

Einführung in die Computerlinguistik

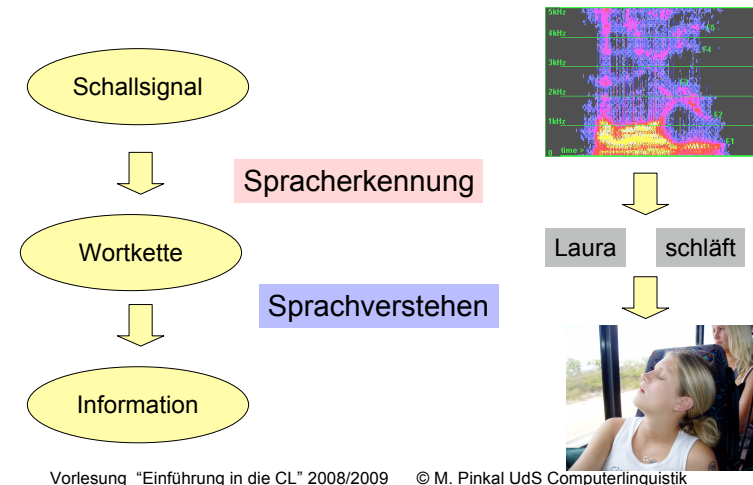
Semantik

WS 2008/2009

Manfred Pinkal

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Was ist Sprachverarbeitung?



Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Semantische Interpretation: Ein Beispiel

Sie beginnt um 10.

- Wozu semantische Interpretation?
 - Generalisierung/ Abstraktion von „Zufälligkeiten“ der sprachlichen Oberfläche
 - Disambiguierung/Resolution: Bestimmung der vom Sprecher intendierten/ kontextspezifischen Äußerungsbedeutung
 - Auswertung/Inferenz: Übergang von der kontextspezifischen Bedeutung zur (in der Situation/ für die Aufgabe) relevanten Information

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Semantik für Dialogverarbeitung

U: Ist der Flug um 2 nach London pünktlich?

S: Ihr Flug nach London Heathrow geht pünktlich um 14:05 ab und steht zum Einsteigen bereit.

U: Und der nächste?

S: Der nächste Flug nach London Gatwick geht um 15.20.

U: Gibt es auch einen nach Heathrow?

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Semantik für Maschinelle Übersetzung

D: *Geht es um 2 bei Ihnen?*

E1: *Is 2 p.m. ok for you?*

E2: *Can we meet at 2 in your office?*

Semantik für den Informationszugriff

Der US-Flugzeughersteller Lockheed hat von Grossbritannien den Auftrag fuer 25 Transportflugzeuge des Typs Hercules C130J erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.

Wieviel kosten die Flugzeuge, die Lockheed an Großbritannien verkauft?

Informationszugriff: Zwei grundlegende Verfahren

- Der Zugriff auf explizit oder implizit in einem Dokument/ einer Äußerung bzw.
 - Der Test, ob eine Information H in einem Dokument/ einer Äußerung T enthalten ist.
- Die Aufgabe kann auf zwei Arten gelöst werden, die grundlegend unterschiedlich sind:
 - Berechnung semantischer Ähnlichkeit zwischen T und H: Enthalten T und H ausreichend ähnliches lexikalisches Material?
 - Bestimmung, ob T und H in einer logischen Folgerungsbeziehung stehen.

Ähnlichkeitsbasierter Ansatz

Question: *Are dolphins mammals?*

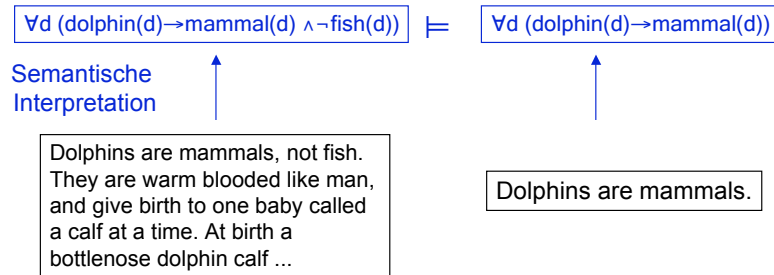
Dolphins are mammals, not fish.
They are warm blooded like man,
and give birth to one baby called
a calf at a time. At birth a
bottlenose dolphin calf ...

