Evaluating the Relationships Instantiated by Semantic Associates of Verbs

Sabine Schulte im Walde & Alissa Melinger

Universität des Saarlandes

July 12, 2005

Semantic Associates

- Semantic associates are words that are called to mind by another word.
- Semantic relatedness is a measure of the likelihood that one word will be called to mind by another word.
- Semantic associates have been used for the last 30+ years to investigate semantic memory.
- Implicit notion that associates reflect meaning components of words.

Recent Trends

- There has been a move away from associates:
  - Concern about what types of semantic information semantic associates reflect.
  - Some evidence that associates may reflect surface co-occurrence patterns more than semantic memory.
  - Heightened interest in semantic features.
- If semantic associates reflect semantic memory, then they could be a valuable tool, if we understood more about them.
  Characterisation of verb and noun responses,
  To evaluate the co-occurrence assumption.

Motivation for Computational Linguistics

- NLP tasks vary in usage of semantic relations, e.g. thesaurus extraction, summarisation
  ⇒ Identification of semantic verb relations
- Data-intensive lexical semantics:
  words ⇔ distributional vectors,
  relatedness of words ⇔ vector similarity
  ⇒ Identification of nominal features to describe verbs

Elicitation Material

- 330 German verbs
- Variety of semantic verb classes, possible ambiguity:
  - self-motion: *gehen* 'walk', *schwimmen* 'swim'
  - cause: *verbrennen* 'burn', *reduzieren* 'reduce'
  - experiencing: *lachen* 'laugh', *überraschen* 'surprise'
  - communication: *erzählen* 'tell', *klagen* 'complain'
  - body: *schlafen* 'sleep', *abnehmen* 'lose weight'
- Variety of frequency ranges (1 < freq < 71,604)
- Random distribution: 6 data sets à 55 verbs, balanced for class affiliation and frequency ranges

Elicitation Procedure

- Web elicitation over Internet
- Bibliographic information:
  linguistic experience, age, regional accent, profession
- Instructions and example page
- Web page for each verb
- Association input:
  spontaneous, exhaustive, one word per line, capitalisation
- 30 sec. for each verb; 2 sec. break; total: ca. 30 min.

schneien

- kalt
- rodeln
- Schneemann
- weiß
- dämmern
-
-

Participants and Data

- 299 accepted data files:
  native German speakers, 80% of target verbs
- Expertise of participants: 166 experts vs. 132 non-experts
- Participants per data set: between 44 and 54
- Number of trials: 16,445
- Number of associations per target verb:
  range 0-16, average: 5.16
- All associations: 81,373 tokens for 18,884 types
  (first) 15,780 tokens for 4,856 types

Data Preparation

1. Lexicon look-up
2. (Semi-automatic) Data correction
3. Quantification over responses

klagen 'complain, moan'		
Gericht	'court'	19
jammern	'moan'	18
weinen	'cry'	13
Anwalt	'lawyer'	11
Richter	'judge'	9
Klage	'complaint'	7
Leid	'suffering'	6
Trauer	'mourning'	6
Klagemauer	'Wailing Wall'	5
laut	'noisy'	5

Linguistic Analyses

- Preference for morpho-syntactic category of responses?
  → distinguish major parts-of-speech:
  nouns, verbs, adjectivs, adverbs
- Encoding of particular semantic relations?
  → look up relation between target and response verb:
  GermaNet (Kunze, 2000/2004)
- Typical argument holders of verb valency?
  → investigate linguistic functions realised by nouns:
  empirical grammar model (Schulte im Walde, 2003)

Excursus: Statistical Grammar Model

- Lexicalised probabilistic context-free grammar
  (Charniak, 1995; Carroll and Rooth, 1998)
- 35 million words of German newspaper corpora
- Unsupervised training by *EM-Algorithm* (Baum, 1972)
- Robust statistical parser *LoPar* (Schmid, 2000)
- Corpus-based quantitative lexical information:
  word frequencies, linguistic functions, head-head relations

Morpho-Syntactic Analysis of Responses

- Source: machine-readable quantitative dictionary
- Dictionary information:
  word forms, part-of-speech tags, lemmas, frequencies
- Ambiguous part-of-speech tags;
  examples: *Rauchen* 'smoke' (V/N)
  überlegen `think about/superior´ (V/ADJ)
- Result: distinction and quantification of
  morpho-syntactic categories of responses

