# Compositionality of German Noun-Noun Compounds and German Particle Verbs: Experiential Data and Distributional Models

PD Dr. Sabine Schulte im Walde

Heisenberg-Gruppe *SemRel*
Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart

*Heinrich-Heine-Universität Düsseldorf, SFB 991*
July 4, 2013

# Overview

1. Phenomena and Framework
2. Distributional Models of Lexical Semantics
3. German Noun-Noun Compounds
4. German Particle Verbs

# Compounds

- Compounds are combinations of two or more simplex words.
- Compounds represent a recurrent focus of attention within theoretical, cognitive, and computational linguistics.
  - Handbook of Compounding (Lieber & Stekauer, 2009)
  - Series of workshops and special journal issues focusing on multi-word expressions, cf. `multiword.sourceforge.net`:
    - Journal of Computer Speech and Language, 2005
    - Language Resources and Evaluation, 2010
    - ACM Transactions on Speech and Language Processing, t.a.
- Our research focus: compositionality

# German Noun-Noun Compounds

- Composition:
  - two-part compounds, i.e., compounds consisting of two simplex constituents
  - both modifiers and heads are nouns
- Examples:
  - *Postbote* 'post man': *Post* 'mail' + *Bote* 'messenger'
  - *Löwenzahn* 'dandelion': *Löwe* 'lion' + *Zahn* 'tooth'
  - *Fliegenpilz* 'toadstool': *Fliege* 'fly/bow tie' + *Pilz* 'mushroom'
  - *Feuerzeug* 'lighter': *Feuer* 'fire' + *Zeug* 'stuff'
- References: Fleischer & Barz (2012); Klos (2011)

# German Particle Verbs (PVs)

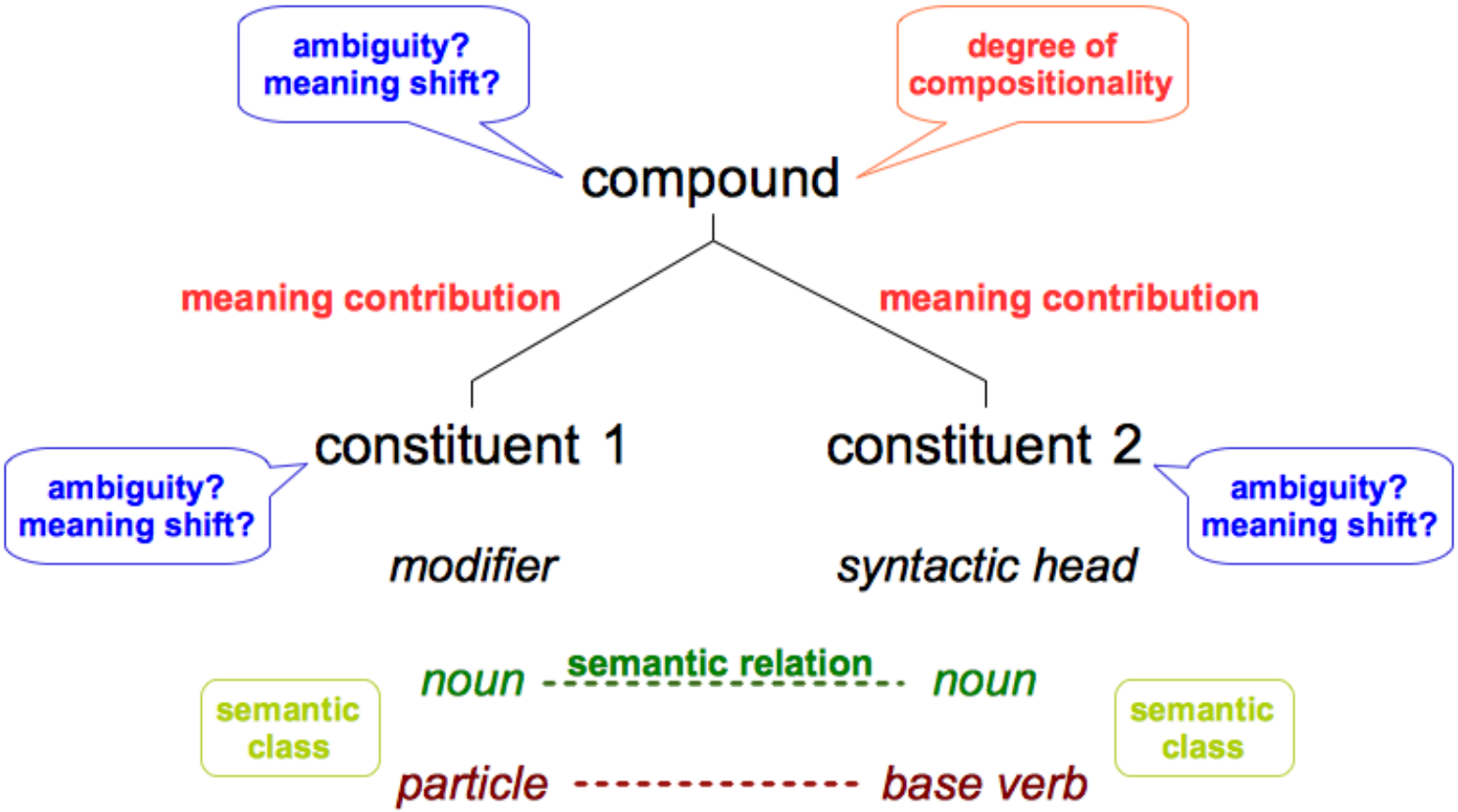
- Composition:
  - composition of base verbs (BVs) and prefix particles
  - focus: preposition particles
- Examples:
  - *abholen* 'fetch': *ab* + *holen* 'fetch'
  - *anfangen* 'begin': *an* + *fangen* 'catch'
  - *einsetzen* 'insert'/'begin': *ein* + *setzen* 'put/sit (down)'
- References:
  - Stiebels (1996); Lüdeling (2001); Dehe et al. (2002)
  - Lechler & Roßdeutscher (2009); Kliche (2011); Springorum (2011)

