Distributional Analyses of Semantic Associates

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

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

Dipartimento di Linguistica, Università di Pisa
February 6, 2008

Semantic Associates

- Semantic associates: words that are spontaneously called to mind in response to a given stimulus

  cook → kitchen, bake, hot, soup, yummy, ...

- Cognitive science: investigate mechanisms underlying the semantic memory
  (representation and access of semantic information)

- Computational linguistics: empirical instantiations of semantic meaning and semantic relatedness

Motivation

- Assumption: semantic associates reflect highly salient linguistic and conceptual features of the stimulus word
- Goals:
  - » identify types of information provided by speakers
  - » distinguish and quantifying relationships between stimulus and response
  - » support creation of NLP resources and definition and application of NLP techniques

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.

Excursus: Distributional Word Meaning

Distributional Word Meaning

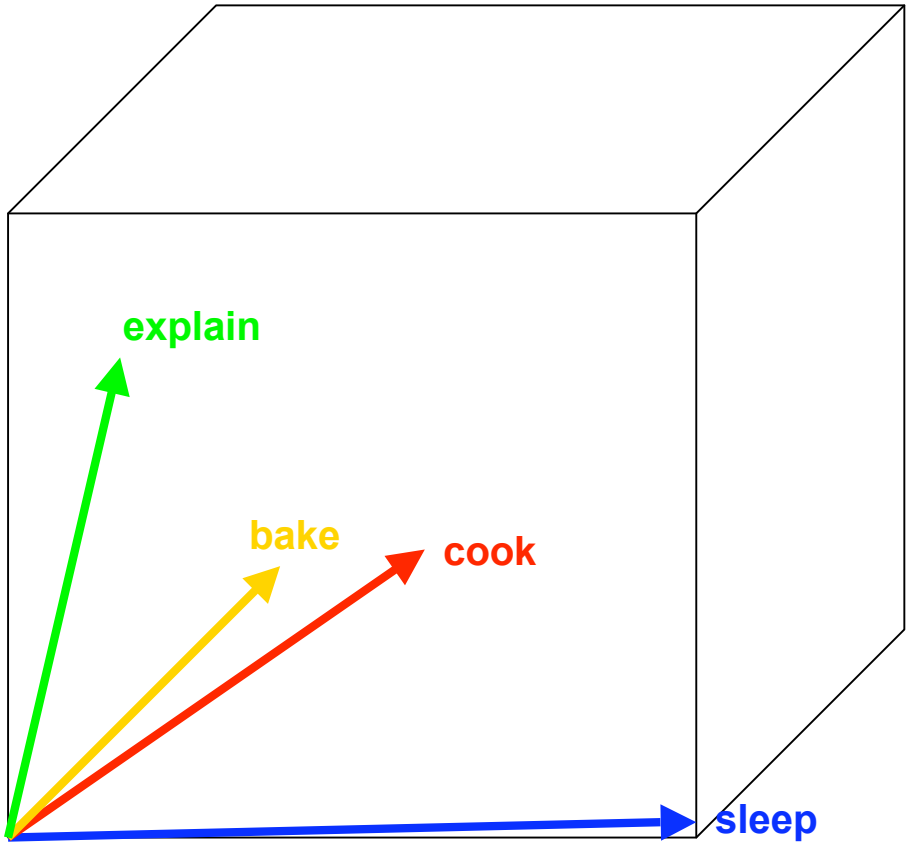
- Words are represented by distributional features
- Basis: distributional hypothesis (Harris, 1968), that *each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts*
- Hypothesis: contextual embedding is related to meaning
- Examples of distributional features:
  - » subcategorisation frames,
  - » bag-of-words in sentences, paragraphs, documents,
  - » words in specific syntactic functions, e.g., direct nouns

Distributional Word Meaning

Subcategorisation frames of verbs

	NP_{nom}	NP_{nom} NP_{acc}	NP_{nom} NP_{acc} NP_{dat}
schlafen *'sleep'*	98	1	1
kochen *'cook'*	35	50	15
backen *'bake'*	14	70	16
erklären *'explain'*	10	32	58