≈

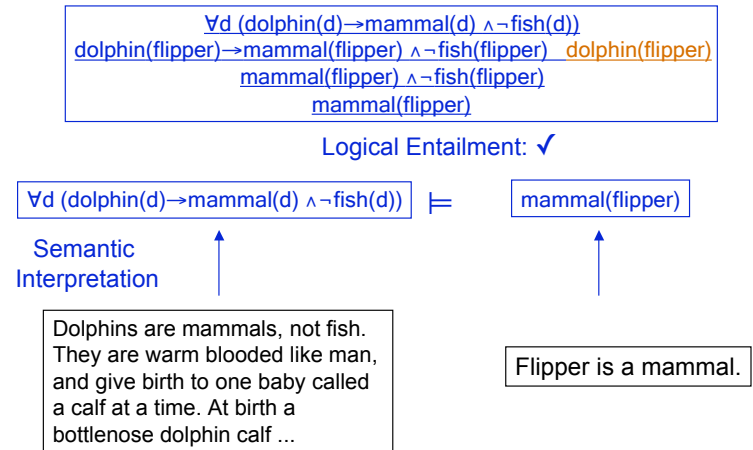
Dolphins are mammals.

Logik-basierter Ansatz

Logische Folgerung: ✓



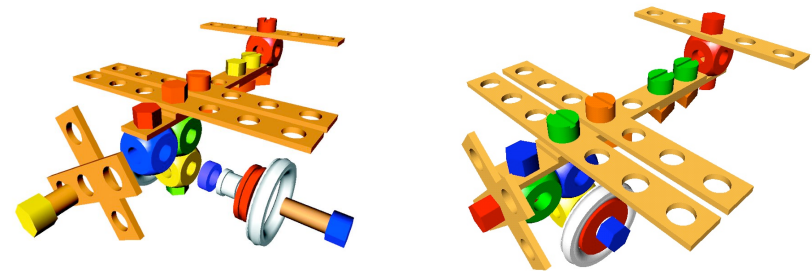
Ist Flipper ein Säugetier?



Was ist Wortbedeutung?



Eine Robotik-Anwendung



Sonderforschungsbereich
 „Situierter Künstliche Kommunikatoren“
 Bielefeld



Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Wortbedeutung ist vielschichtig und heterogen

- Zur Wortbedeutung können gehören:
 - Propositionale/ konzeptuelle Information – begrifflich erfassbar, in einem logischen Framework darstellbar
 - Visuelle (und andere perzeptuelle) prototypische Information
 - Handlungsbezogene Information
 - Sterotypische Information – nur im Regelfall gültig (Default-Information)
- Es gibt keine scharfe Grenze zwischen sprachlicher Bedeutung und nicht-sprachlichem Wissen

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Beschreibung des semantischen Lexikons

- Vollständige semantische Beschreibung eines Wortschatzes ist nicht möglich.
- Lexikalische Bedeutungsbeschreibung ist notwendigerweise partiell - auf eine Schicht/ bestimmte Schichten semantischer Information eingeschränkt.
- Lexikalisch-semantische Beschreibung für die Sprachtechnologie sollte sich an den Erfordernissen von Anwendungen bzw. Anwendungstypen orientieren.
- Die Computerlinguistik hat sich bisher hauptsächlich an der Aufgabe des Informationszugriffs orientiert, und sich mit der Schicht der lexikalisch-semantischen Relationen befasst.

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Lexikalische Semantik und Informationszugriff

*Der US-Flugzeughersteller Lockheed hat von Grossbritannien den Auftrag fuer 25 **Flugzeuge** des Typs Hercules C130J erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.*

- *Wieviele kosten die **Maschinen**, die Lockheed an Großbritannien verkauft hat?*

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Elementare semantische Relationen

- **Synonymie**: Flugzeug - Maschine - Flieger
- **Hypernymie / Hyponymie**, die Ober-/Unterbegriffsrelation:
 - *Flugzeug* - *Propellermaschine*
 - *Tier* - *Hund*
 - *töten* - *umbringen*
- **Meronymie/ Holonymie** die Teil-von-Relation:
 - Ast - Baum
 - Holz - Baum
 - Baum - Wald
- **Antonymie**: Kontrastrelation
 - gut - schlecht
 - teuer - billig
 - Freund - Feind

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Die Wort-Bedeutungs-Relation