# Research Focuses

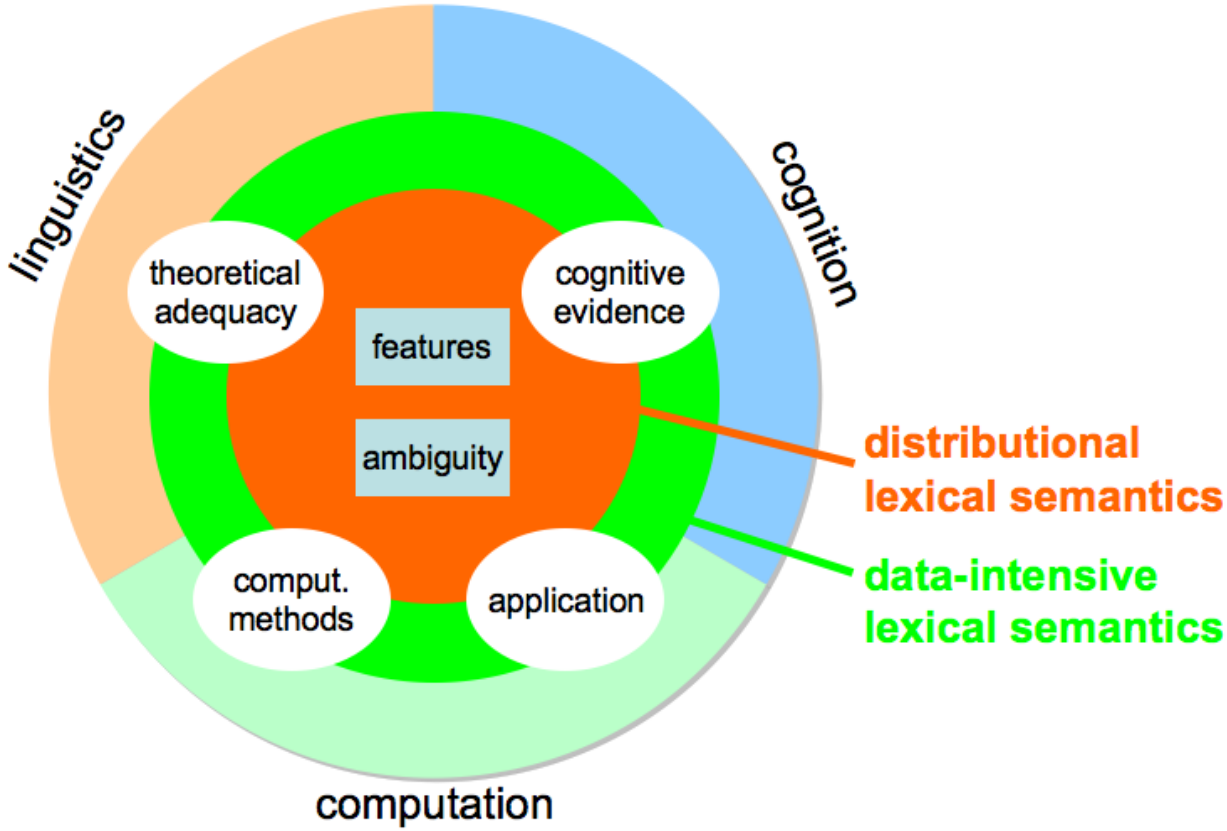
1. Degree of compositionality of compounds
2. Contribution of meaning aspects of constituents to compound meaning
3. Role of ambiguity

4. Role of modifier vs. head constituents (in noun-noun compounds)
5. Metaphorical shifts of particle (verb) meaning aspects

# Research Focuses



# Research Framework



# Distributional Models

# Distributional Semantics

- Distributional Hypothesis:

  *You shall know a word by the company it keeps.* (Firth, 1957)

  *Each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts.* (Harris, 1968)

- Distributional Semantics exploits the *distributional hypothesis* to identify contextual features for vector space models that best describe the words, phrases, sentences, etc. of interest.

# Vector Space Models

- Vector Space Models (VSMs): explore the notion of "similarity" between a set of target objects within a geometric setting. (Turney and Pantel, 2010; Erk, 2012)
- Idealised concept: a lexical item is defined by the total of contextual features (co-occurrence).
- Co-occurrence features: corpus-based, salient contextual properties of the target lexical items.

# Vector Space Models: Example 1

- Matrix:

| | grün | gelb | schälen | fallen | Baum |
|---|---|---|---|---|---|
| Apfel | 80 | 1 | 311 | 22 | 105 |
| Banane | 13 | 56 | 83 | 2 | 8 |
| Blatt | 258 | 0 | 1 | 98 | 244 |

# Vector Space Models: Example 1

- Matrix:

| | grün | gelb | schälen | fallen | Baum |
|---|---|---|---|---|---|
| Apfel | 80 | 1 | 311 | 22 | 105 |
| Banane | 13 | 56 | 83 | 2 | 8 |
| Blatt | 258 | 0 | 1 | 98 | 244 |

- Vector:

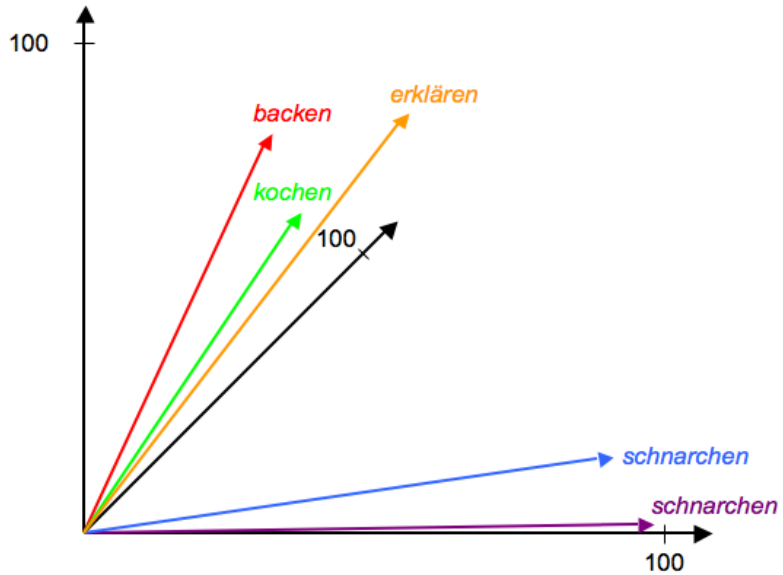
  Apfel: $\langle 80,1,311,22,105 \rangle$
  Banane: $\langle 13,56,83,2,8 \rangle$
  Blatt: $\langle 258,0,1,98,244 \rangle$

# Vector Space Models: Example 2

| | ⟨NPnom⟩ | ⟨NPnom,NPacc⟩ | ⟨NPnom,NPacc,NPdat⟩ |
|---|---|---|---|
| schlafen | 98 | 1 | 1 |
| kochen | 35 | 50 | 15 |
| backen | 14 | 70 | 16 |
| erklären | 10 | 32 | 58 |
| schnarchen | 90 | 1 | 9 |

