EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT
FACHBEREICH INFORMATIK

Lexikalische Semantik: WordNet

Skript, Version 2019-11-10

Timm Lichte

Outline

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT
FACHBEREICH INFORMATIK

Lexikalische Semantik: WordNet

Skript, Version 2019-11-10

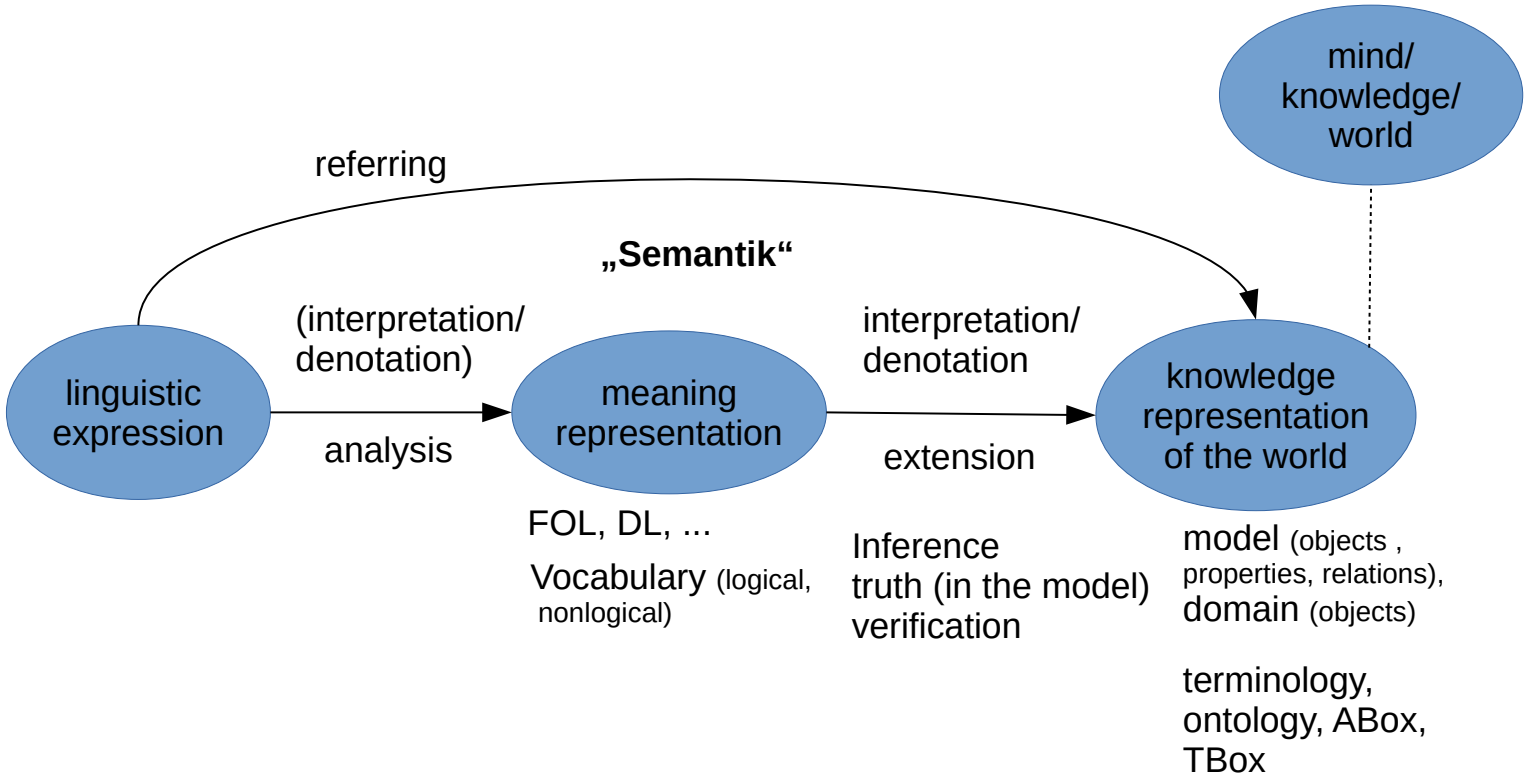
Timm Lichte

Intro

Die **SEMANTIK** beschäftigt sich mit der Bedeutung von Symbolen bzw. Zeichen (Symbole, die eine Bedeutung haben). Sind diese Symbole linguistische Wortformen wie ***Bank***, dann haben wir es genauer mit der **LEXIKALISCHEN SEMANTIK** zu tun.

Vielleicht die wichtigste Ressource im Bereich der lexikalischen Semantik ist **WORDNET**. Dank der relativ einfachen, theorieneutralen Grundprinzipien genießt WordNet auch nach 30 Jahren immer noch einen Sonderstatus in der NLP.

Semantik allgemein I



Semantik/Bedeutung stellt eine Verbindung her zwischen einem Zeichen und der “Welt”.

Unterstützende Literatur

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Miller, George A. 1995. WordNet: A lexical database for English. Communications of the ACM 38(11). 39–41.
- Jurafsky, Daniel & James H. Martin. 2018. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third edition draft of September 23, 2018. `https://web.stanford.edu/` jurafsky/slp3/ed3book.pdf.

