

# Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Ressourcen zur Wortbedeutung

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung  
Universität Stuttgart

15. Januar 2010

- **Lexikon**: Bestandteil einer linguistischen Theorie, die sich mit den Eigenschaften von Lexemen befaßt (Engelberg & Lemnitzer, 2004)
- Ressourcen zu lexikalischem Wissen:
  - **manuelle Ressourcen**, z.B. Wörterbücher, Thesauri, Taxonomien, Ontologien, Klassifikationen etc.
  - **automatische Erstellung von Ressourcen** auf der Basis von (annotierten) Korpus-Daten

- **Wörterbuch**: Datensammlung mit äußerer Zugriffsstruktur, die sprachliche Angaben zu lexikalischen Einheiten wie Wörtern, Wendungen, Morphemen etc. oder zu Begriffen enthält (Engelberg & Lemnitzer, 2004)
- Beispiele: Bedeutungswörterbuch, Rechtschreibwörterbuch, Stilwörterbuch, etymologisches Wörterbuch, Fremdwörterbuch, Synonymenwörterbuch, Fachwörterbuch etc.

# Wörterbücher: Beispiele

- Duden
- Wissenmedia
- Langenscheidt
- Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS)
- Deutscher Wortschatz Leipzig
- [canoonet](#): Deutsche Wörterbücher und Grammatik
- Wiktionary
- LEO: Deutsch, Englisch, Französisch, Spanisch, Chinesisch
- Dictionary.com

- **Enzyklopädie**: Datensammlung mit äußerer Zugriffsstruktur, die Sachinformationen zu lexikalischen Einheiten enthält (Engelberg & Lemnitzer, 2004)
- Inhaltliche Unterscheidung zwischen Wörterbuch und Enzyklopädie (sprachlich-lexikalisch vs. sachlich-enzklopädisch) ist nicht immer unproblematisch
- *Inferential* (how to use language) vs. *referential* (incorporating a knowledge of the world) abilities (Marconi 1997)

- Brockhaus
- Britannica
- Wikipedia
- Reference.com

- **Thesaurus:** inhaltsparadigmatisches Wörterbuch (Engelberg & Lemnitzer, 2004)
- **Thesaurus:** 1. wissenschaftliches Wörterbuch mit dem Ziel, den Gesamtwortschatz einer Sprache zu kodifizieren; 2. nach Sachgebieten bzw. Bedeutungsähnlichkeit gegliedertes Wörterbuch (Bußmann, 1990)
- [Roget's Thesaurus](#)
- [openthesaurus](#)
- [Thesaurus.com](#)

- **Ontologie:** explizite Spezifikation einer gemeinsamen Konzeptualisierung; formale intersubjektive sprachunabhängige Repräsentation der Welt, die von den Aufgaben, Zielen, Handlungen, Einstellungen etc. eines Agenten sowie von den Spezifika einzelner Situationen abstrahiert (Carstensen, 2004)
- beschreibt sowohl die Primitive der konzeptuellen Ebene bzgl. ihrer möglichen Interpretationen, als auch die möglichen Optionen auf der epistemologischen Ebene
- Unterscheidung zwischen bereichsspezifischen und allgemeinen Aspekten
- DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering)
- SUMO (Suggested Upper Merged Ontology)
- Cyc

# Semantische Relationen

- Relationen zwischen zwei Wortbedeutungen
- Beispiele:
  - **Synonymie/Plesionymie**: identische oder sehr ähnliche Wortbedeutung, z.B. *Kleinkrieg–Fehde, hacke–betrunken, helfen–unterstützen*
  - **Antonymie**: gegensätzliche Wortbedeutungen, z.B. *gelingen–misslingen, richtig–falsch, auftauchen–verschwinden, rot–grün, Lehrer–Schüler*
  - **Hyperonymie**: Über-/Unterordnung von Wortbedeutungen, z.B. *Amsel–Vogel, Kaffee–Getränk, joggen–rennen; kochen–kreieren*
  - **Meronymie**: Teil/Ganzes-Beziehung von Wortbedeutungen, z.B. *Tür–Hause, Blatt–Baum*
  - **Kausalität**: Ursächlichkeit, z.B. *zeigen–sehen, geben–haben*
  - **Logische Konsequenz**, z.B. *töten–sterben*
- Repertoire abhängig von Wortklasse