# Vector Space Models: Example 2

| | ⟨NPnom⟩ | ⟨NPnom,NPacc⟩ | ⟨NPnom,NPacc,NPdat⟩ |
|---|---|---|---|
| schlafen | 98 | 1 | 1 |
| kochen | 35 | 50 | 15 |
| backen | 14 | 70 | 16 |
| erklären | 10 | 32 | 58 |
| schnarchen | 90 | 1 | 9 |



# German Noun-Noun Compounds

# Phenomenon

- Composition:
  - two-part compounds, i.e., compounds consisting of two simplex constituents
  - both modifiers and heads are nouns
- Examples:
  - *Postbote* 'post man': *Post* 'mail' + *Bote* 'messenger'
  - *Löwenzahn* 'dandelion': *Löwe* 'lion' + *Zahn* 'tooth'
  - *Fliegenpilz* 'toadstool': *Fliege* 'fly/bow tie' + *Pilz* 'mushroom'
  - *Feuerzeug* 'lighter': *Feuer* 'fire' + *Zeug* 'stuff'
- References: Fleischer & Barz (2012); Klos (2011)

# Dataset

- Original dataset:
  - selection of 450 concrete, depictable German noun compounds by von der Heide & Borgwaldt (2009)
  - four compositionality classes (O=opaque; T=transparent): O+O, T+T, O+T, T+O
- Our dataset:
  - subset of above, comprising 244 two-part noun-noun compounds

# Experiential Data

- Human ratings on the degree of compositionality:
  - compound–constituent ratings
  - compound 'whole' ratings
- Association norms
  - compounds
  - modifiers
  - heads
- Feature norms
  - compounds
  - modifiers
  - heads

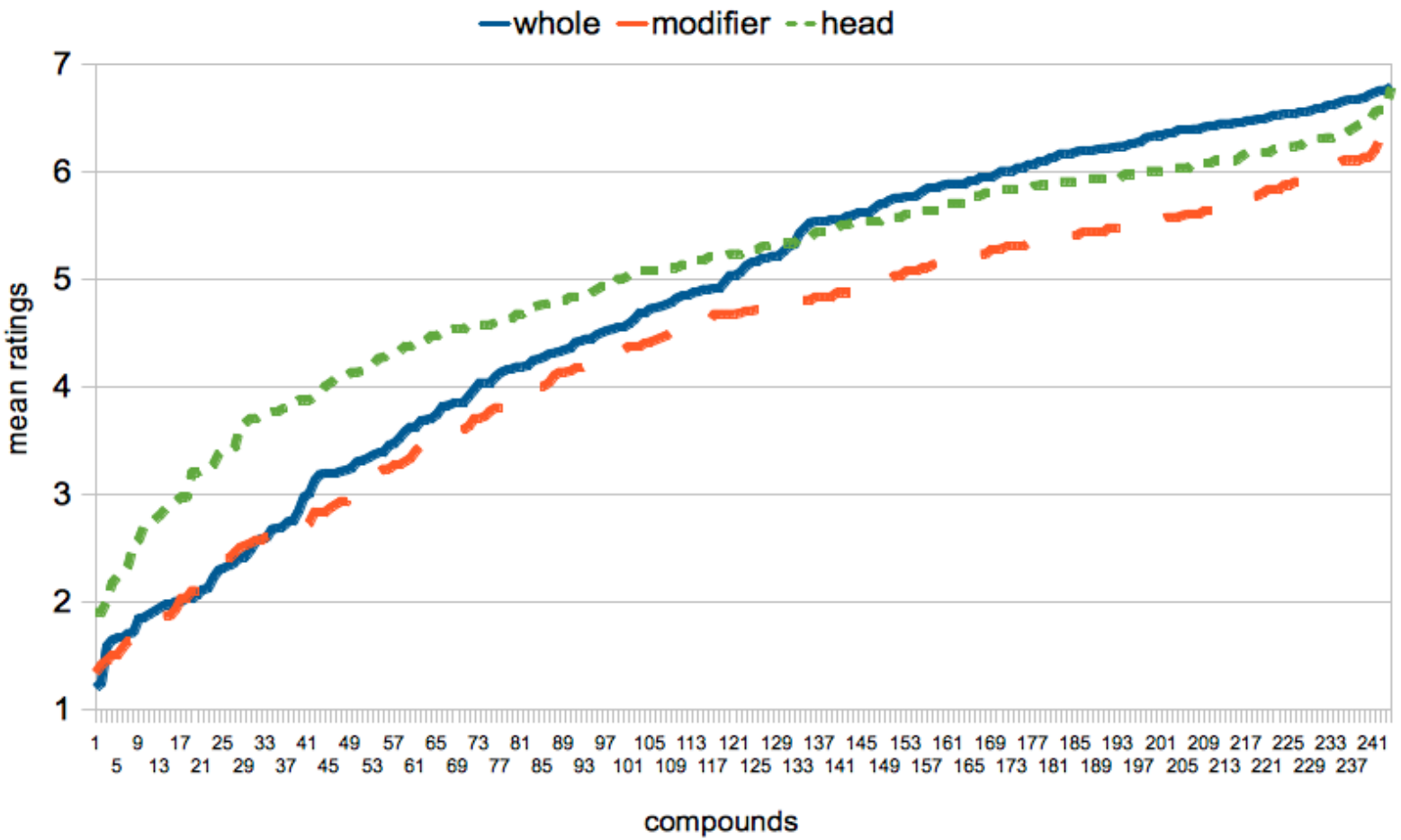
# Compositionality Ratings

- Degree of compositionality: semantic relatedness between compound meaning and meanings of constituents
- Two collections:
  1. Compound–Constituent Ratings (v.d. Heide/Borgwaldt, 2009)
     - Task: degree of compositionality of the compounds with respect to their first as well as their second constituent
     - Scale: 1 (definitely opaque) to 7 (definitely transparent)
  2. Compound Whole Ratings (*SemRel* group)
     - Task: degree of compositionality of the compounds as a whole
     - Scale: 1 (definitely opaque) to 7 (definitely transparent)