- Genau genommen bestehen semantische Relationen nicht zwischen **Wörtern**, sondern zwischen **Konzepten** bzw. **Wortbedeutungen** ("word senses"): Die Abbildung zwischen phonologischen/ orthographischen Wörtern und Konzepten ist in beiden Richtungen nicht eindeutig.
- Ein Konzept kann in verschiedenen Wörtern kodiert sein: **Synonymie**
- Ein Wort ist in der Regel mit verschiedenen Wörtern assoziiert: **Lexikalische Ambiguität**
- Ambiguität zwischen nicht-verwandten Konzepten heißt **Homonymie**
 - *Bank*: *Geldinstitut* / *Bank*: *Sitzmöbel*
- Ambiguität zwischen semantisch verwandten Konzepten heißt **Polysemie**
 - *Maschine*: *Flugzeug* / *Maschine*: *Motorrad*
 - *Sitz im Auto* / *im Bundestag*
 - *einen Plan* / *einen Koffer aufgeben*

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

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.
- Synsets liefern direkte Information zur Synonymierelation und zur Wort-Konzept-Abbildung: Ein orthographisches Wort hat genau die Synsets als Lesarten, in denen es als Element vorkommt.
- Dies gilt im Prinzip. praktisch sind meist keine oder zu wenige Synonyme vorhanden, um Konzepte eindeutig zu charakterisieren. Deshalb enthält WordNet für alle Synsets „Glossen“ (Umschreibungen) und Anwendungsbeispiele.

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Beispiel: Lesarten von car

- **S:** (n) **car**, auto, automobile, machine, motorcar
- **S:** (n) **car**, railcar, railway car, railroad car
- **S:** (n) **car**, gondola
- **S:** (n) **car**, elevator car
- **S:** (n) cable car, **car**

Vorlesung "Einführung in die CL" 2008/2009 © M. Pinkal UdS Computerlinguistik

Beispiel: Synset, Glosse, Sprachbeispiel

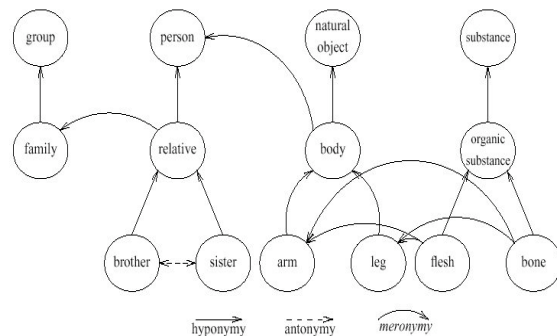
- **car**
 - { car, auto, automobile, machine, motorcar }
 - a motor vehicle with four wheels; usually propelled by an internal combustion engine
 - "he needs a car to get to work"

Hyponyme von *motor vehicle*

- **S:** (n) **motor vehicle**, **automotive vehicle** (a self-propelled wheeled vehicle that does not run on rails)
- **direct hyponym / full hyponym**
 - **S:** (n) **amphibian**, **amphibious vehicle** (a flat-bottomed motor vehicle that can travel on land or water)
 - **S:** (n) **bloodmobile** (a motor vehicle equipped to collect blood donations)
 - **S:** (n) **car**, **auto**, **automobile**, **machine**, **motorcar** (a motor vehicle with four wheels; usually propelled by an internal combustion engine) "he needs a car to get to work"
 - **S:** (n) **doodlebug** (a small motor vehicle)
 - **S:** (n) **four-wheel drive**, **4WD** (a motor vehicle with a four-wheel drive transmission system)
 - **S:** (n) **go-kart** (a small low motor vehicle with four wheels and an open framework; used for racing)
 - **S:** (n) **golfcart**, **golf cart** (a small motor vehicle in which golfers can ride between shots)
 - **S:** (n) **hearse** (a vehicle for carrying a coffin to a church or a cemetery; formerly drawn by horses but now usually a motor vehicle)
 - **S:** (n) **motorcycle**, **bike** (a motor vehicle with two wheels and a strong frame)
 - **S:** (n) **snowplow**, **snowplough** (a vehicle used to push snow from roads)
 - **S:** (n) **truck**, **motortruck** (an automotive vehicle suitable for hauling)

Ein kleines Fragment von WordNet

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



WordNet-Daten

- Englisch WordNet hat zurzeit einen Umfang von
 - 150.000 lexikalischen Einträgen (Wörtern)
 - 120.000 Synsets
- WordNet-Versionen gibt es für etwa 45 Sprachen
- WordNet wird in vielen sprach- und informationstechnologischen Anwendungen erfolgreich genutzt.
- "GermaNet": eine deutsche WordNet-Version mit etwa 100.000 lexikalischen Einträgen
- Englisch WordNet
 - Web Interface: <http://wordnet.princeton.edu/perl/webwn>
 - Allgemeine Information: <http://wordnet.princeton.edu>

