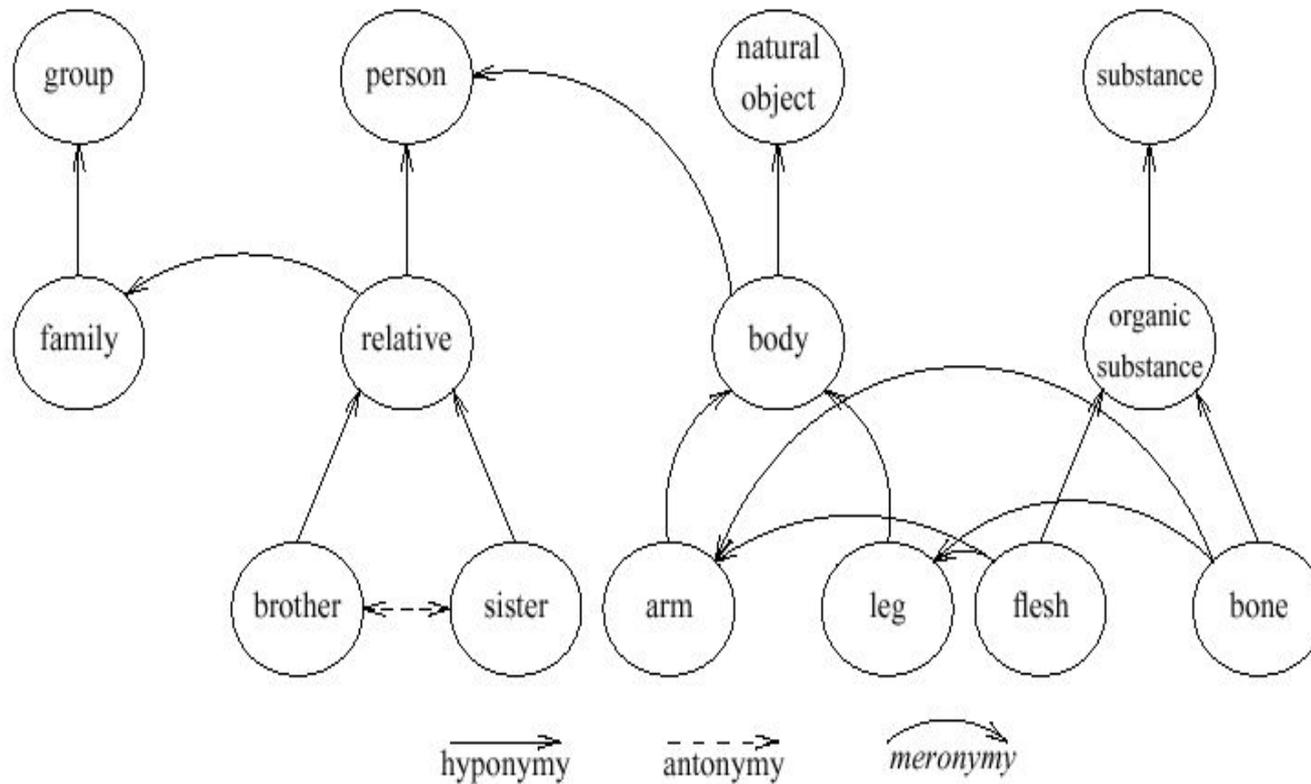
- {Absatz, Abschnitt, Paragraph}
- {Absatz, Treppenabsatz}
- {Absatz, Verkauf, Verkaufsziffern}
- ...

- <http://wordnetweb.princeton.edu/perl/webwn>

[Englisches WordNet](#)

Ein kleines Fragment von WordNet

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts



WordNet-Daten

- Englischsprachiges WordNet hat zurzeit einen Umfang von
 - 170.000 lexikalischen Einträgen (Wörtern)
 - 120.000 Synsets
- WordNet-Versionen gibt es für etwa 45 Sprachen (für das Deutsche: "GermaNet")
- WordNet wird in vielen sprach- und informationstechnologischen Anwendungen erfolgreich genutzt, insbesondere als Grundlage von Inferenz
- Englischsprachiges WordNet
 - Web Interface: <http://wordnet.princeton.edu/>
 - Allgemeine Information: <http://wordnet.princeton.edu>

Nochmal: Was ist Wortbedeutung eigentlich?



Sprachliche Bedeutung ist vielschichtig und heterogen

Zur Bedeutung können gehören:

- **Propositionale/ konzeptuelle Information** – begrifflich erfassbare, z.B. definitorische Information
- Visuelle (und andere sensorische) **prototypische Information**
- Handlungs-/ funktionsbezogene Information
- Sterotypische Information – nur im Regelfall gültig (Default-Information)

Bedeutung in Kontext und Situation

Sie beginnt um 10.