Wird jeweils angezeigt!

Fakten und Zahlen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

WordNet:

- entwickelt seit 1986 unter der Leitung von George A. Miller (†)
- entstanden im Bereich der Künstlichen Intelligenz und Psychologie
- Umfang Version 3.1 von 2012 (Quelle Wikipedia):
  - 155 327 Lemmata
  - 175 979 Synsets (~Bedeutungen)
  - 207 016 Lemma-Synset-Paare
  - durchschn. Lemma-Ambiguität: 1,34
  - durchschn. Synset-Umfang: 1,176

Knoten = Synsets

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Synset

Each member of a given synset expresses the same concept, though not all synset members are interchangeable in all contexts.[2]

Tests für (Fast-)Synonymie (“near synonymy”):

- “interchangeability in some contexts”

- ***Mein Wagen / Auto ist kaputt.***
- ***Ich brauche eine Münze für einen Wagen /#ein Auto .***

Synset-Attribute

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Jedes Synset hat die folgenden Attribute:

- Synset-Label `hello.n.01`
- Lemmata ***hello***, ***hullo***, ***hi***, ***howdy***
- Definition/Glosse ‘an expression of greeting’
- Examples ***every morning they exchanged polite hellos***
- Frequenz (in SemCor)

Kanten = semantische Relationen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Semantische Relationen bestehen zwischen **nominalen** Synsets:

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} $\rightarrow$ *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} $\rightarrow$ *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} $\rightarrow$ *author*^{1}
Instance Hyponym	Has-Instance	From concepts to concept instances	*composer*^{1} $\rightarrow$ *Bach*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} $\rightarrow$ *professor*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} $\rightarrow$ *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} $\rightarrow$ *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} $\rightarrow$ *meal*^{1}
Substance Meronym		From substances to their subparts	*water*^{1} $\rightarrow$ *oxygen*^{1}
Substance Holonym		From parts of substances to wholes	*gin*^{1} $\rightarrow$ *martini*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} $\Longleftrightarrow$ *follower*^{1}
Derivationally Related Form		Lemmas w/same morphological root	*destruction*^{1} $\Longleftrightarrow$ *destroy*

Figure C.2 Noun relations in WordNet.

(aus Jurafsky & Martin [3])

Beispiel: *deceiver*

(aus Fellbaum [2])