Morpho-Syntactic Distribution

	V	N	ADJ	ADV	
Freq	19.863	48.905	8.510	1.268	TOKEN
Prob	25	62	11	2	all assoc

Morpho-Syntactic Correlations/Tests

- Correlation +
  target verb frequency ⇔ verb/adverb responses
- Correlation -
  target verb frequency ⇔ noun responses
- Variation across verb classes

Morpho-Syntactic Distribution Examples

	V	N	ADJ	ADV
Total Prob	25	62	11	2
aufhören 'stop'	49	39	4	6
aufregen 'be upset'	22	54	21	0
backen 'bake'	7	86	6	1
bemerken 'realise'	52	31	12	2
dünken 'seem'	46	30	18	1
flüstern 'whisper'	19	43	37	0
nehmen 'take'	60	31	3	2
radeln 'bike'	8	84	6	2
schreiben 'write'	14	81	4	1

Semantic Relations between Verbs

- Semantic relations between target and response verbs
- Source: lexical semantic taxonomy GermaNet (GWN)
- Organisation of verbs, nouns, adjectives, adverbs
- Classes of synonyms: synsets
- Ambiguous words: assignment to multiple classes
- Lexical and conceptual relations between synsets:
  antonymy, hypernymy, entailment, cause, etc.

Semantic Relations between Verbs

- Synonymy: target and response verb in common synset
- Other semantic relations:
  look up GermaNet semantic relations between
  - target verb synsets
  - response verb synsets
- Quantification of target-response relation:
  association frequency
- No relation: target and response verb in GWN, no relation
- Unknown relation: response verb not in GWN

Semantic Relations: Distributions

	SYN	ANT	HYP	HYPO	CAUSE	ENTAIL	ALSO	NONE	?	
Freq	1,194	252	2,807	3,016	49	0	0	10,509	1,726	TO
Prob	6	1	14	16	0	0	0	54	9	all

Semantic Relations: Correlations/Tests

- Correlation +
  target verb frequency ⇔ synonym, antonym, hyponym
- Variation across verb classes (example):
  significant differences in antonyms
  » aspectual verb *aufhören* 'stop'
  ⇒ *anfangen* 'begin', *weitermachen* 'go on'
  » creation verb *backen* 'bake'

No Semantic Relation in GermaNet

54%

- *schwitzen 'sweat'*
  stinken 'stink' (8), *laufen* 'run' (5), *frieren* 'be cold' (2)
- *adressieren* 'address'
  schreiben 'write' (15), *schicken* 'send' (6)
- *backen* 'bake'
  essen 'eat' (6), *schmecken* 'taste' (2), *kneten* 'knead' (2)
- *setzen* 'seat'
  legen 'lay' (8), *stellen* 'place' (7), *ausruhen* 'rest' (3)

Window Approach for Semantic Relations

- Corpus data from statistical grammar model
- Window (left+right): 5/20/50 words, excluding symbols
- Basis: association tokens

window	*pos (37%)*	*neg (63%)*	*all*
5	*57*	*29*	*39*
20	*75*	*46*	*57*
50	*82*	*54*	*64*

Window Approach for GWN Relations

Relation		window: 5		window: 20		window: 50	
Synonymy	1,194	652	55%	923	77%	1,018	85%
Antonymy	252	200	79%	209	83%	213	85%
Hypernymy	2,807	1,564	56%	2,103	75%	2,254	80%
Hyponymy	3,016	1,695	56%	2,232	74%	2,463	82%
Cause	49	39	80%	41	84%	42	86%
Entailment	0	0	0%	0	0%	0	0%
Also see	0	0	0%	0	0%	0	0%

Missing Semantic Relations

» tagging:
auftauen 'defreeze' - *wärme* 'warmth' (1)

» missing links in GermaNet:
erhitzen 'heat' - *abkühlen* 'cool' (2)

» domain:
paddeln 'paddel' - *rudern* 'row' (22)

» non-classical semantic relation/scene information:
initiieren 'initiate' - *anfangen* 'start' (21)
auftauen 'defrost' - *essen* 'eat' (8)
paddeln 'paddle' - *schwimmen* 'swim' (7)