Ebenen der Bedeutung

- Linguistischer Bedeutungsgehalt, das "**semantische Potenzial**" einer Äußerung
 - Wie repräsentieren wir **Satzbedeutung**?
 - Wie **berechnen** wir die Satzbedeutung aus der syntaktischen Struktur?
 - Wie repräsentieren wir **Wortbedeutung**?
 - **Wie bringen wir Wortbedeutung und Satzbedeutung zusammen?**
- **Relevante Äußerungsinformation**
 - Wie kommen wir von der Äußerungsbedeutung zur relevanten Äußerungsinformation?
Inferenz
- **Intendierte Äußerungsbedeutung**
 - Wie ermitteln wir die Äußerungsbedeutung auf der Grundlage von Kontextinformation?
Disambiguierung oder **Ambiguitätsresolution**

Wortbedeutung und Inferenz

- Wollen wir uns um 12 in der Stadt zum Essen treffen?
Meine Vorlesung fängt um 10 an.
- Mag Peter Lammkoteletts?
Peter ist Vegetarier.
- Besitzt Maria einen Wagen?
Maria hat sich kürzlich ein Cabrio gekauft.

Inferenz

- Um **Inferenzen** auf sprachlichen Ausdrücken durchzuführen, benötigen wir **nicht unbedingt** die Repräsentation der **vollen Wortbedeutung** – nicht einmal die volle "propositionale" Information
- Hilfreich ist bereits die Kenntnis von **semantischen Relationen**, d.h., der systematischen Beziehungen, in denen Wörter zueinander stehen.
- **Beispiel:**
 - Besitzt Maria einen Wagen?
 - Maria besitzt ein Cabrio.
 - Cabrios sind Wagen.**
 - Also besitzt Maria einen Wagen.

Inferenz

- Das Standardwerkzeug für die Modellierung für Inferenzen ist die Prädikatenlogik (Deduktion)

- **Beispiel:**

Maria besitzt ein Cabrio.

Cabrios sind Wagen.

Also besitzt Maria einen Wagen.

$\exists y(\text{cabrio}'(y) \wedge \text{besitzen}'(\text{maria}', y))$

$\forall x(\text{cabrio}'(x) \rightarrow \text{wagen}'(x))$

$\exists x(\text{wagen}'(x) \wedge \text{besitzen}'(\text{maria}', x))$

Semantische Relationen

- **Synonymie:** *Wagen – Auto*
- **Hyponymie/Hypernymie**, die Unter-/Oberbegriffsrelation:
 - *Luftfrachter – Flugzeug*
 - *Dackel – Hund*
 - *Cabrio – Wagen*
- **Antonymie:** Kontrastrelation
 - *teuer – billig*
 - *Freund – Feind*
 - *Säugetier – Fisch*
- **Meronymie/ Holonymie** die Teil-von-Relation:
 - *Ast – Baum*
 - *Holz – Baum*
 - *Baum – Wald*

WordNet und Inferenz

- WordNet-Relationen können als logische Formeln reformuliert werden:

$\forall x(\text{cabrio}'(x) \rightarrow \text{wagen}'(x))$

$\forall x(\text{delphin}'(x) \rightarrow \text{säugetier}'(x))$

- Damit können sie unmittelbar für Inferenzprozesse nutzbar gemacht werden.

Ebenen der Bedeutung

- Linguistischer Bedeutungsgehalt, das "**semantische Potenzial**" einer Äußerung
 - Wie repräsentieren wir **Satzbedeutung**?
 - Wie **berechnen** wir die Satzbedeutung aus der syntaktischen Struktur?
 - Wie repräsentieren wir **Wortbedeutung**?
 - Wie bringen wir Wortbedeutung und Satzbedeutung zusammen?
- **Relevante Äußerungsinformation**
 - Wie kommen wir von der Äußerungsbedeutung zur relevanten Äußerungsinformation?
Inferenz
- **Intendierte Äußerungsbedeutung**
 - Wie ermitteln wir die Äußerungsbedeutung auf der Grundlage von Kontextinformation?
Disambiguierung oder **Ambiguitätsresolution**

Mehrdeutigkeit

Lexikalische Mehrdeutigkeit:

Bank, Absatz, Baum

Referenzielle Mehrdeutigkeit:

er, sie, es, dort, damals, der Präsident, die Vorlesung

Hans mag seinen Hund, obwohl er ihn manchmal beißt

Strukturelle Mehrdeutigkeit:

Peter sieht den Mann mit dem Teleskop (Anbindungsambiguität)

Zwei Fremdsprachen spricht jeder Linguist (Skopusambiguität)

1,2 Millionen Besucher tranken 800000 Tassen Kaffee

(distributive und kollektive Interpretation)

Automatische Disambiguierung

Sie beginnt um 10.

Ich gehe jetzt zur Bank, denn ich muss nachher einkaufen.