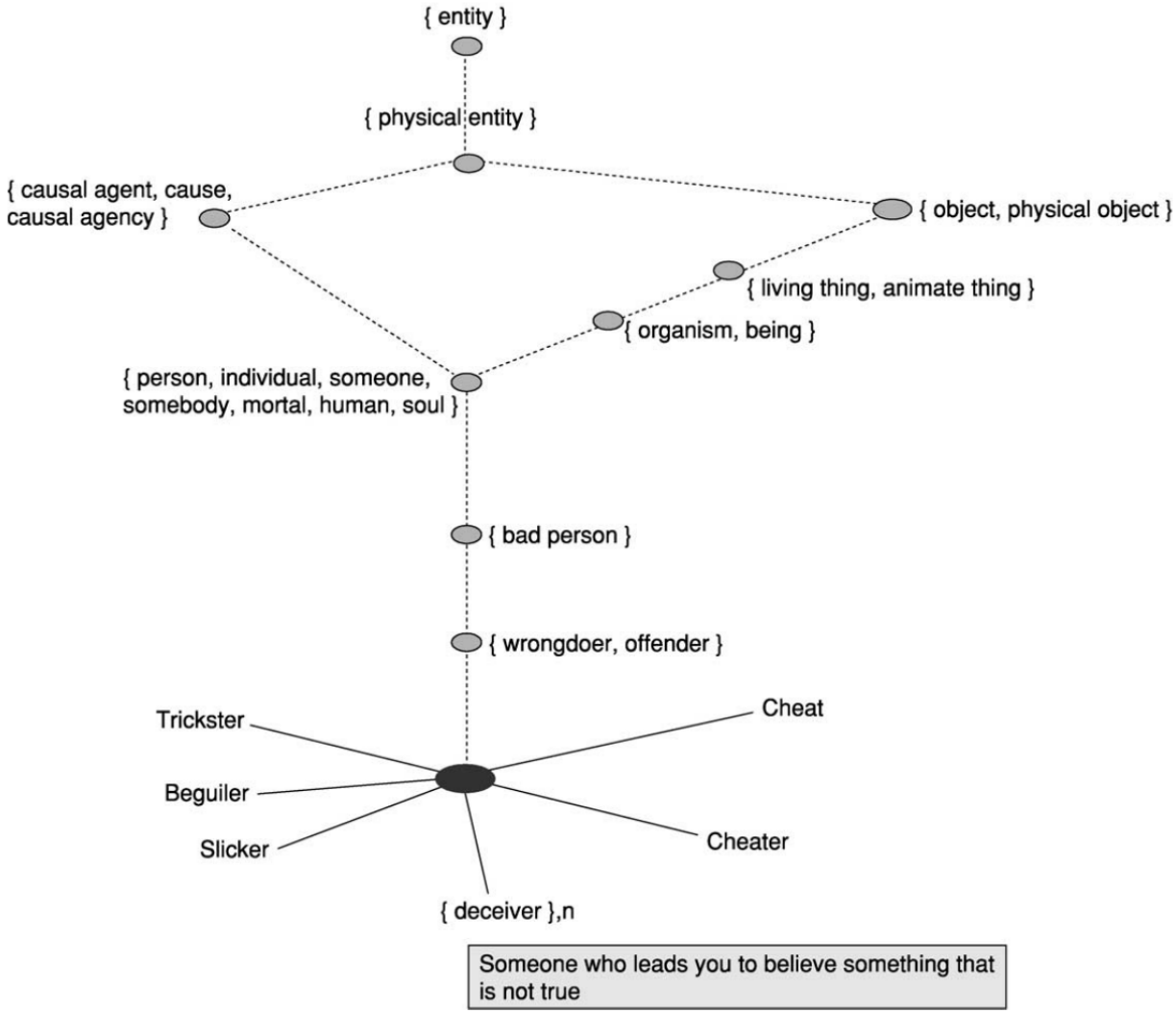


Figure 1 A WordNet noun tree.

Kanten = semantische Relationen

Semantische Relationen bestehen zwischen **verbalen** Synsets:

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From events to subordinate event (often via specific manner)	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Semantic opposition between lemmas	$increase^1 \Longleftrightarrow decrease^1$
Derivationally Related Form	Lemmas with same morphological root	$destroy^1 \Longleftrightarrow destruction^1$

Figure C.3 Verb relations in WordNet.

(aus Jurafsky & Martin [3])

Beispiel: *deceive*

(aus Fellbaum [2])

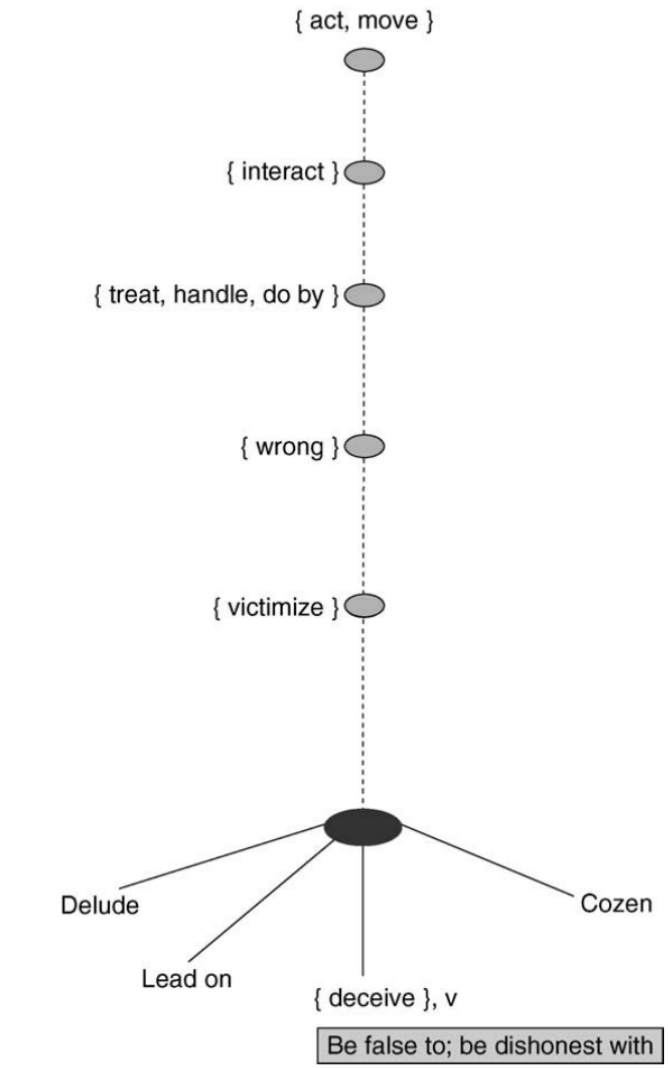


Figure 2 A WordNet verb tree.

Beispiel: *dry* versus *wet*

(aus Fellbaum [2])