# Semantische Relationen

- **Paradigmatische vs. syntagmatische Relationen:**  
vertikale, Mengen-basierte Ersetzbarkeit von Elementen derselben Wortklasse (*Katze/Hund/Tier*) vs.  
horizontale, syntaktische Sequenz von Elementen (*Eis essen; aus und vorbei*)
- **Assoziative Relationen:** unterspezifizierte (semantische) Relation, z.B. *abgewöhnen–aufhören, auftauen–Wasser, fliegen–Luft, Affe–Zoo, Badewanne–Schaum, Polizist–grün*
- **Situationenabhängige Relationen:** Wortbedeutungen (z.B. Teilnehmer, Orte, Geschehnisse), die einer bestimmten Situation eigen sind, z.B. *backen, Küche, Koch, Herd*

- Lexikalisch-semantisches Online-Wortnetz zu Wortbedeutungen und ihren semantischen Relationen
- Ursprung: psycholinguistische Theorien des lexikalischen Gedächtnisses
- Wortbedeutung ist als Konzeptknoten mit seinen semantischen Verknüpfungen repräsentiert
- Englische Nomen, Verben, Adjektive und Adverbien werden in Mengen von Synonymen (*synsets*) organisiert
- Jedes Synset repräsentiert ein lexikalisches Konzept
- Ursprüngliche Entwicklung: Cognitive Science Laboratory at Princeton University
- WordNet existiert mittlerweile für mehr als 50 Sprachen

# WordNet: Synsets

- Synsets sind Mengen von Synonymen/Plesionymen
- Ambige Wörter werden entsprechend ihrer Mehrdeutigkeit mehreren Synsets zugeordnet
- Beispiele:
  - {*Kaffee*} (nichtalkoholisches Getränk) vs.  
{*Kaffee, Kaffeetrinken, Kaffeeklatsch, Kaffeetafel*}
  - {*kalt*} (körpergefühlsspezifisch) vs.  
{*kalt*} (temperaturspezifisch) vs.  
{*kalt, hartherzig, hart, kaltherzig*}
  - {*abnehmen*} (verringern) vs. {*abnehmen, abmachen*} (entfernen) vs.  
{*abnehmen, wegnehmen*} (nehmen) vs. {*abnehmen*} (Diät) vs.  
{*ab-, herunternehmen*} (abhängen) vs. {*abnehmen*} (helfen) vs.  
{*abnehmen, abkaufen*} (glauben) vs. {*abnehmen*} (verlangen) vs.  
{*abnehmen, checken, begutachten, abchecken*} (kontrollieren)

# WordNet: Relationen

Innerhalb von Synsets:

- Synonymie/Plesionymie

Zwischen Synsets oder Teilen von Synsets derselben Wortklasse:

- Antonymie
- Hypernymie/Hyponymie
- Meronymie/Holonymie
- ...

Zwischen Synsets oder Teilen von Synsets verschiedener Wortklassen:

- Evokation

Morphologie:

- Komposita
- Derivation

# WordNet: Synset-Beschreibung

- Eindeutige Synset-Nummer (*offset*)
- Liste von Wörtern im Synset
- Liste von Relationen zu anderen Wörtern/Synsets
- Erläuterung (*gloss*)
- Verwendungsbeispiel
- (Subkategorisierung)

Beispiel:

01546841 01 kontrollieren

008 @ 01546607 ~ 01547143 ~ 01547358 ~ 01547526 ~ 01547673 ~ 01547894  
~ 01548245 ~ 01548506

'zur Überwachung/Untersuchung o.ä. Kontrollen durchführen',  
'Der Pilot kontrollierte seine Instrumente.'