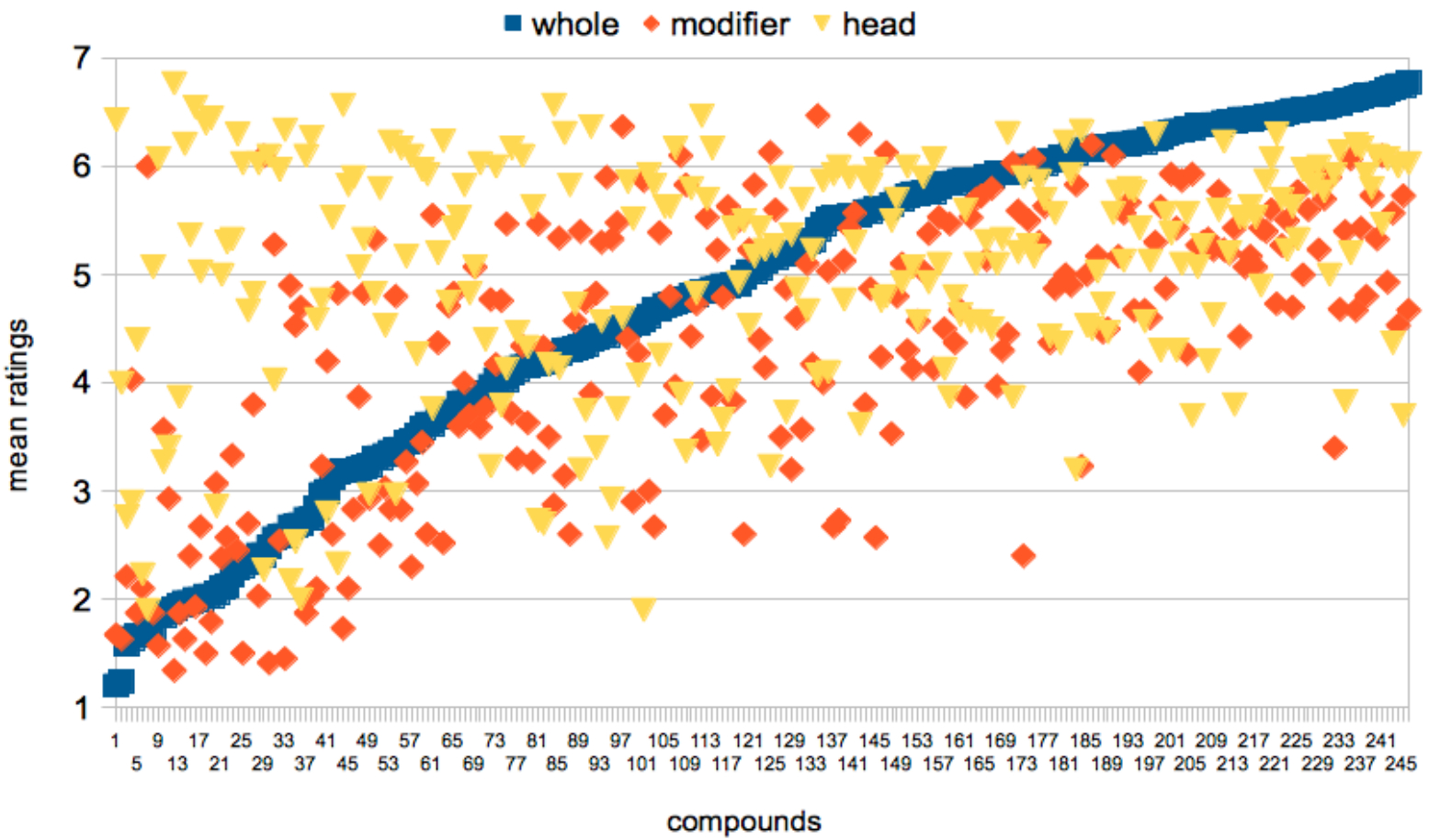
# Compositionality Ratings: Examples

| Compounds | | | Mean Ratings and Standard Deviations | | |
|---|---|---|---|---|---|
| whole | literal meanings of constituents | | whole | modifier | head |
| *Ahornblatt* 'maple leaf' | maple | leaf | 6.03 ± 1.49 | **5.64** ± 1.63 | **5.71** ± 1.70 |
| *Postbote* 'post man' | mail | messenger | 6.33 ± 0.96 | **5.87** ± 1.55 | **5.10** ± 1.99 |
| *Seezunge* 'sole' | sea | tongue | 1.85 ± 1.28 | **3.57** ± 2.42 | **3.27** ± 2.32 |
| *Windlicht* 'storm lamp' | wind | light | 3.52 ± 2.08 | **3.07** ± 2.12 | **4.27** ± 2.36 |
| *Löwenzahn* 'dandelion' | lion | tooth | 1.66 ± 1.54 | **2.10** ± 1.84 | **2.23** ± 1.92 |
| *Maulwurf* 'mole' | mouth | throw | 1.58 ± 1.43 | **2.21** ± 1.68 | **2.76** ± 2.10 |
| *Fliegenpilz* 'toadstool' | fly/bow tie | mushroom | 2.00 ± 1.20 | **1.93** ± 1.28 | **6.55** ± 0.63 |
| *Flohmarkt* 'flea market' | flea | market | 2.31 ± 1.65 | **1.50** ± 1.22 | **6.03** ± 1.50 |
| *Feuerzeug* 'lighter' | fire | stuff | 4.58 ± 1.75 | **5.87** ± 1.01 | **1.90** ± 1.03 |
| *Fleischwolf* 'meat chopper' | meat | wolf | 1.70 ± 1.05 | **6.00** ± 1.44 | **1.90** ± 1.42 |

# Compositionality Ratings: Distribution (1)



# Compositionality Ratings: Distribution (2)



# Association Norms

*Example*: Associations of *snow*?