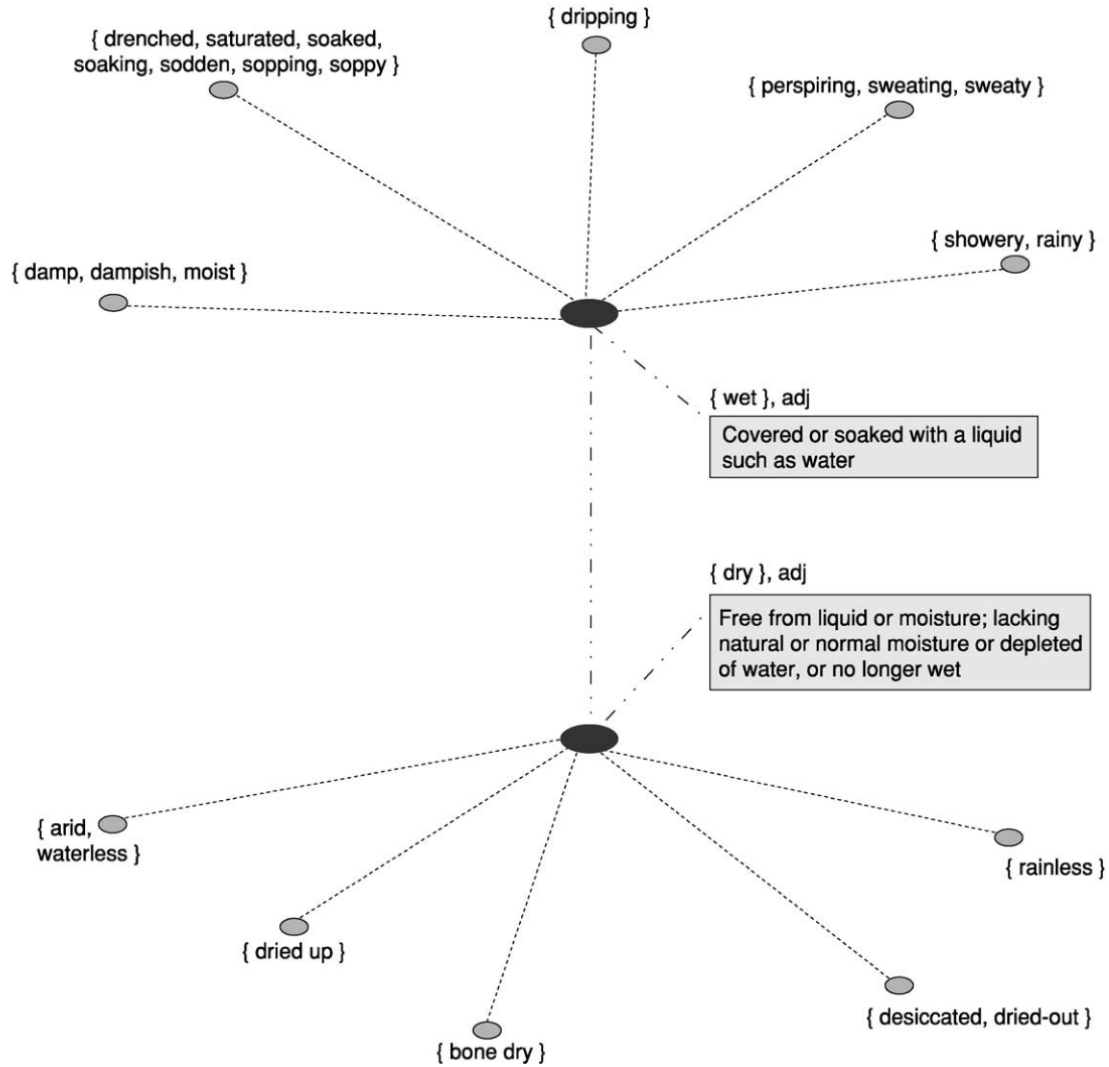


Figure 3 An adjective cluster in WordNet.

Globale Struktur

Unterschiedliche Dateien je nach syntatkischer Kategorie (Zahlen aus Jurafsky & Martin [3]):

- Nomen: 117 798 (1,23 Bedeutungen/Lemma)
- Verben: 11 529 (2,16 Bedeutungen/Lemma)
- Adjektive: 22 479
- Adverbien: 4 481

Unique Beginners

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Die WordNet-Hierarchie für Nomen kann in mehrere Hierarchien partitioniert werden, die “relatively distinct semantic fields representieren”

Die Wurzel (“top”) einer solchen Teilhierarchy wird auch “unique beginner” genannt (aus Fellbaum [1: §1.2]):

Table 1.1
List of 25 unique beginners for noun source files

{act, activity}	{food}	{possession}
{animal, fauna}	{group, grouping}	{process}
{artifact}	{location}	{quantity, amount}
{attribute}	{motivation, motive}	{relation}
{body}	{natural object}	{shape}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	{person, human being}	{substance}
{event, happening}	{plant, flora}	{time}
{feeling, emotion}		

Unique Beginners – Taxonomic Sisters – Semantic Atoms

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Das entspricht in gewisser Weise den “**taxonomic sisters**” bei Kroeger (ohne “mutually exclusive” zu sein), oder den **semantischen Atomen** in der Komponentenanalyse.

Die Herausforderung ist offensichtlich:

> *The problem, of course, is to **decide what the primitive semantic components should be**. Different workers make different choices; one important criterion is that, collectively, they should provide a place for every English noun. (Miller in Fellbaum [1])*

Gruppierungen von Unique Beginners

(aus Fellbaum [1])