# GermaNet 5.2 (Dez. 2009): Statistik Synsets

	Wörter	Synsets	Wortbedeutungen
Adjektive	7,650	5,550	8,130
Nomen	60,851	46,735	64,315
Verben	8,480	9,290	12,414
Summe	76,981	61,575	84,859

Lesarten pro Wort: 1,1

Wortbedeutungen pro Synset: 1,38

Lexikalische Relation	Anzahl
Synonymie	23,284
Antonymie	1,579
Pertonymie	1,701

Konzeptuelle Relation	Anzahl
Hyperonymie	66,887
Meronymie/Holonymie	1,005 / 4,152
Entailment	21
Assoziation	1,358
Causation	201

- WordNet existiert mittlerweile für mehr als 50 Sprachen
- Sprachübergreifend werden WordNets zu multi-lingualen Ressourcen zusammengefasst: EuroWordNet, BalkaNet, AsianWordNet, etc.
- EuroWordNet verwaltet die WordNets als autonome, sprachspezifische Strukturen; die sprachübergreifende Verbindung ist durch einen *inter-lingual index (ILI)* realisiert
- WordNet-Webseiten:
  - Global WordNet Association
  - WordNet
  - GermaNet
  - EuroWordNet

- Lexikalische Dokumentation der syntaktischen und semantischen Valenzen eines Wortes für sämtliche Wortbedeutungen
- Ressourcen: Lexikalische Datenbank und Korpus-Annotation
- Dokumentation für englische Verben, Nomen und Adjektive
- Theorie: Frame-Semantik (Fillmore, 1977)
- Sprachübergreifende Vernetzung (englisch-deutsch)
- Vernetzung mit Ontologie

- **Frame-Semantik:** Theorie, die linguistische Semantik mit enzyklopädischem Wissen verknüpft
- Beschreibung von Wortbedeutungen durch das Hintergrundwissen, das notwendig ist, um das Wort bzw. den Satz zu verstehen
- **Frame:** konzeptuelle Struktur, die prototypische Situationen modelliert
- **Frame-Element:** Wort oder Ausdruck, der Frame hervorruft
- **Frame-Rolle:** Teilnehmer und Eigenschaften einer Frame-Situation
- Jede Bedeutung eines ambigen Wortes gehört zu einem anderen semantischen Frame, z.B. *cut.v*: intentional traversing, change direction, cutting, cause change of position on a scale, cause harm, experience bodily harm

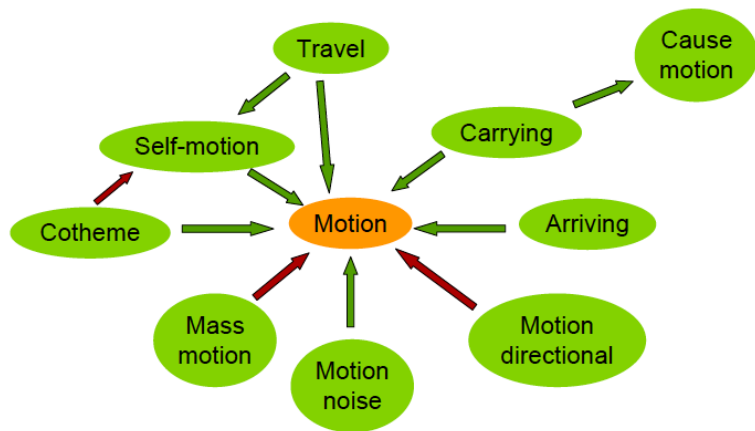
# FrameNet: Frame-Beispiele

- **apply heat**: common situation involving a **cook**, some **food**, and a **heating instrument**;  
Elemente: *bake, blanch, boil, broil, brown, simmer* etc.
- **change position on a scale**: situation involving the change of an **item**'s position on a scale (the **attribute**) from a starting point (**initial value**) to an end point (**final value**);  
Elemente: *decline, decrease, gain, rise* etc.
- **damaging**: an **agent** affects a **patient** in such a way that the **patient** (or some **subregion** of the **patient**) ends up in a non-canonical state;  
Elemente: *damage, sabotage, scratch, tear, vandalise* etc.