# Association Norms

*Example*: Associations of *snow*? *white, winter, sledge, . . .*

# Association Norms

*Example*: Associations of *snow*? *white, winter, sledge, ...*

Associations to German noun compounds collected in 2010–2012:

- web experiment with 996 compounds+constituents for 442 noun compounds (Schulte im Walde et al., 2012):
  - 10–36 participants per stimulus
  - 28,238/47,249 stimulus–association types/tokens
- AMT experiment with 571 compounds+constituents for 246 noun-noun compounds (unpublished):
  - 2–120 (in general: 30) participants per stimulus
  - 26,415/59,444 stimulus–association types/tokens
- web data + AMT data contains a total of 47,523/106,693 stimulus–association types/tokens

# Association Norms

| *Fliegenpilz* 'fly agaric' | | | *Fliege* 'fly/bow tie' | | | *Pilz* 'mushroom' | | |
|---|---|---|---|---|---|---|---|---|
| *giftig* | 'poisonous' | 12 | *nervig* | 'annoying' | 4 | *Wald* | 'forest' | 13 |
| *rot* | 'red' | 7 | *summen* | 'buzz' | 2 | *Fliegenpilz* | 'fly agaric' | 4 |
| *Wald* | 'forest' | 5 | *lästig* | 'annoying' | 2 | *sammeln* | 'collect' | 3 |
| *Gift* | 'poison' | 2 | *Insekt* | 'bug' | 2 | *giftig* | 'poisonous' | 3 |
| *Hut* | 'cap' | 1 | *Tier* | 'animal' | 2 | *Schimmel* | 'mould' | 2 |
| *Glück* | 'fortune' | 1 | *Fliegenklatsche* | 'fly flap' | 2 | *Suche* | 'search' | 2 |
| *Kinderbuch* | 'children's book' | 1 | *Krawatte* | 'tie' | 2 | *Hut* | 'cap' | 2 |
| *Pflanze* | 'plant' | 1 | *Sommer* | 'summer' | 2 | *Pilzpfanne* | 'mushroom pan' | 2 |
| *Muster* | 'pattern' | 1 | *Anzug* | 'suit' | 1 | *essbar* | 'eatable' | 1 |
| *weiß* | 'white' | 1 | *fangen* | 'catch' | 1 | *Suppe* | 'soup' | 1 |

# Feature Norms

*Example*: Typical features of *dog*?

# Feature Norms

*Example*: Typical features of *dog*?

*is an animal, has four legs, barks, . . .*

# Feature Norms

*Example*: Typical features of *dog*?

*is an animal, has four legs, barks, ...*

Features of German noun compounds collected in 2012–2013:

- AMT experiment with 571 compounds+constituents for 246 noun-noun compounds (unpublished):
  - 1–63 features per stimulus
  - 7,985/12,660 stimulus–feature types/tokens

Examples:

- *Schneeball* 'snow ball' → *ist kalt* 'is cold' (7), *ist rund* 'is round' (7), *ist weiß* 'is white' (6)
- *Schnee* 'snow' → *ist kalt* 'is cold' (13), *ist weiß* 'is white' (13), *gibt es im Winter* 'exists in winter' (3)
- *Ball* 'ball' → *ist rund* 'is round' (14), *zum Spielen* 'for playing' (3), *kann rollen* 'can roll' (2)

# Models

1. Distributional model of lexical, corpus-based co-occurrence (Schulte im Walde et al., 2013):
   - Task: predict the degree of compositionality of the compounds
   - Subtask 1: compare window-based vs. syntax-based features
   - Subtask 2: compare contributions of modifiers vs. heads

2. Multi-modal model incorporating lexical data (co-occurrence), experiential data (associations, features), and visual data (pictorial features); Roller & Schulte im Walde, submitted
   - Task: predict the degree of compositionality of the compounds

# Lexical Model: Hypotheses

1. Targets in the vector space models are nouns
   (compound nouns, modifier nouns, head nouns).

   → adjectives and verbs provide most salient features,

   → syntax-based outperforms window-based.

# Lexical Model: Hypotheses

1. Targets in the vector space models are nouns
   (compound nouns, modifier nouns, head nouns).

   → adjectives and verbs provide most salient features,

   → syntax-based outperforms window-based.

2. Contributions of *modifier* noun vs. *head* noun:

   distributional properties of heads are more salient than distributional properties of modifiers

   in predicting the degree of compositionality of the compounds.

# Vector Space Models: Setup

- Goal: use VSM to identify salient distributional features to predict the degree of compositionality of the compounds
- Corpora: two German web corpora
- Feature Values: local mutual information (Evert, 2005) of co-occurrence counts (between target nouns and features):

$$LMI = O \times log \frac{O}{E}$$

- Measure of Relatedness: cosine $\sim$ degree of compositionality
- Evaluation: cosine against human ratings;
  Spearman Rank-Order Correlation Coefficient $\rho$
  (Siegel and Castellan, 1988)

# Baseline and Upper Bound

| Function | $\rho$ | |
|---|---|---|
| | Baseline | Upper Bound |
| modifier only | .0959 | .6002 |
| head only | .1019 | .1385 |
| addition | .1168 | .7687 |
| multiplication | .1079 | .7829 |

