An Empirical Characterisation of Response Types in German Association Norms

Sabine Schulte im Walde
(Universität Stuttgart)

In collaboration with
Alissa Melinger
(University of Dundee)
Michael Roth, Andrea Weber
(Universität des Saarlandes)

Saarbrücken, July 12, 2007

Motivation

- Semantic associates are concepts spontaneously called to mind by a stimulus word
- Basis: collection of semantic associates evoked by German verbs and nouns
- Goal: empirical characterisation of verb and noun properties
- Assumption: semantic associates reflect highly salient linguistic and conceptual features of the stimulus word

Motivation

- Two issues related to word properties and word relations:
  1. Modelling word meaning by empirical features
  2. Definition of semantic relations between words/contexts

- Analyses:
  - » Motivation by potential NLP uses
  - » Exploration of relationships between stimuli and responses
  - » Basis: large-scale lexicographic databases and empirical, corpus-based resources

Distributional Word Meaning

- Data-intensive lexical semantics:
  empirically define and induce features that
  - » capture various word meaning aspects
  - » can be obtained automatically from corpus-data

  → similarity of words, sentences, paragraphs, etc.

  Examples: clustering, word sense discrimination, anaphora resolution, multi-word expressions, text indexing, etc.

- Distributional descriptions: contextual features, such as words co-occurring in a document, in a context window, or with respect to a word-word relationship, such as syntactic structure, syntactic and semantic valency, etc.

Distributional Word Meaning

- Little effort has been spent on investigating the eligibility of the various types of features

  Examples: Pereira, Tishby and Lee (1993) and Rooth et al. (1999) refer to a direct object noun for describing verbs; Curran (2003) to subjects and direct objects; Lin (1998) and McCarthy et al. (2003) used any dependency relation detected by the chunker or parser

- Assumption: semantic associates identify contextual functions for empirical feature descriptions

- Procedure: examine which functions are activated by associates and therefore contribute to salient meaning components of individual words and across words

Semantic Relations

- For many NLP resources and applications, it is crucial to define and use semantic relations between words or contexts.

  Examples: creation of lexical taxonomies (Fellbaum, 1998) and ontologies (Maedche and Staab, 2000; Navigli and Velardi, 2004; Kavalek and Svatek, 2005), thesaurus extraction (Lin, 1999; McCarthy et al., 2003), semantic lexicons used in e.g. information retrieval (Roark and Charniak, 1998; Riloff and Jones, 1999), question answering (Girju, 2003), summarisation (Barzilay et al., 2002), text understanding (Lapata, 2002; Beigman and Shamir, 2006)

- Limited work has been spent on specifying the range of relations.

Semantic Relations

- Assumption: semantic associates provide a means to investigate the range of semantic relations
- Procedure: analysis of semantic relations inter-categorical, i.e., verb-verb and noun-noun relations
- Assumption: examine types of relations that are captured by semantic associations, identified as important or salient by the speakers of the language

Overview

1. Data collection and preparation
2. Resources for data investigation
3. Linguistic analyses of experimental data
   (a) NLP motivation
   (b) analyses
   (c) interpretation

Data Collection and Preparation

Experiment Material: Verbs

- 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

schneien

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

Experiment Data: Verbs

- 299 accepted data files:
  native German speakers; threshold: 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: 79,480 tokens for 39,254 types

Data Preparation: Verbs

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

Experiment Material: Nouns

- 409 German nouns
- Depictable objects
- Variety of semantic categories:
  - » plants: *Rose* `rose´, *Baum* `tree´, *Zweig* `branch´
  - » professions: *Doktor* `doctor´, *Bäcker* `baker´
  - » instruments: *Klavier* `piano´, *Trommel* `drums´
  - » body parts: *Auge* `eye´, *Kopf* `head´, *Fuß* `foot´ ...
- Homophones: ca. 10% of the nouns
- Variety of frequency ranges according to CELEX

Experiment Procedure: Nouns

- 409 stimuli divided into 3 questionnaires
- Each set presented in two formats:
  with and without pictures
- 300 native German participants;
  50 participants for each questionnaire
- Maximum of three associates per stimulus
- No time limit
- Total number of responses: 116,714 Tokens
  31,035 Types