# FrameNet: Annotationsbeispiele

- Verben:  
[*Cook* Matilde] **fried** [*Food* the catfish] [*Heating instrument* in a heavy iron skillet].  
[*Item* Colgate's stock] **rose** [*Difference* \$3.64] [*Final value* to \$49.94].
- Nomen:  
... the **reduction** [*Item* of debt levels] [*Final value* to \$25] [*Initial value* from \$2066]
- Adjektive:  
[*Sleeper* They] were **asleep** [*Duration* for hours].

# FrameNet: *Inheritance* und *Use*



Use

Inheritance

[Abbildung von Katrin Erk]

# FrameNet: Vererbungsbeispiel

- **Frame:** *Transportation*
  - **Frame-Elemente:** mover, means, path
  - **Situation:** mover moves along path by means
- **Frame:** *Driving*
  - Vererbung von *Transportation*
  - **Frame-Elemente:** driver=mover, rider=mover, cargo=mover, vehicle=means
  - **Situationen:** driver starts vehicle, driver controls vehicle, driver stops vehicle
- **Korpus-Beispiel:**

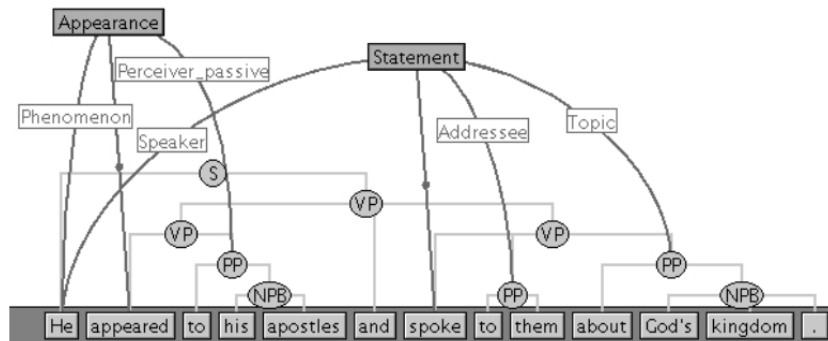
Now [*Driver* Tim] was **driving** [*Rider* his guest] [*Path* to the station].

# FrameNet: Sprachen

- [Englisches FrameNet](#)
- [Deutsches FrameNet \(SALSA\)](#)
- [Japanisches FrameNet](#)
- [Spanisches FrameNet](#)
  
- [Kicktionary: mehrsprachiges elektronisches Fußballwörterbuch](#)

- FrameNet für deutsche Verben, Nomen und Adjektive
- Verlinkung mit englischer lexikalischer Datenbank
- Übertragung von Frames ins Deutsche
- Vorgehensweise: Korpus-gesteuert statt Frame-gesteuert
- Annotation der TIGER-Treebank mit semantischen Rollen

# TIGER/SALSA: Beispiel



[Abbildung von Sebastian Padó]

# PropBank und NomBank

- Korpus-Annotation mit semantischen Propositionen; Fokus: Verben (PropBank) und Nomen (NomBank)
- Prädikat-Argument-Strukturen (semantische Rollen) werden zu den syntaktischen Bäumen der Penn Treebank hinzugefügt
- Jede Verb-/Nomen-Instanz wird berücksichtigt
- Ziel: Modellierung von Alternationsverhalten durch breite Korpus-Annotation
- Link zwischen semantischen Rollen und syntaktischer Realisierung (Levin, 1993)
- Zweck: nützliche Ebene für empirische Studien

- Framing:
  - ① Untersuchung von Korpus-Beispielen für ein bestimmtes Verb
  - ② Gruppierung der Beispiele in eine oder mehrere Verb-Bedeutungen
  - ③ Jede wesentliche Bedeutung wird durch eine Argumentstruktur modelliert
  