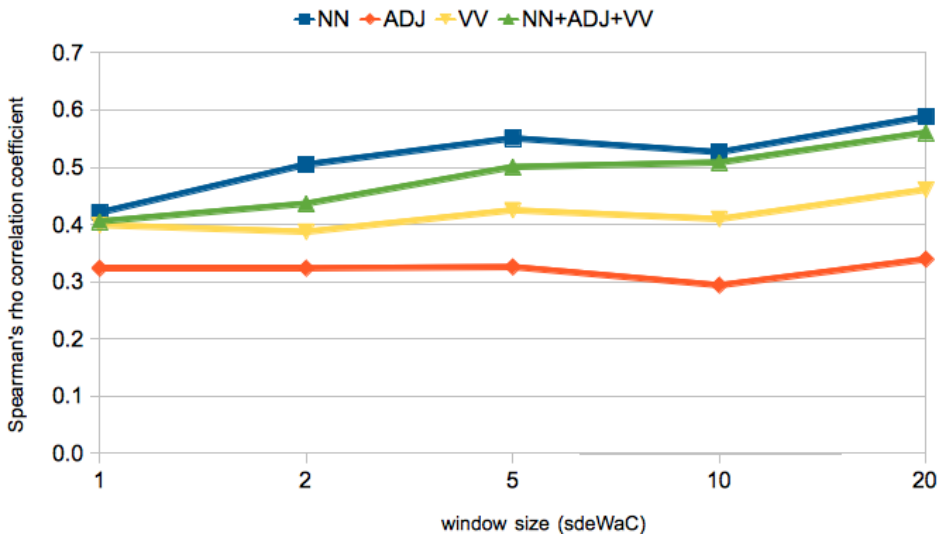
# Corpus Data: German Web Corpora

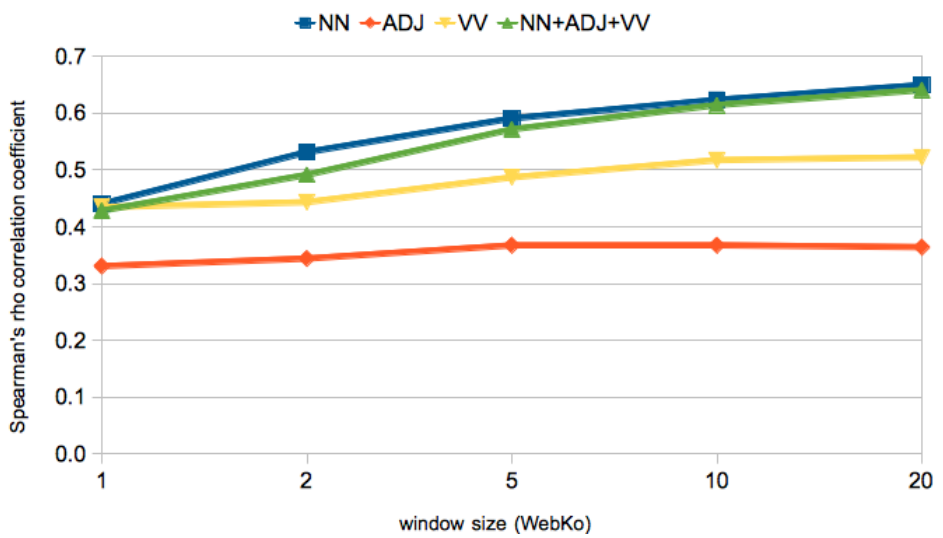
1. *sdeWaC* (Faaß et al., 2010; Faaß & Eckart, 2013)
   - cleaned and parsed version of the German web corpus *deWaC* created by the *WaCky* group (Baroni et al., 2009)
   - corpus cleaning: removing duplicates; disregarding syntactically ill-formed sentences; etc.
   - size: approx. 880 million words
   - disadvantage: sentences in the corpus are sorted alphabetically
     $\rightarrow$ window co-occurrence refers to $x$ words to left and right BUT within the same sentence
2. *WebKo*
   - predecessor version of *sdeWaC*
   - size: approx. 1.5 billion words
   - disadvantage: less clean and not parsed

# Window-based VSMs

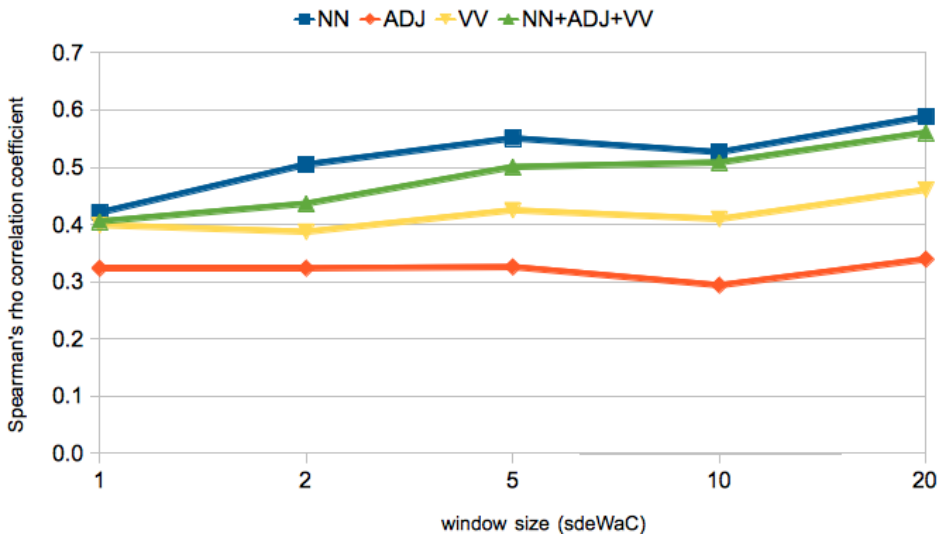
- Hypothesis 1 (i):
  adjectives and verbs provide most salient features
- Task: compare parts-of-speech in predicting compositionality
- Setup:
  - specification of corpus, part-of-speech and window size
  - determine co-occurrence counts and calculate lmi values
  - parts-of-speech: common nouns, adjectives, main verbs
  - window sizes: 1, 2, 5, 10, 20 (, ...100)
  - basis: lemmas; no punctuation
  - *example vector:* adjectives, window of 5 words, WebKo corpus