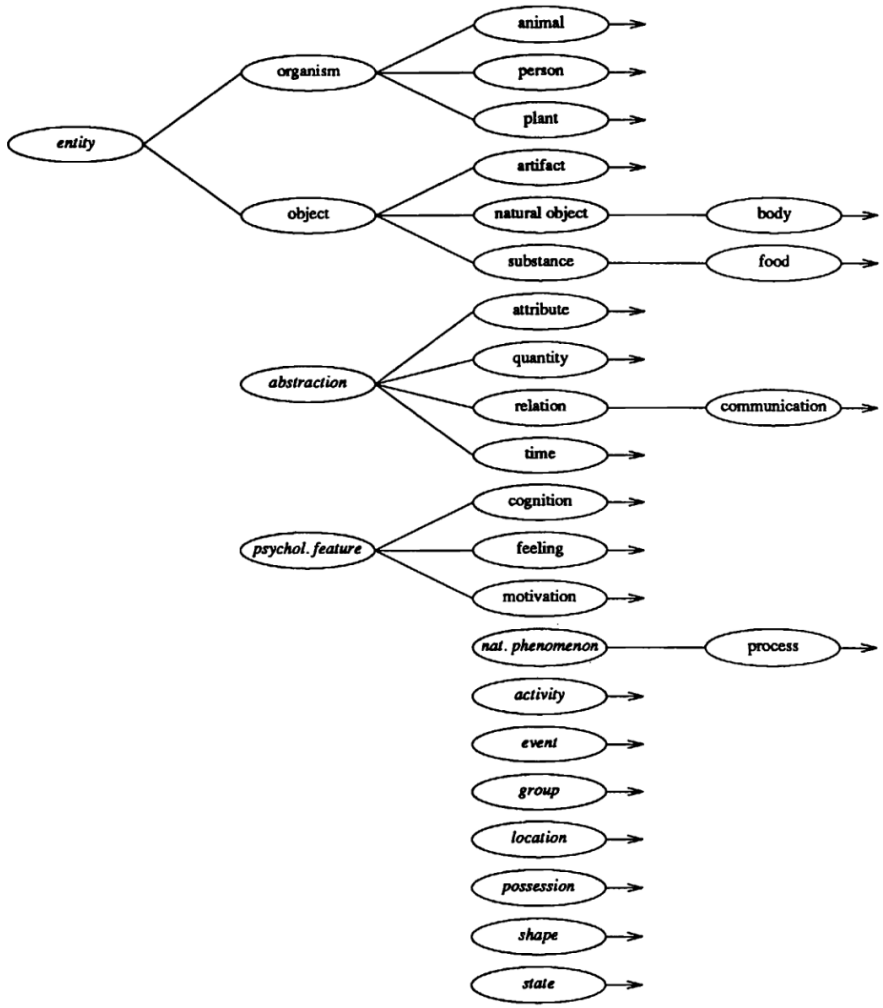


Figure 1.1
Diagrammatic representation of relations that reduce the 25 noun source files to 11 unique beginners. The unique beginners are italicized.

Tiefe

Fellbaum [1: §1.2]:

- seldom more than **10 or 12 levels deep**
- The deepest examples usually contain technical distinctions that are not part of the everyday vocabulary.

Visualisierung

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

http://wordventure.eti.pg.gda.pl/wne/wne.html

Erstellung von WordNet

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

ganz viel Handarbeit . . . .

$\Rightarrow$ teuer und zeitaufwendig, aber hohe Qualität

Bemühungen, den Erstellungsprozess so weit wie möglich zu automatisieren.

Lücken in WordNet

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Miller in Fellbaum [1: §1.4]

- `is-not-a`-Relationen / Ausnahmen:
  - ***A whale is not a fish.***
- Unterschiedliche Arten von Hyponymie:
  - **taxonomic** (`is-a-kind-of`)
    functional (`is-used-as-a-kind-of`)
  - {***chicken***} @→ {***bird***} versus {***chicken***} @→ {***food***}
- Unterscheidung zwischen Eigennamen und Gattungsname, Maßnomen und zählbaren Nomen
- andere semantische Relationen
  - implizit verfügbar über Glossen

Lücken in WordNet

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Fellbaum [2]:

- syntagmatische Relationen:
  - thematische Rollen (AGENS, PATIENS, ...)
  - grammatische Funktionen (Subjekt, Objekt, ...)

The Tennis Problem[1]

- Semantische Felder wie ***racquet***, ***ball***, and ***net*** werden nicht direkt repräsentiert.
- ad-hoc/diskursgetriggerte Relationen: ***racquet***, ***ball***, ***net***, ***court game***, ***physician***, ***hospital*** ...

Aber vielleicht indirekt? $\longrightarrow$ Relationspfade

Relationspfade und Ähnlichkeit

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Intuition

Synsets sind konzeptuell um so ähnlicher, je näher sie in WordNet beieinander liegen, d.h. je **kürzer** der Relationspfad zwischen ihnen ist