- Annotation:
  - ① Wortarten-Annotation durch einen Regel-basierten Tagger; Genauigkeit: 83%
  - ② Manuelle Korrektur der Tagger-Ausgabe
  - ③ Adjudikation der Annotatoren-Ergebnisse

- Semantische Rollen werden auf Verb-Basis definiert
- Verb-Argumente werden numeriert, beginnend mit 0.  
*Arg0* bezieht sich normalerweise auf den prototypischen Agens, *Arg1* auf den prototypischen Betroffenen oder das Thema
- **Rollen-Menge**: Menge semantischer Rollen für die verschiedenen Realisierungen eines Verbs
- **Rahmen-Menge**: Rollen-Menge, die mit einer Menge von syntaktischen Rahmen assoziiert wird
- **Modifikator-Rollen**: semantische Rollen, die sich auf alle Verben beziehen können, z.B. *LOC* (*location*), *NEG* (*negation*), *CAU* (*cause*)

# PropBank: Beispiele

- Frameset **accept<sub>1</sub>** 'take willingly'

Arg0: acceptor

Arg1: thing accepted

Arg2: accepted-from

Arg3: attribute

[Arg0 He] [ArgM-mod would] [ArgM-neg n't] accept [Arg1 anything of value] [Arg2 from those he was writing about].

- Frameset **kick<sub>1</sub>** 'drive or impel with the foot'

Arg0: kicker

Arg1: thing kicked

Arg2: instrument (default: foot)

[Arg0 John<sub>i</sub>] tried [Arg0 \*trace\*<sub>i</sub>] to kick [Arg1 the football].

- Ein mehrdeutiges Verb kann mehr als eine Frame-Menge haben

- Frameset **decline<sub>1</sub>** 'go down incrementally'

Arg0: entity going down

Arg1: amount gone down by EXT

Arg2: start point

Arg3: end point

... [*Arg1* its income] declined [*Arg2-EXT* 42%] [*Arg4* to \$2,420].

- Frameset **decline<sub>2</sub>** 'demure, reject'

Arg0: agent

Arg1: thing rejected

[*Arg0* A spokesman<sub>*i*</sub>] declined [*Arg1* \*trace\*<sub>*i*</sub> to elaborate].

- Alternationen, die die Verb-Bedeutung erhalten, beziehen sich auf eine gemeinsame Frame-Menge
- Frameset **open<sub>1</sub>** 'cause to open'
  - Arg0: agent
  - Arg1: thing opened
  - Arg2: instrument

[Arg0 John] opened [Arg1 the door].

[Arg1 The door] opened.

[Arg0 John] opened [Arg1 the door] [Arg2 with his foot].

# PropBank vs. FrameNet

- Gemeinsames Ziel: Dokumentation der syntaktischen Realisierungen von Argumenten durch Annotation von semantischen Rollen
- Unterschiedliche Vorgehensweisen:
  - FrameNet geht Frame-gesteuert vor.
  - PropBank annotiert ein vollständiges Korpus, die Penn Treebank.
  - PropBank legt weniger Wert auf die Semantik der Klassen, mit denen Verben assoziiert werden.
  - Die PropBank- und die FrameNet-Rollen korrespondieren nicht notwendigerweise miteinander.

- Ausnutzen der PropBank-Definitionen für Verben in Bezug auf ihre Übertragbarkeit zu Nomen, z.B. *decide* → *decision*
- Fokus: die meisten Argument-subkategorisierenden Nomen sind Nominalisierungen
- Abdeckung:
  - Verben: *decision, helper, nominee* ∼ PropBank
  - Adjektive: *incompetence, ability, wisdom* ∼ COMLEX-Syntax plus Heuristiken, dann manuell korrigiert
  - Relationale Nomen: *Präsident, Freund, Vater* ∼ Annahme: eine Rollen-Menge beschreibt alle Elemente einer Nomenklasse
  - Partitive Nomen: *barrage, clump, variety* etc.