WordNet und Inferenz

- **WordNet-Relationen** können als **logische Formeln** reformuliert werden:
 $\forall x(\text{family}(x) \rightarrow \text{group}(x))$
 $\forall x(\text{person}(x) \rightarrow \exists y(\text{substance_m}(y, x) \wedge \text{body}(y)))$
 $\forall x(\text{body}(x) \rightarrow \exists y(\text{part_m}(y, x) \wedge \text{leg}(y)))$
 $\forall x(\text{body}(x) \rightarrow \exists y(\text{part_m}(y, x) \wedge \text{arm}(y)))$
- Für die semantische Verarbeitung in einem **logischen Framework** kann WordNet als große Datenbasis verwendet werden, die zusätzliche Axiome für die Inferenz bereitstellt.
- In einem ähnlichkeitsbasierten Framework kann WordNet-Information verwendet werden, um bei der Suche nach relevanter Information größere Abdeckung zu erreichen: Die Suchanfrage kann z.B. um Synonyme, Ober- und Unterbegriffe erweitert werden („query expansion“).

Lexikalische Semantik und Informationszugriff

*Der US-Flugzeughersteller Lockheed hat von Grossbritannien den Auftrag fuer 25 **Flugzeuge** des Typs Hercules C130J erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.*

- *Wieviel kosten die **Maschinen**, die Lockheed an Großbritannien verkauft hat?*

Lexikalische Semantik und Informationszugriff

*Der US-Flugzeughersteller Lockheed hat von Grossbritannien den Auftrag fuer 25 **Abfangjäger** erhalten. Vertreter des Verteidigungsministeriums bezifferten den Wert des Auftrags mit umgerechnet 2,5 Mrd. DM.*

- *Wieviel kosten die **Flugzeuge**, die Lockheed an Großbritannien verkauft hat?*

Satzsemantik: Wie kommen wir von Wortbedeutungen zu Satzrepräsentationen?

- Der Standardformalismus für die Repräsentation satzsemantischer Information ist die **Prädikatenlogik**.
- **Prädikat-Argument-Strukturen** als elementare Satzbausteine der logischen Repräsentation
- Kombination zu komplexen Repräsentationen durch **logische Operatoren** (Junktoren und Quantoren)

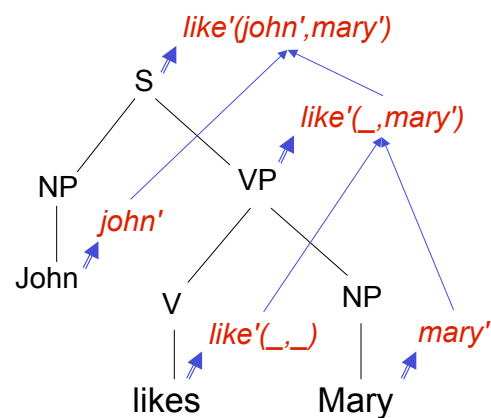
Prädikat-Argument-Strukturen

- **Prädikat-Argument-Strukturen** als elementare Satzbausteine der logischen Repräsentation
 - arbeiten(peter)
 - kennen(hans, maria)
 - bruder-von(peter, maria)
 - geben(maria, peter, das-buch)

Satzsemantik: Komposition

- Die **Bedeutung eines Satzes** berechnet sich aus
 - den **Bedeutungen der Wörter**, aus denen er sich zusammensetzt, und
 - seiner **syntaktischen Struktur**