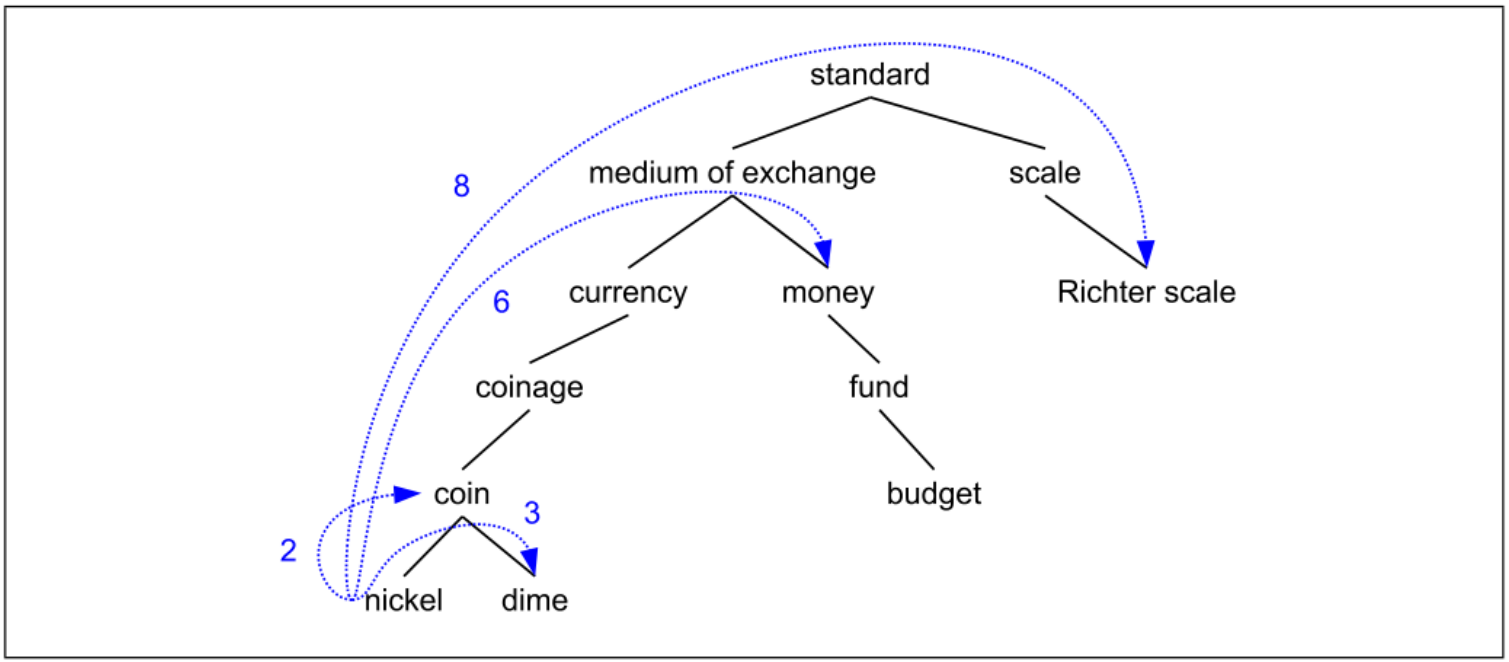


Figure C.5 A fragment of the WordNet hypernym hierarchy, showing path lengths (number of edges plus 1) from *nickel* to *coin* (2), *dime* (3), *money* (6), and *Richter scale* (8).

Pfadlänge und Wortähnlichkeit

Das lässt sehr einfach auf Worte/Lemmata übertragen:

- $pathlen(c_1, c_2) = 1+$ the number of edges in the shortest path between the sense nodes c_1 and c_2
- $sim_{path}(c_1, c_2) = \frac{1}{pathlen(c_1, c_2)}$
- $wordsim(w_1, w_2) = \max_{c_1 \in senses(w_1), c_2 \in senses(w_2)} sim(c_1, c_2)$