Distributional Word Meaning



Distributional Word Meaning

Nouns in verb context

	bed	*kitchen*	*task*	*child*
schlafen *'sleep'*	250	30	10	111
kochen *'cook'*	5	384	60	30
backen *'bake'*	8	498	3	80
erklären *'explain'*	60	20	445	98

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 functions activated by associates

Procedure

- Basis: collection of semantic associates evoked by German verbs and nouns
- Goal: empirical characterisation of verb and noun properties
- Focus: modelling word meaning by empirical features
- Analyses:
  - » Motivation by potential NLP uses
  - » Exploration of relationships between stimuli and responses
  - » Basis: large-scale lexicographic databases and empirical, corpus-based resources to characterise the associations

Overview

1. Data collection and preparation
2. Resources for data investigation
3. Linguistic analyses of experimental data
   (a) NLP motivation
   (b) analyses: *verbs / nouns*
   (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

Experiment Procedure: Verbs

- Web experiment over Internet
- Bibliographic information:
  linguistic experience, age, regional accent, profession
- Instructions and example page
- Experiment 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 `to snow´

kalt	`cold´
rodeln	`sledge´
Schneemann	`snowman´
weiß	`white´
dämmern	`dawn´

Experiment Data: Verbs

- 299 accepted data files from native German speakers
- 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
  (first) 15,788 tokens for 7,425 types

Data Preparation: Verbs

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

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
- 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´ ...
- Depictable objects
- 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
- All associations: 116,714 tokens for 31,035 types
  (first) 39,727 tokens for 11,389 types

Modality: *word (+ picture)*

Witch

magic

wizard

broom

Data Preparation: Nouns

Schloss `lock´ (depicted), `castle´

Association		POS	PW	W	PW&W
Schlüssel	‘key’	N	38	13	51
Tür	‘door’	N	10	5	15
Prinzessin	‘Princess’	N	0	8	8
Burg	‘castle’	N	0	8	8
sicher	‘safe’	ADJ	7	0	7
Fahrrad	‘bike’	N	7	0	7
schließen	‘close’	V	6	1	7
Keller	‘cellar’	N	7	0	7
König	‘king’	N	0	7	7
Turm	‘tower’	N	0	6	6
Sicherheit	‘safety’	N	5	1	6

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

Linguistic Analyses of Experiment Data

Overview of Analyses

- Morpho-syntactic analysis
- Syntax-semantic noun functions
- Co-occurrence analysis