Bedeutungskomposition: Ein Beispiel



Grammatikfragment 3 (Semantik): Lexikon

arbeitet:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{VFORM} \\ \text{SUBJ} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{SEM} <1> \\ \text{PRED} \text{arbeiten'} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$	Studentin:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{NUM} \text{sg} \\ \text{GEN} \text{f} \end{array} \right] \end{array} \right] \end{array} \right]$
	$\left[\begin{array}{c} \text{SUBCAT} \left[\begin{array}{c} \text{SEM} \left[\begin{array}{c} \text{arg} <1> \\ \text{empty} \end{array} \right] \end{array} \right] \end{array} \right]$		$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{NUM} \text{sg} \\ \text{GEN} \text{n} \end{array} \right] \end{array} \right] \end{array} \right]$
		Fach:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{NUM} \text{sg} \\ \text{GEN} \text{n} \end{array} \right] \end{array} \right] \end{array} \right]$
			$\left[\begin{array}{c} \text{SUBCAT} \left[\begin{array}{c} \text{SEM} \left[\begin{array}{c} \text{fach'} \end{array} \right] \end{array} \right] \end{array} \right]$
wählt:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{VFORM} \\ \text{SUBJ} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{SEM} <1> \\ \text{PRED} \text{wählen'} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$	das:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{NUM} \text{sg} \\ \text{GEN} \text{n} \end{array} \right] \end{array} \right] \end{array} \right]$
	$\left[\begin{array}{c} \text{SUBCAT} \left[\begin{array}{c} \text{SEM} \left[\begin{array}{c} \text{ARG1} <1> \\ \text{ARG2} <2> \end{array} \right] \end{array} \right] \end{array} \right]$	die:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{GEN} \text{f} \end{array} \right] \end{array} \right] \end{array} \right]$
		der:	$\left[\begin{array}{c} \text{CAT} \\ \text{HEAD} \left[\begin{array}{c} \text{AGR} \left[\begin{array}{c} \text{GEN} \text{n} \end{array} \right] \end{array} \right] \end{array} \right]$

Grammatikfragment 3 (Semantik): Regeln und Constraints

$S \rightarrow NP VP$
 $\langle S HEAD \rangle = \langle VP HEAD \rangle$
 $\langle S HEAD SUBJ \rangle = \langle NP HEAD \rangle$
 $\langle S HEAD AGR \rangle = \langle S HEAD SUBJ AGR \rangle$

$VP \rightarrow V$
 $\langle VP HEAD \rangle = \langle V HEAD \rangle$
 $\langle V SUBCAT \rangle = \text{empty}$

$VP \rightarrow V NP$
 $\langle VP HEAD \rangle = \langle V HEAD \rangle$
 $\langle V SUBCAT \rangle = \langle NP \rangle$

$NP \rightarrow DET N$
 $\langle NP HEAD \rangle = \langle N HEAD \rangle$
 $\langle NP HEAD AGR \rangle = \langle DET AGR \rangle$

Komplexe prädikatenlogische Formeln

- Kombination zu komplexen Repräsentationen durch logische Operatoren (Junktoren und Quantoren)

The big red block is in the box.

$\text{big}(b) \wedge \text{red}(b) \wedge \text{block}(b) \wedge \text{box}(x) \wedge \text{in}(b,x)$

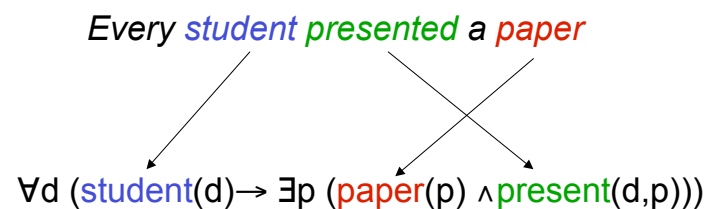
Delfine sind Säugetiere.

$\forall d (\text{delfin}(d) \rightarrow \text{säugetier}(d))$

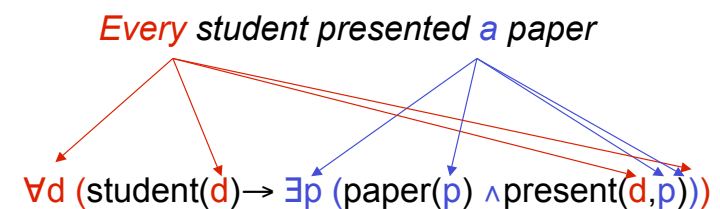
Delfine leben im Meer.

$\forall d (\text{delfin}(d) \rightarrow \exists m (\text{meer}(m) \wedge \text{lebt-in}(d,m)))$

Bedeutungskomposition: Eine Herausforderung



Eine Herausforderung für die semantische Komposition



Wie werden komplexe semantische Repräsentationen berechnet?

- Pragmatische Näherungslösung: Semantik-Konstruktion durch Unifikationsgrammatik mit Reentrancy
- Grundsätzliche Lösung:
Semantik-Konstruktion durch höherstufige Logik, λ -Kalkül

Semantik-Konstruktion