- Frameset **destruction**

Arg0: agent of destruction

Arg1: patient of destruction

[Arg0 Richard]'s destruction of [Arg1 the secret tape]

- Frameset **anniversary**

Arg0: agent

Arg1: thing remembered

Arg2: times celebrated

[Arg0 Investors] celebrated the [Arg2 second] anniversary  
of [Arg1 Black Monday].

# Assoziationsnormen

- **Assoziation:** (Psychologie) Bewusstseinsverknüpfung von zwei oder mehreren Vorstellungsaspekten (Bußmann, 1990)
- **Assoziation:** (Psycholinguistik) Verknüpfung zwischen Reiz (Stimulus) und Reaktion (Response)
- Assoziationsnorm: quantifizierte Sammlung von Assoziationen
- Beispiele:
  - Assoziationen zu deutschen Nomen
  - Assoziationen zu deutschen Verben
  - University of South Florida Free Association Norms
  - Edinburgh Word Association Thesaurus

# Ressourcen zur Wortbedeutung: Zusammenfassung

- Manuelle und (semi-)automatische Ressourcen
- Verschiedene Ressourcen stellen unterschiedliche Perspektiven bereit
- Fokus: Ressourcen, die durch Computer genutzt werden können
- Wort-Einträge unterscheiden Wortbedeutungen
- Übermacht an englischen Ressourcen;  
andere Sprachen verlinken ihre Ressourcen teilweise

-  Stefan Engelberg and Lothar Lemnitzer.  
*Lexikographie und Wörterbuchbenutzung*, volume 14 of *Stauffenburg Einführungen*.  
Stauffenburg-Verlag, Tübingen, 2 edition, 2004.
-  Diego Marconi.  
*Lexical Competence*.  
MIT Press, Cambridge, MA, 1997.
-  Kai Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde, and Hagen Langer, editors.  
*Computerlinguistik und Sprachtechnologie – Eine Einführung*.  
Spektrum Akademischer Verlag, Heidelberg, 2nd edition, 2004.

# Referenzen: WordNet



Christiane Fellbaum, editor.

*WordNet – An Electronic Lexical Database.*

Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.



Lothar Lemnitzer and Claudia Kunze.

*Computerlexikographie*, chapter 6.

Gunter Narr Verlag, Tübingen, Germany, 2007.



George A. Miller, editor.

*WordNet: An On-line Lexical Database*, volume 3 (4).

Oxford University Press, 1990.

Special Issue of the International Journal of Lexicography.



Piek Vossen.

Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index.

*International Journal of Lexicography*, 17(2):161–173, 2004.



Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck.

Background to FrameNet.

*International Journal of Lexicography*, 16:235–250, 2003.



Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal.  
Towards a Resource for Lexical Semantics: A Large German Corpus  
with Extensive Semantic Annotation.

In *Proceedings of the 41st Annual Meeting of the Association for  
Computational Linguistics*, pages 537–544, Sapporo, Japan, 2003.



Martha Palmer, Daniel Gildea, and Paul Kingsbury.  
The Proposition Bank: An annotated Resource of Semantic Roles.  
*Computational Linguistics*, 31(1):71–106, 2005.

-  Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman.  
The NomBank Project: An Interim Report.  
*In Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation, Boston, MA, 2004.*
-  Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman.  
Annotating Noun Argument Structure for NomBank.  
*In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.*

# Referenzen: Assoziationen



James Deese.

*The Structure of Associations in Language and Thought.*

The John Hopkins Press, Baltimore, MD, 1965.



Herbert H. Clark.

Word Associations and Linguistic Theory.

In John Lyons, editor, *New Horizon in Linguistics*, chapter 15, pages 271–286. Penguin, 1971.



Sabine Schulte im Walde, Alissa Melinger, Michael Roth, and Andrea Weber.

An Empirical Characterisation of Response Types in German Association Norms.

*Research on Language and Computation*, 6(2):205–238, 2008.

DOI 10.1007/s11168-008-9048-4.