The box was in the pen – The pen was in the box

- Disambiguierung durch Inferenz über Alltagswissen.
- Das Inferenzproblem ist grundsätzlich lösbar (Logik, Deduktion, Theorembeweiser)
- Das unlösbare Problem ist die Kodierung von Alltagswissen in ausreichendem Umfang und in der angemessenen Form. („Knowledge Bottleneck“)
- Alternative: Statistische Modellierung

WSD und Weltwissen

- Wissensbasierte Disambiguierung durch Inferenz mit Weltwissen:
Ich gehe nachher einkaufen, deshalb muss ich jetzt dringend zur Bank.
- Aber: Unterschiedliche Verwendungen von „Bank“ benötigen unterschiedliches Wissen:
 - *Ich will ein Haus bauen, ...*
 - *Ich habe geerbt, ...*
 - *Ich brauche eine neue Kreditkarte, ...*
- Viele Wörter sind mehrdeutig, viele sind vielfach mehrdeutig, und jedes Wort erfordert spezifisches Wissen: *Bank – Absatz*
- Wissensbasierte Disambiguierung ist praktisch nicht machbar: Riesige Mengen an handkodiertem Weltwissen wären nötig: „**Knowledge Bottleneck**“
- Attraktive Alternative: **Statistische Modellierung**

WSD: Korpusannotation

- Spezifikation des **Annotationsschemas**: Übernahme von Wortbedeutungen aus einem Wörterbuch oder Thesaurus (Standard: WordNet-Synsets)
- **Annotation** aller Zielwort-Instanzen im Trainingskorpus mit einer Wortbedeutung

Trainingskorpus

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

Trainings-Korpus: Annotation mit WSD-Information

...

- (A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank **<bank1>** mit leicht schwingender Rückenlehne entwickelt. ...
- (A295) ... Ich suche noch eine Bank **<bank1>** für meinen Garten und sondiere deshalb gerade Angebote. ...
- (A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank**<bank2>** angelegt. ...
- (A452) ... Beim Test Anlageberatung der Banken **<bank2>** löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

WSD: Merkmalspezifikation

- Wir verwenden Kontextwörter als Merkmale für die Disambiguierung. Präziser ausgedrückt:
- Wir nehmen für alle Zielwörter eine gemeinsame Merkmalsmenge an: Vorkommen/ Nichtvorkommen der n (z.B. $n=1000$) häufigsten **Inhaltswörter** (Substantive, Verben, Adjektive) im Kontext des Zielwortes.
- Den Kontext einer Instanz legen wir als den Satz fest, in dem die Instanz vorkommt (alternativ: das Fenster mit fester Länge von k Wörtern rechts und links von der Instanz (z.B. $k=10$)).

Zielwörter und Kontextwörter

...

- (A237) ... Für diejenigen, denen *Komfort wichtig ist*, haben wir eine Bank **<bank1>** mit *leicht schwingender Rückenlehne* entwickelt. ...
- (A295) ... Ich *suche* noch eine Bank **<bank1>** für meinen *Garten* und *sondiere* deshalb gerade *Angebote*. ...
- (A303) ... Habe im *März* 2000 einen höheren *Betrag* bei einer Bank **<bank2>** *angelegt*. ...
- (A452) ... Beim *Test Anlageberatung* der Banken **<bank2>** *löste* kein *Institut* die einfache *Frage* nach einer *sicheren Anlage* *wirklich gut*. ...

...

WSD: Merkmalsextraktion

- Wir **lemmatisieren** die Kontextwörter (Bestimmung des Wortstamms durch morphologische Analyse)
- Wir ermitteln für jedes Vorkommen eines Zielwortes ein spezifisches Merkmalsmuster v , indem wir für jedes i : $1 \leq i \leq 1000$ setzen:
 - $v_i = 1$, wenn das Wort w_i als Kontextwort im Satz auftritt
 - $v_i = 0$ sonst.
- Alle Merkmale sind Boole'sche Merkmale ($\in \{0, 1\}$). Das Merkmalsmuster v kann als Vektor in einem hochdimensionalen Raum betrachtet werden.

Merkmalsextraktion, Beispiel

Instanz Nr	A237	A295	A303	A452	...
Annotation	bank1	bank1	bank2	bank2	...
Frage	0	0	0	1	...
Komfort	1	0	0	0	...
anlegen	0	0	1	0	...
Betrag	0	0	1	0	...
Garten	0	1	0	0	...
suchen	0	1	0	0	...
fahren	0	0	0	0	...
richtig	0	0	0	0	...
Test	0	0	0	1	...
...

Neue Instanz:

*Keine Frage: In
einen ordentlichen
Garten gehören
neben einer Bank
auch die richtigen
Möbel.*

Neu
?
1
0
0
0
1
0
0
1
0
...