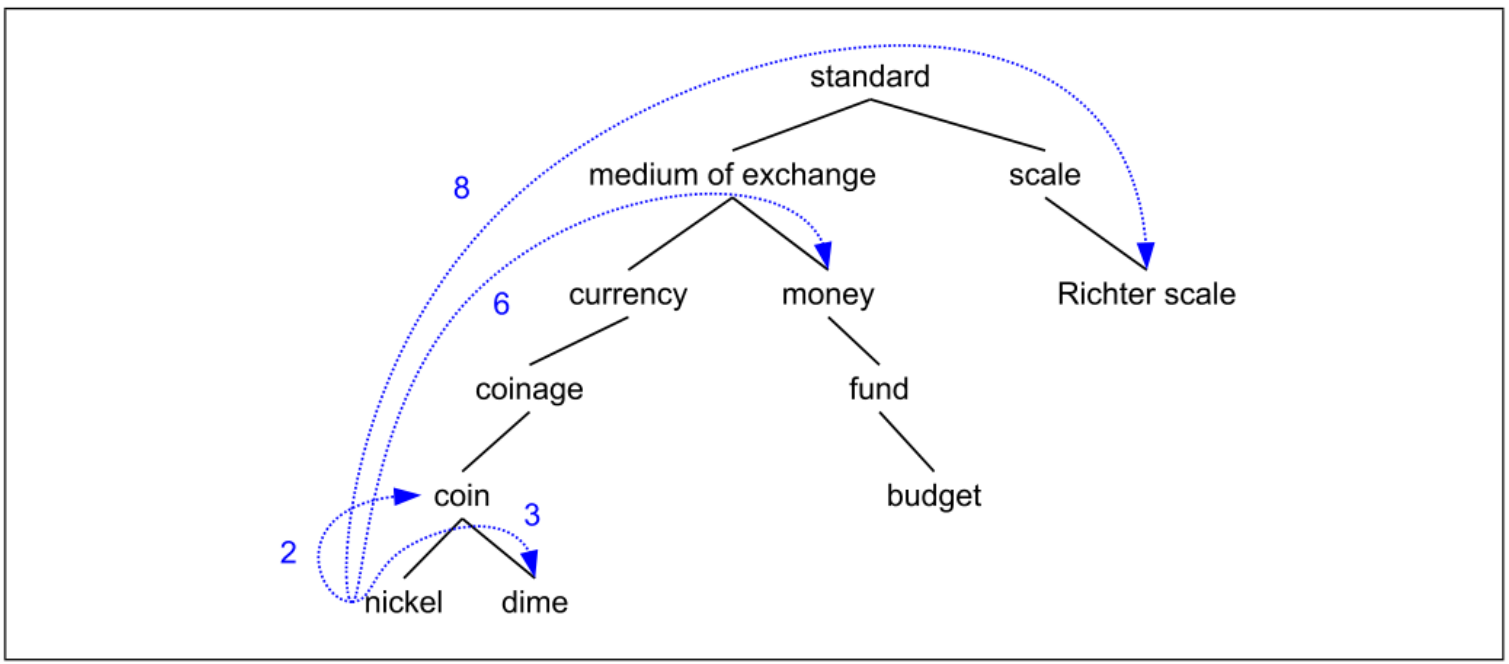


Figure C.5 A fragment of the WordNet hypernym hierarchy, showing path lengths (number of edges plus 1) from *nickel* to *coin* (2), *dime* (3), *money* (6), and *Richter scale* (8).

Gewichtung von Relationspfaden

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Problem

Die Ähnlichkeit zwischen ***nickel*** und ***money*** ist intuitiv viel größer als die zwischen ***nickel*** und ***standard***.

Lösung: unterschiedliche Gewichtung von Relationspfaden

- abhängig von der Tiefe der Einbettung (Wu & Palmer)
- abhängig von der Auftretenshäufigkeit in einem Korpus

Resnik/Information-based Word Similarity:

- $P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$
- $IC(c) = -\log P(c)$
- $LCS(c_1, c_2) =$ the lowest common subsumer
- $sim_{resnik}(c_1, c_2) = -\log P(LCS(c_1, c_2))$

Anwendung: WSD

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mit solchen Ähnlichkeitsmaßen für Bedeutungen in WordNet lassen WSD-Verfahren umsetzen.

Idee:

- Gegeben ein Target-Lemma in einem Satz mit möglichen Bedeutungen $\sigma_1, ..., \sigma_n$.
- Man wähle die Bedeutung σ_i so, dass die Ähnlichkeit zu den Bedeuntungen der anderen Lemmata im Satz maximiert wird.
- ***Ich ging zur Bank, um Geld abzuheben.***

Andere Sprachen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

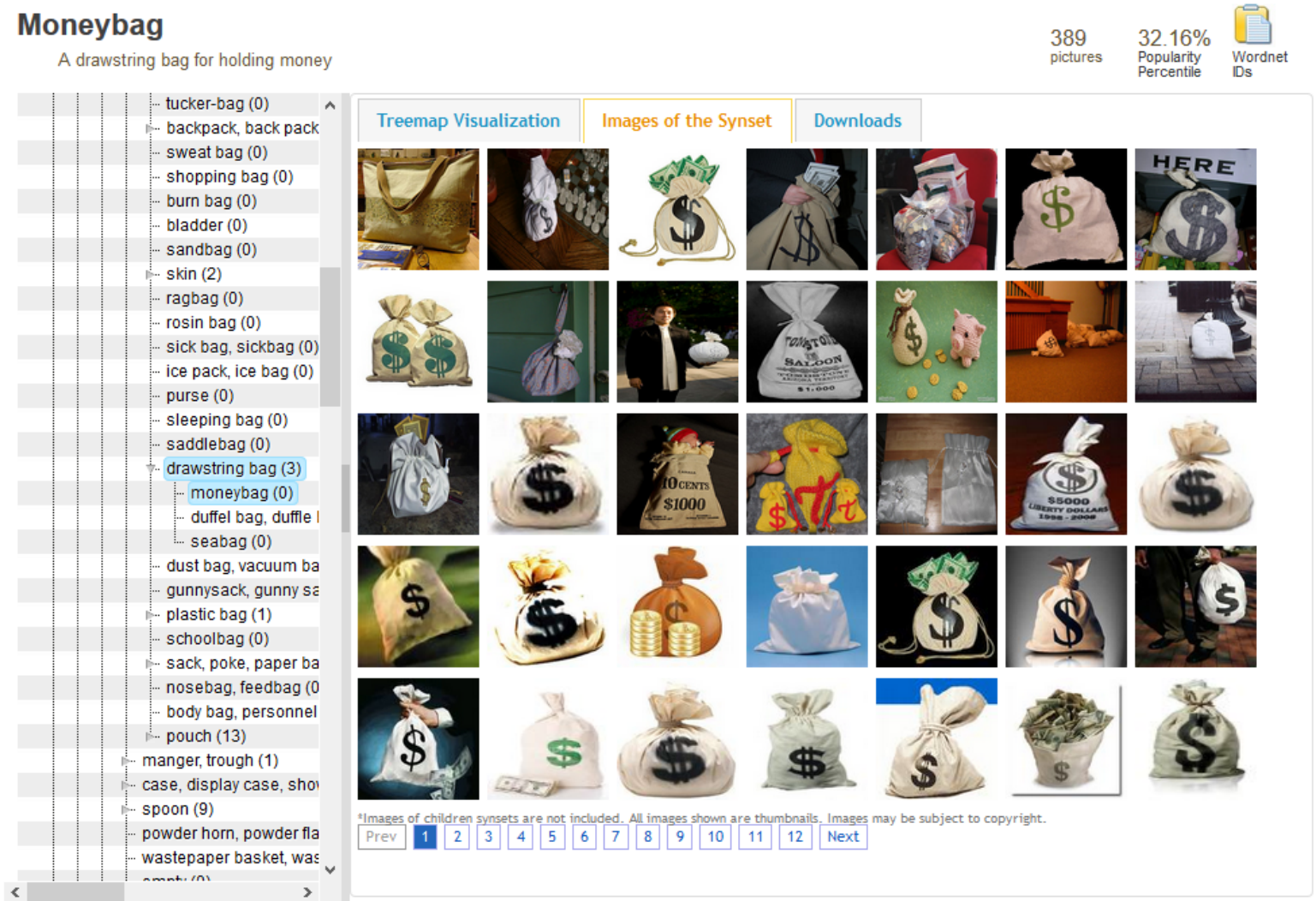
- Deutsch: GermaNet (in Tübingen)
- Polnisch
- Japanisch
- Hindu
- ...