Modality: *word (+ picture)*

Witch

magic

wizard

broom

Data Preparation: Nouns

Schloss `lock´ (depicted), `castle´

Association		POS	PW	W
Schlüssel	‘key’	N	38	13
Tür	‘door’	N	10	5
Prinzessin	‘Princess’	N	0	8
Burg	‘castle’	N	0	8
sicher	‘safe’	ADJ	7	0
Fahrrad	‘bike’	N	7	0
schließen	‘close’	V	6	1
Keller	‘cellar’	N	7	0
König	‘king’	N	0	7
Turm	‘tower’	N	0	6
Sicherheit	‘safety’	N	5	1

Resources for Data Investigation

Resources for Data Investigation

- Corpus data:

  German newspaper corpus from the 1990s;
  approx. 200 million words

  → co-occurrence analyses between stimuli and responses
  → training data for the statistical grammar model

- Statistical grammar model:

  German lexicalised PCFG; focus on verb subcategorisation;
  unsupervised training on 35 million words from corpus

  → corpus-based quantitative lexical information

- GermaNet:
  lexical semantic taxonomy → semantic relations

Linguistic Analyses of Experiment Data

Linguistic Analyses of Experiment Data

1. Distributional word meaning
    - » Morpho-syntactic analysis
    - » Syntax-semantic noun functions
    - » Co-occurrence analysis
2. Semantic relations

Morpho-Syntactic Analyses

Motivation

- Focus: feature choice in distributional descriptions to model word meaning
- Distinguish and quantify the part-of-speech categories of the associate responses
  - » preparatory step for the analyses to follow
  - » insight into the relevance of predominant POS categories with respect to meaning aspects

Procedure

- Assign part-of-speech to each response to the stimuli
- Basis: empirical grammar dictionary (verb stimuli), database (noun stimuli)
- 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

Results: Verbs

	V	N	ADJ	ADV	
Freq	19.863	48.905	8.510	1.268	TOKEN
Prob	25	62	11	2	
Freq	9.317	23.524	4.983	802	TYPES
Prob	24	61	13	2	

Examples: Verbs

	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

Results: Nouns

	V	N	PN	ADJ	
Freq	13,905	80,419	3,147	19,075	TOKEN
Prob	12	69	3	16	
Freq	3,601	20,389	1,275	5,658	TYPES
Prob	12	66	4	18	

Examples: Nouns

	V	N	PN	ADJ
Total Prob	12	69	3	16
Ananas 'pineapple'	1	51	3	45
Esel 'donkey'	6	42	4	45
Kopf 'head'	6	89	0	5
Löffel 'spoon'	8	86	0	6
Mund 'mouth'	34	65	0	11
Telefon 'telephone'	41	53	2	4
Tempel 'temple'	5	58	24	13
Wecker 'alarm clock'	36	42	0	22
Zwiebel 'onion'	31	54	0	15

Interpretation

- Nouns play a major role among verb and noun features.
- Correspondence to predominant use of nominal features in distributional descriptions.
- Relevance of part-of-speech categories varies according to the semantic class of the word to model.
- Restricting the categories to nominal features restricts the feature sets to „average“ relevance, does not cover the meaning aspects of all semantic word classes.