Semantic Relations: Summary

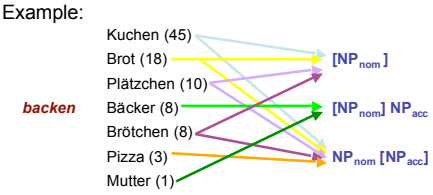
- Characterisation of verb-verb associations
- Proportion and distinction of 'classical' verb relations
- Location of verb-verb co-occurrences in corpus window
- Detection of missing links in GermaNet
- Non-classical verb relations as major part of associations
- Associations as basis for defining non-classical relations

Syntax-Semantic Functions of Nouns

- Source: statistical grammar model
- Verb valency:
  - 38 syntactic subcategorisation frames
  - plus PP information (case+preposition) → 178 frames
  - subcategorised nouns
- Example: *backen* 'bake'
  - frames: NP_{nom}
    NP_{nom} NP_{acc} …
  - filler examples for NP_{nom} $[NP_{acc}]$: *Brötchen* 'rolls'
    Brot 'bread'
    Kuchen 'cake' …

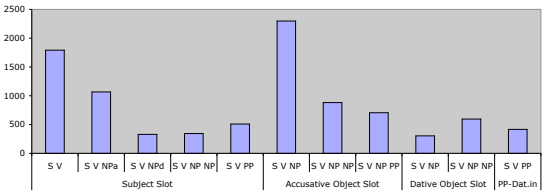
Syntax-Semantic Functions: Analysis

- Typical conceptual roles which speakers have in mind
- Look up syntactic relationships between verb and nouns
- Example:



Syntax-Semantic Frame Inspection

28%



Syntax-Semantic Functions: Tests

- Association of verb classes and frame slots:

  significant differences in direct objects

  » creation verb *backen* 'bake'

  » aspectual verb *aufhören* 'stop'

No Linguistic Function in Grammar

72%

- *backen* 'bake'
  Ofen 'oven' (19), *Mehl* 'flour' (17), *Weihnachten* 'Xmas' (15)
- *fliegen* 'fly'
  Urlaub 'vacation' (11), *Flügel* 'wings' (9)
- *anfangen* 'begin'
  Start 'start' (14), *Motivation* 'motivation' (3)
- *enden* 'end'
  Feierabend 'leisure-time' (4), *Rente* 'retirement' (2)

Window Approach for Linguistic Functions

- Corpus data from statistical grammar model
- Window (left+right): 5/20/50 words, excluding symbols
- Basis: association tokens

window	*pos (28%)*	*neg (72%)*	*all*
5	*87*	*19*	*39*
20	*95*	*37*	*54*
50	*97*	*47*	*62*

Missing Linguistic Functions

» lemmatisation and tagging:
e.g. composita: *Übermacht, Zeitspanne, Autorennen*

» argument vs. adjunct:
erhitzen 'heat' - *Pfanne* 'pan' (2)

» domain:
radeln 'bike' - *Oma* 'grand-mom' (1)
stoppen 'stop' - *Plosiv* 'plosive' (1)

» scene information/world knowledge:
trocknen 'dry' - *Trockner* 'dryer' (11)
auftauen 'defrost' - *Wasser* 'water' (14)

Linguistic Functions: Summary

- Characterisation of verb-noun associations
- Properties of nouns represent conceptual roles of verbs
- Scene information in addition to subcategorisation
- Co-occurrence counts to supplement argument counts
- Usage of roles and window-based nouns for distributional verb descriptions

Summary and Discussion

• This work examined the functional, ontological and co-occurrence relations of a large set of stimulus-response pairs.
  • Characterize associates: noun responses largely reflected non-argument concepts, and verb responses often reflected non-classical semantic relations.
  • Evaluate co-occurrence assumptions: 67% of first responses occurred in the co-occurrence window of ±20 words, 51% occurred in the co-occurrence window of ± 5 words.

Summary and Discussion

• The analyses of noun and verb responses provides an objective characterization of associate relation types which can inform the interpretation of experimental results.

• Is the co-occurrence assumption still tenable, or do the analyses suggest that semantic memory may be tapped by associate elicitation? What is the appropriate baseline to use as an evaluation metric?

• Prior research focused on associates of noun targets. Do the above patterns generalize to nouns?

END