`http://globalwordnet.org/wordnets-in-the-world/`

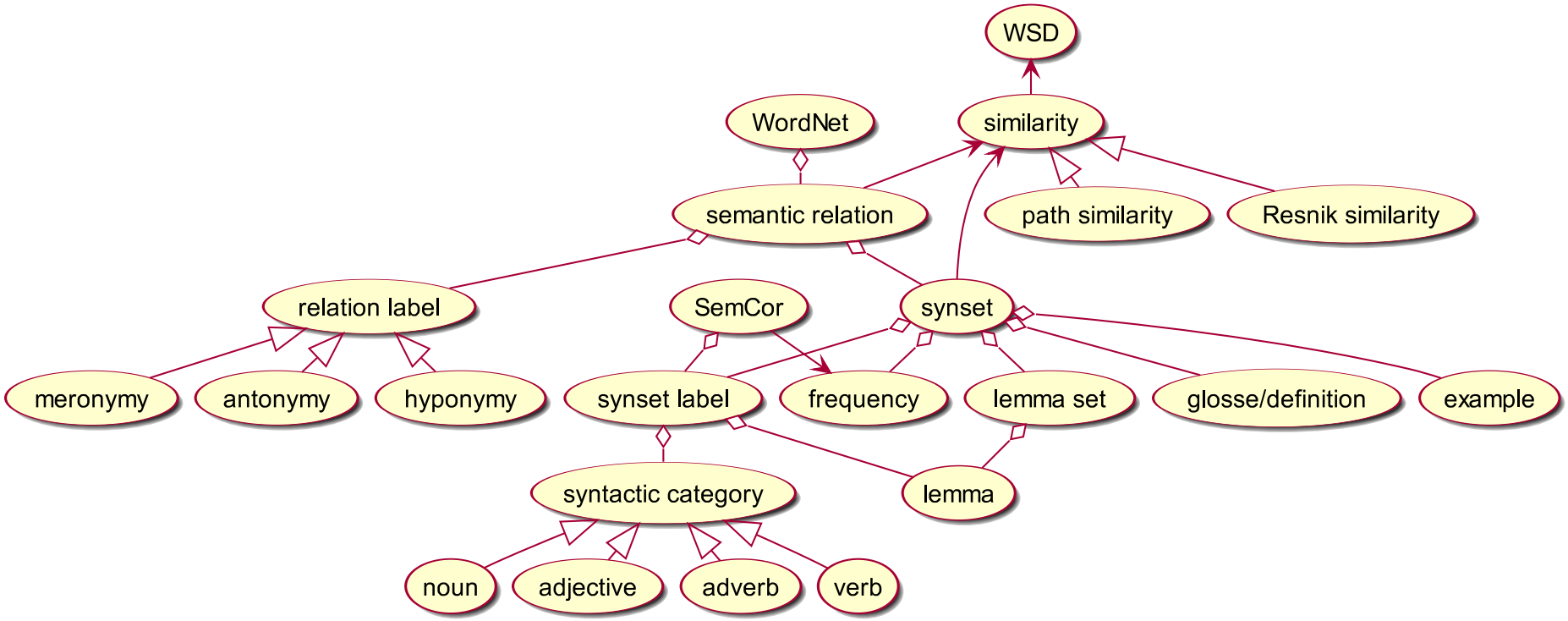
Multimodale Erweiterungen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

IMAGENET: Photodatenbank mit Verknüpfungen zu WordNet-Bedeutungen



Begriffsdiagramm



Literaturangaben I

[1] Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. (Language, Speech, and Communication). Cambridge, MA: MIT Press.

[2] Fellbaum, Christiane. 2006. WordNet(s). *Encyclopedia of Language & Linguistics*. 665–670.

[3] Jurafsky, Daniel & James H. Martin. 2018. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Third edition draft of September 23, 2018. `https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf`.

[4] Miller, George A. 1995. WordNet: A lexical database for english. *Communications of the ACM* 38(11). 39–41.