Syntax-Semantic Noun Functions

Motivation

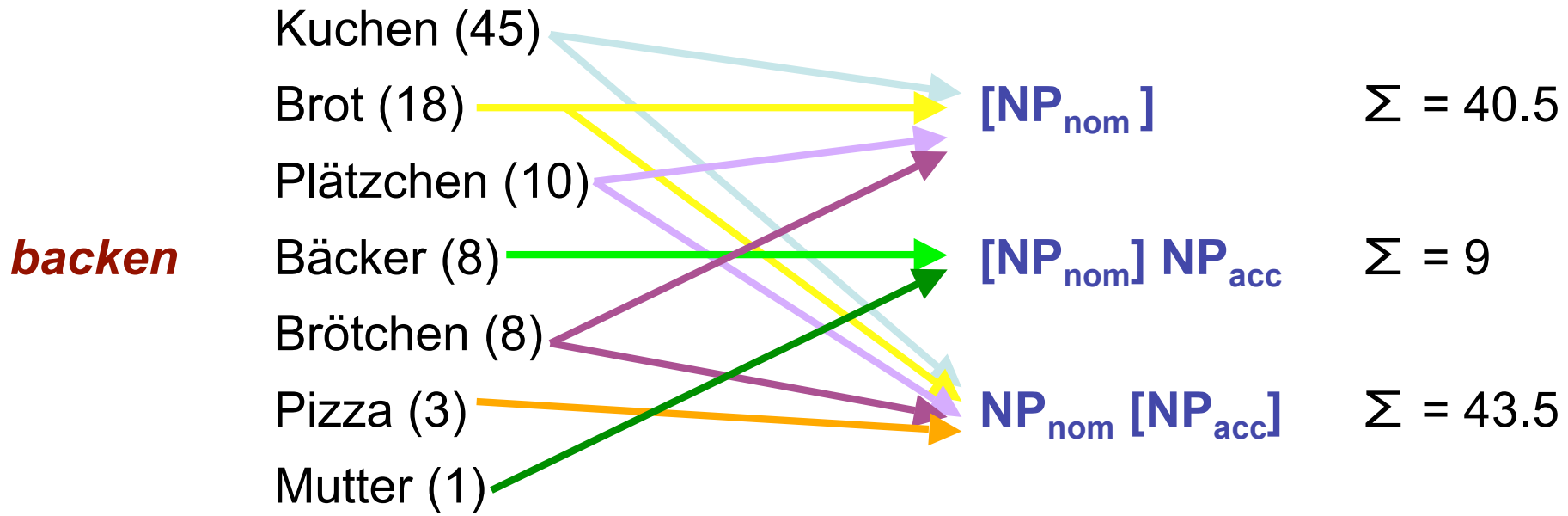
- Focus: feature choice in distributional descriptions to model word meaning → conceptual roles of nouns
- Assumption: noun responses to verb stimuli and verb responses to noun stimuli relate to conceptual roles required by the verbs
- Identify prominent roles for distributional verb descriptions by evaluating which functional roles are highlighted by verb-noun pairs
- Basis: empirical grammar model

Procedure

- Source: statistical grammar model
- Verb valency:
  - » 38 syntactic subcategorisation frames
  - » plus PP information (case+preposition) → 178 frames
  - » subcategorised nouns → 592 roles
- Example: *backen* ‘bake’
  - » frames: **NP_{nom}**
    NP_{nom} NP_{acc} ...
  - » filler examples for **NP_{nom} [NP_{acc}]**: *Brot* ‘bread’
    Kuchen ‘cake’ …

Procedure

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



Results: Verbs

Function		TOKEN (all)	
S	**S** V	1,792	4
	S V AO	1,040	2
	S V DO	265	1
	S V PP	575	1
AO	S V **AO**	3,124	6
	S V **AO** DO	824	2
	S V **AO** PP	653	1
DO	S V **DO**	268	1
	S V AO **DO**	468	1
PP	S V **PP:in$_{Dat}$**	487	1
Total (of these 10)		9,496	**19**
Total found in grammar		13,527	**28**
Unknown verb or noun		10,964	**22**
Unknown function		24,250	**50**

Results: Nouns

Function		TOKEN (all)	
S	**S** V	1,095	8
	S V AO	300	2
	S V PP	406	3
	S V C-2	103	1
	S V INF	71	1
AO	S V **AO**	1,480	11
	S V **AO** DO	206	1
	S V **AO** PP	218	2
DO	S V **DO**	144	1
	S V AO **DO**	99	1
PP	S V **PP:auf$_{Dat}$**	263	2
	S V **PP:in$_{Dat}$**	193	1
Total (of these 12)		4,578	**33**
Total found in grammar		5,661	**41**
Unknown verb or noun		1,505	**11**
Unknown function		6,712	**48**