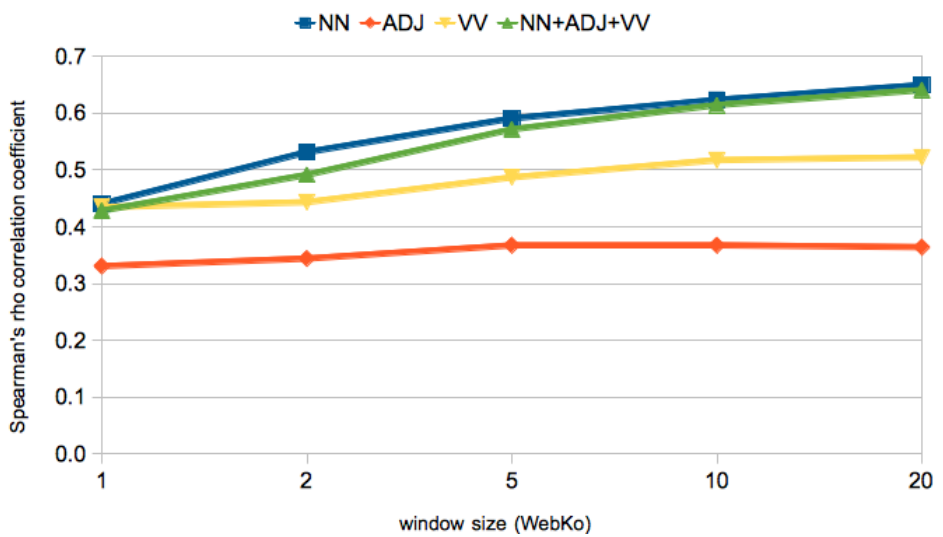
# Window-based VSMs: Results



# Window-based VSMs: Results



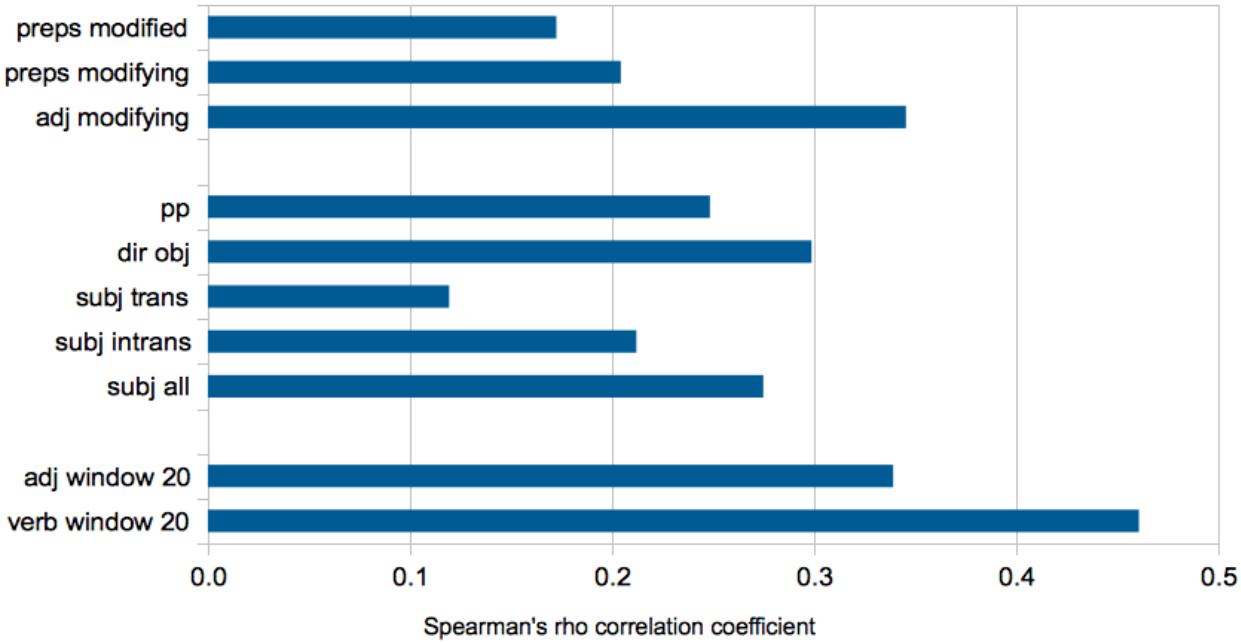




- NN > NN+ADJ+VV > VV > ADJ (significant)
- window sizes: 100 = 50 ∼ 20 > 10 > 5 > 2 > 1
- WebKo > sdeWaC (significant; also with sentence-internal windows)
- best result: $\rho = 0.6497$ (WebKo, NN, window size: 20)

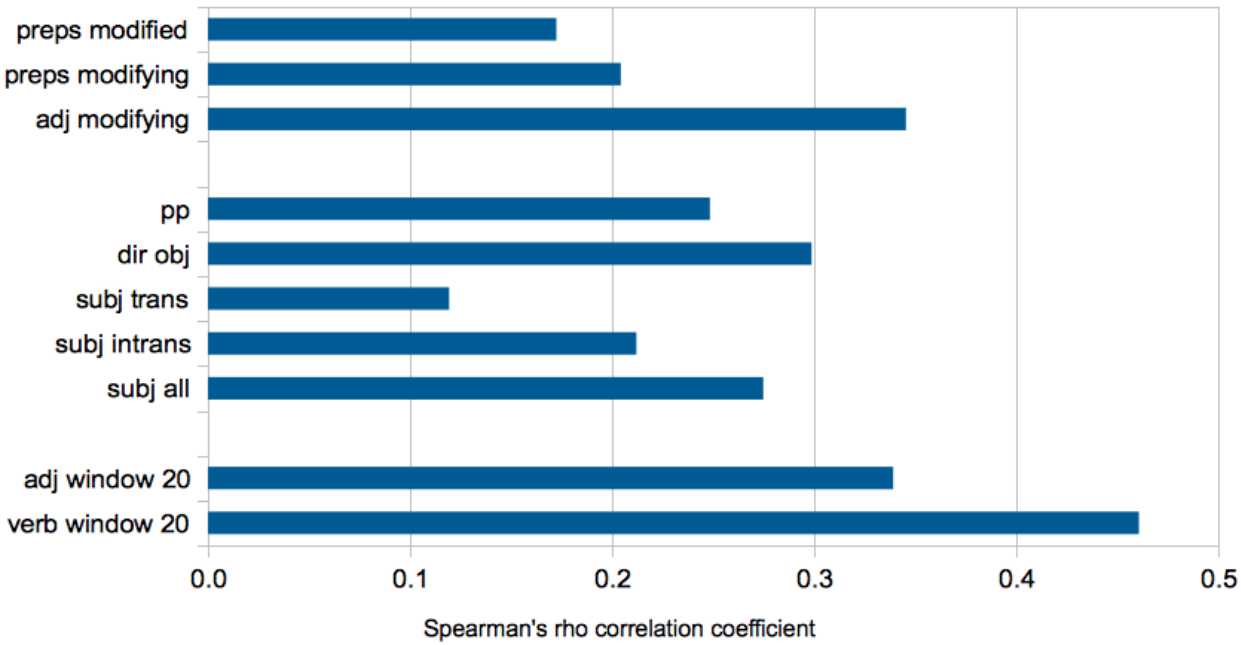
# Syntax-based VSMs

- Hypothesis 1 (ii):
  syntax-based features outperform window-based features
- Task: compare the two co-occurrence conditions
- Setup:
  - corpus choice: sdeWaC (parsed)
  - specification of syntactic function
  - determine co-occurrence counts and calculate lmi values
  - functions:
    - nouns in verb subcategorisation:
      intransitive and transitive subjects; direct and PP objects
    - noun-modifying adjectives
    - noun-modifying and noun-modified prepositions
    - concatenation of all function features