Morpho-Syntactic Analysis

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

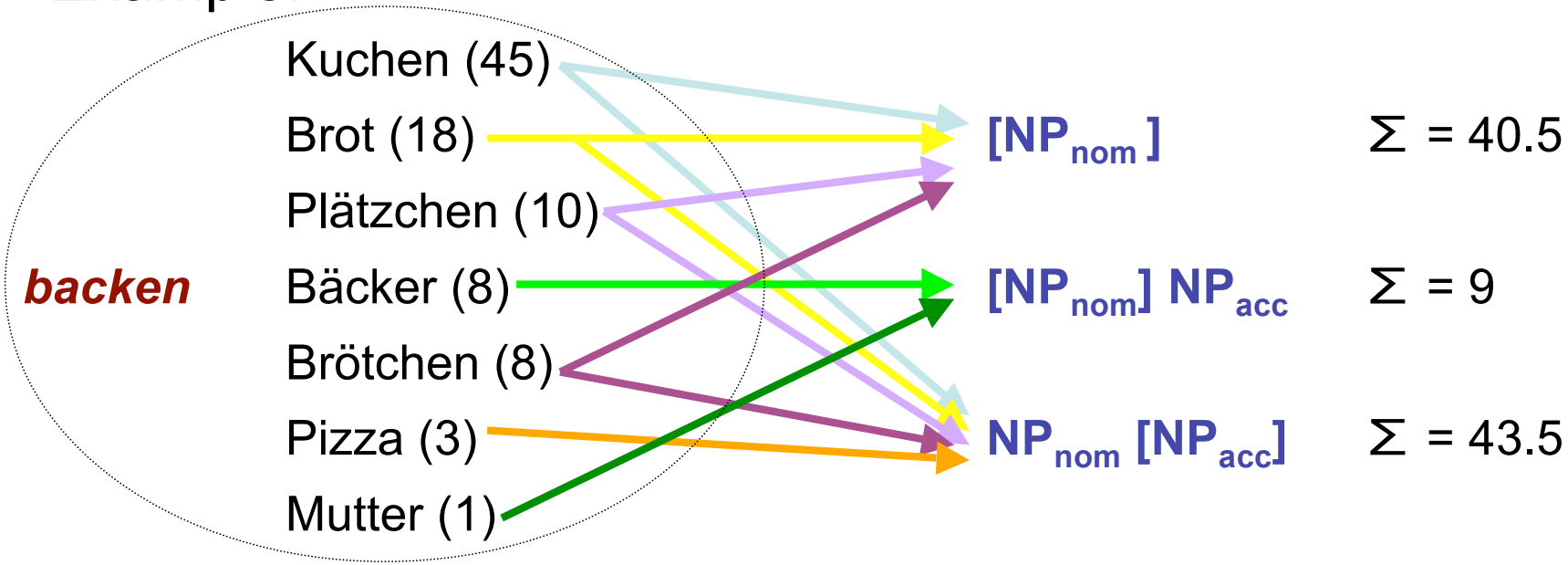
- Focus: feature choice in distributional descriptions to model word meaning → conceptual roles of nouns
- Assumption: noun responses to verb stimuli and verb responses to noun stimuli relate to conceptual roles required by the verbs
- Identify prominent roles for distributional verb descriptions by 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)		TYPES (all)	
S	**S** V	1,792	4	479	2
	S V AO	1,040	2	371	2
	S V DO	265	1	82	0
	S V PP	575	1	208	1
AO	S V **AO**	3,124	6	972	4
	S V **AO** DO	824	2	234	1
	S V **AO** PP	653	1	205	1
DO	S V **DO**	268	1	102	0
	S V AO **DO**	468	1	141	1
PP	S V **PP:in$_{Dat}$**	487	1	98	0
Total (of these 10)		9,496	19	2,892	12
Total found in grammar		**13,527**	**28**	**4,210**	**18**
Unknown verb or noun		10,964	22	6,951	30
Unknown function		24,250	50	12,255	52

Results: Nouns

Function		TOKEN (all)		TYPES (all)	
S	**S** V	1,095	8	173	5
	S V AO	300	2	58	2
	S V PP	406	3	69	2
	S V C-2	103	1	11	0
	S V INF	71	1	10	0
AO	S V **AO**	1,480	11	241	7
	S V **AO** DO	206	1	35	1
	S V **AO** PP	218	2	44	1
DO	S V **DO**	144	1	15	0
	S V AO **DO**	99	1	16	0
PP	S V **PP:auf$_{Dat}$**	263	2	18	0
	S V **PP:in$_{Dat}$**	193	1	22	1
Total (of these 12)		4,578	33	712	19
Total found in grammar		**5,661**	**41**	**933**	**26**
Unknown verb or noun		1,505	11	430	12
Unknown function		6,712	48	2,212	62

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‘),
    dialect expressions (*Ausstecherle* `cookie-cutter‘),
    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
- 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
- Scene-related information beyond the clause level captured by corpus co-occurrence (vs. subcategorisation)
- Adverbs show strong co-occurrence: token-type ratio in corpus; few grammatical restrictions; relevant if close
- 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.

  bemalen `paint´ → Pinsel `brush´
  nieseln `drizzle´ → nass `wet´
  mampfen `munch´ → lecker `yummy´
  auftauen `defrost´ → Wasse `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.

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.