Interpretation

- Missing nouns/verbs in grammar model (22/11%):
  - » lemmatisation of compound nouns, e.g. *Autorennen*
  - » domain of the training corpus, e.g. slang responses (*Grufties* `old people‘), technical expressions (*Plosiv* `plosive‘)
  - » coverage of corpus: 99% verbs, 78/90% nouns
- Strong correlation between frequency of frame-slot combination in grammar model and number of responses that link to that frame-slot combination in our data
  - → direct object and subject roles are represented proportionate to their frequency in the grammar

Interpretation

- 50/48% verb-noun pairs with no functional relation, e.g.:

  bemalen `paint´ → Pinsel `brush´
  erhitzen `heat´ → Pfanne `pan´
  bemerken `notice´ → Aufmerksamkeit `attention´
  feiern `celebrate´ → Musik `music´

  Handtuch `towel´ → trocknen `dry´
  Zange `pincer´ → biegen `bend´
  Kissen `cushion´ → schlafen `sleep´
  Nase `nose´ → riechen `smell´

- Noun stimuli/responses are not restricted to verb sub-categorisation role fillers
  → clause-internal adjuncts and clause-external, scene-related information or world knowledge as nominal features in distributional descriptions

Co-Occurrence Analysis

Motivation

- Verb-noun pairs within the association norms might co-occur in local contexts even if not related by a sub-categorisation function
- Focus: feature choice in distributional descriptions to model word meaning → role of co-occurrence
- Human associations reflect word co-occurrence probabilities (McKoon and Ratcliff, 1992; Plaut, 1995)
- Observed correlations between associative strength and word co-occurrence (Spence and Owens, 1990)
- Use of low-level co-occurrence information in corpus-based word descriptions?

Procedure

- Use complete newspaper corpus, 200 million words
- Check whether the associate responses occur in a window of 20 words to the left or to the right of the relevant stimulus word
- Determine co-occurrence strength between stimuli and their associations

Results: Verbs

POS	Co-Occurrence Strength						
	1	2	3	5	10	20	50
all	***77***	*70*	*66*	*59*	*50*	*40*	*27*
V	**79**	71	67	60	50	41	29
N	**76**	69	66	59	50	40	27
ADJ	**77**	69	64	57	45	36	22
ADV	**91**	88	85	80	72	62	50

Results: Nouns

POS	Co-Occurrence Strength						
	1	2	3	5	10	20	50
all	***84***	*77*	*72*	*64*	*52*	*38*	*23*
V	**88**	82	77	69	57	44	28
N	**84**	78	72	65	53	39	23
ADJ	**83**	76	70	63	50	36	20

Interpretation

- Co-occurrence assumption holds for our German association data, to a large extent: 77/84\% coverage of response tokens
- Scene-related information beyond the clause level captured by corpus co-occurrence (vs. subcategorisation)
- Co-occurrence information is less expensive than annotated data

  → co-occurrence information as integral component for empirical descriptions of word properties

Interpretation

- Stimulus-associate pairs without co-occurrence, e.g.

  nieseln `drizzle´ → nass `wet´
  mampfen `munch´ → lecker `yummy´
  auftauen `defrost´ → Wasser `water´
  überraschen `surprise´ → Freude `joy´
  leiten `guide´ → Verantwortung `responsibility´

  Ananas `pineapple´ → gelb `yellow´
  Geschenk `present´ → Überraschung `surprise´
  Walnuss `walnut´ → Weihnachten `Christmas´
  Magnet `magnet´ → Physik `physics´

- Challenge to empirical models of word meaning

Summary: Distributional Word Meaning

- Nouns play a major role among verb and noun features.
- Strong correlation between frame-slot combinations in grammar model and in our data → no linguistic functions could be considered to be prominent to represent conceptual nominal roles for verbs.
- Noun associations are not restricted to verb subcategorisation role fillers; clause-internal adjuncts and clause-external, scene-related information or world knowledge should also play a role as features → co-occurrence for empirical descriptions of word properties.

Semantic Relations

Motivation

- Focus: types of relationships between stimulus words and associate responses
- For many NLP resources and applications, it is crucial to define and use semantic relations between words or contexts
- Limited work has been spent on specifying the range of relations
- Semantic associates provide a means to investigate the range of semantic relations

Procedure

- Semantic relations between stimulus and response verb-verb and noun-noun pairs
- Source: lexical semantic taxonomy GermaNet (GWN)
- Synonymy: target and response verb in common synset
- Other semantic relations:

  look up GermaNet semantic relations between
  » stimulus synsets
  » response synsets
- Quantification of target-response relation:

  association frequency

Results: Verbs

Relation	*GermaNet*		*Token*	
Synonymy	4,633		792	4
Antonymy	571	226	209	1
Hypernymy	19,424	9,275	1,343	7
(indirect)			540	3
Hyponymy	19,424	9,275	1,702	9
(indirect)			514	3
Co-Hyponymy	102,018	55,122	2,232	12
(indirect)			1,517	8
Cause	236	95	40	0
Entailment	15	8	0	0
Total in GWN			**8,859**	**46**
Unknown			2,207	12
No relation			7,841	41

Results: Nouns

Relation	*GermaNet*		*Token*	
Synonymy	18,992		533	1
Antonymy	1,553	478	33	0
Hypernymy	82,685	30,707	1,387	2
(indirect)			2,365	3
Hyponymy	82,829	30,708	714	1
(indirect)			289	0
Co-Hyponymy	575,585	302,755	3,584	4
(indirect)			2,964	4
Holonymy	8,625	3,995	579	1
(indirect)			102	0
Meronymy	8,625	3,998	1,171	1
(indirect)			224	0
Total in GWN			**14,028**	**17**
Unknown			13,543	17
No relation			52,814	66

Interpretation

- Distribution of stimulus-response relations is correlated with stimulus frequency:
  synonym, antonym, hyponym ~ verb freq;
  hypernym, (co-)hyponym, hyponym, meronym ~ noun freq
- Distribution of relations varies by verb class
- Unknown cases (12/17%):
  - » part-of-speech confusion, e.g. *wärme* `warmth´ as verb
  - » regional expressions, e.g. *Weck* `roll´
  - » proper names, e.g. *Moses*
  - » production: particle verbs, noun compounds

Interpretation: No Relation

- Incomplete taxonomy, e.g.
  - analysieren `analyse´ → untersuchen `examine´ (synonymy)
  - schwitzen `sweat´ → frieren `be cold´ (antonymy)
  - Anker `anchor´ → Schiff `ship´ (holonymy)
  - Kaktus `cactus´ → Stachel `spine´ (meronymy)
- Other relations, e.g.
  - adressieren `address´ → schicken `send´ (temporal following)
  - schwitzen `sweat´ → stinken `stink´ (consequence)
  - erfahren `get to know´ → wissen `know´ (implication)
  - Kamel `camel´ → Wüste `desert´ (location)
  - Gans `goose´ → Weihnachten `Christmas´ (occasion)
  - Schlitten `sledge´ → Schnee `snow´ (condition)
- Compound nouns (12%), e.g.
  - Melone `melon´ → Honig `honey´ *(Honigmelone `cantaloupe´)*
  - Schale `bowl´ → Obst `fruit´ *(Obstschale `fruit bowl´)*

Summary: Semantic Relations

- Major proportion of stimulus-associate pairs not related via GWN taxonomy:
  - » missing links in GermaNet;
    association data provide a useful starting point to enhance the taxonomy
  - » relations other than those coded in GermaNet,
    such as temporal order, cause, consequence for verb-verb pairs,
    and condition, instrument, result for noun-noun pairs
- Hope: human associations cover the range of possible semantic relations to a large extent, and they represent an excellent basis for defining an exhaustive set

Final Comments

- Association norms have contributed to the under-standing of issues in computational linguistics.
- Results are to a large extent correlated with the semantic classes of the stimuli, and/or with their corpus frequencies. → For specifying word properties and word-word relations with respect to individual words, the semantic class and the frequency range of that word should be taken into account, in order to go beyond an „average“ empirical description.