# Syntax-based VSMs: Results



# Syntax-based VSMs: Results

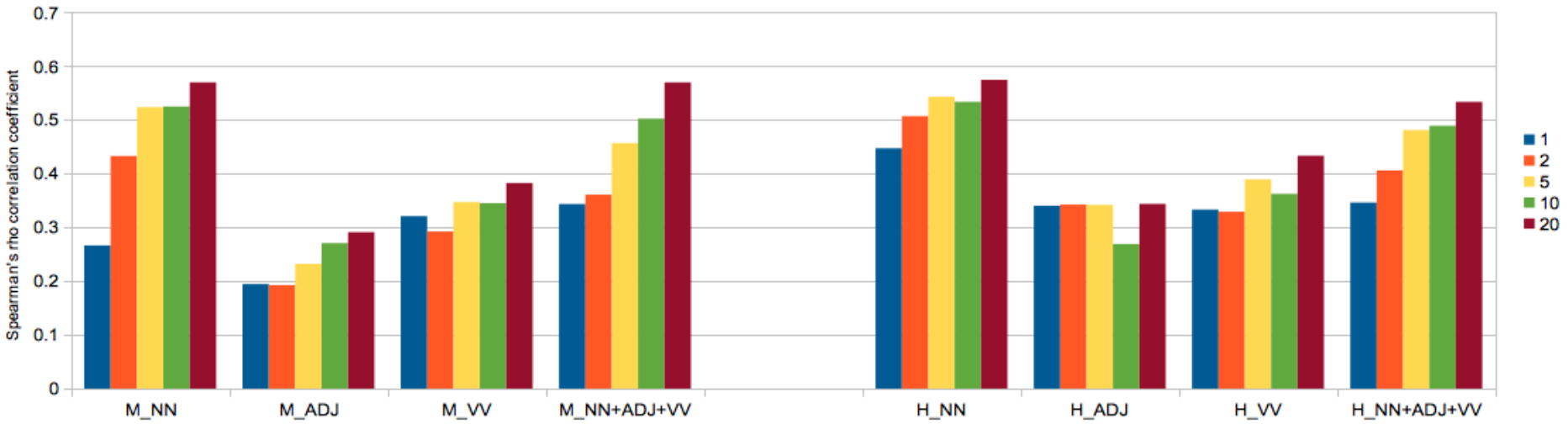



- window-based $>$ syntax-based
- noun-modifying adjectives $\sim$ adjectives in window 20
- verbs in window 20 $>$ verb subcategorisation
- abstracting over subject (in)transitivity $>$ specific functions
- concatenation worse than individual functions

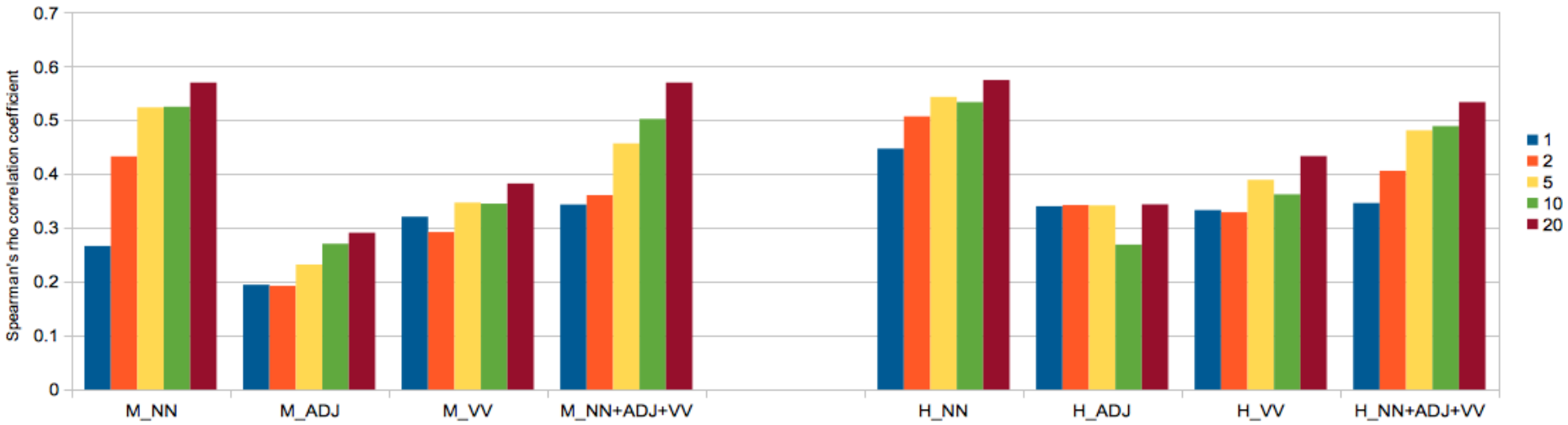
# Role of Modifiers vs. Heads (1)

- Hypothesis 2:
  distributional properties of heads are more salient than distributional properties of modifiers
- Perspective (i): salient features for compound–modifier vs. compound–head pairs
- Setup:
  - same as before (window-based and syntax-based)
  - distinguish evaluation of 244 compound–modifier predictions vs. 244 compound–head predictions (instead of abstracting over the constituent type, using all 488 predictions)

# Role of Modifiers vs. Heads (1): Results



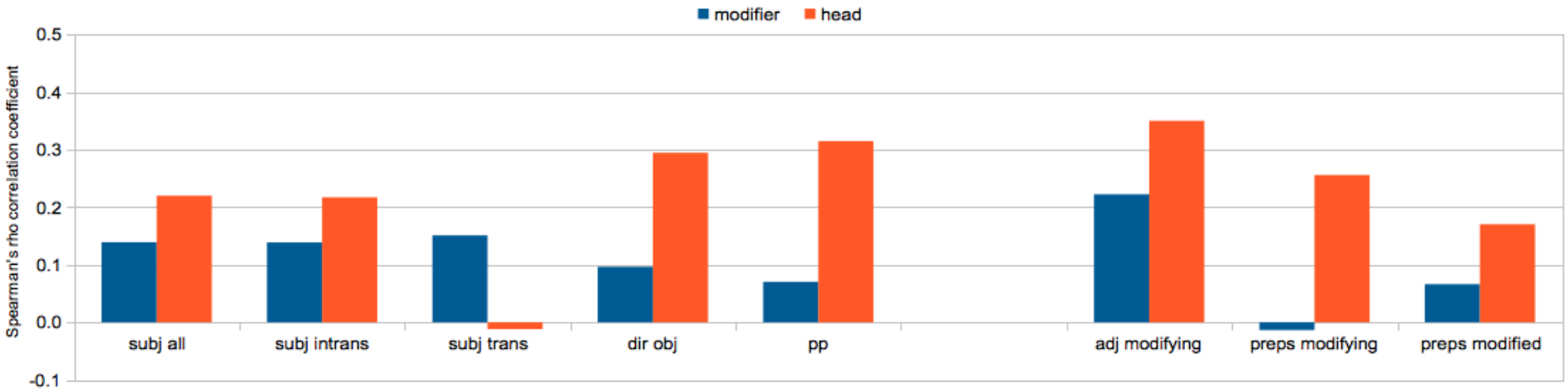
# Role of Modifiers vs. Heads (1): Results



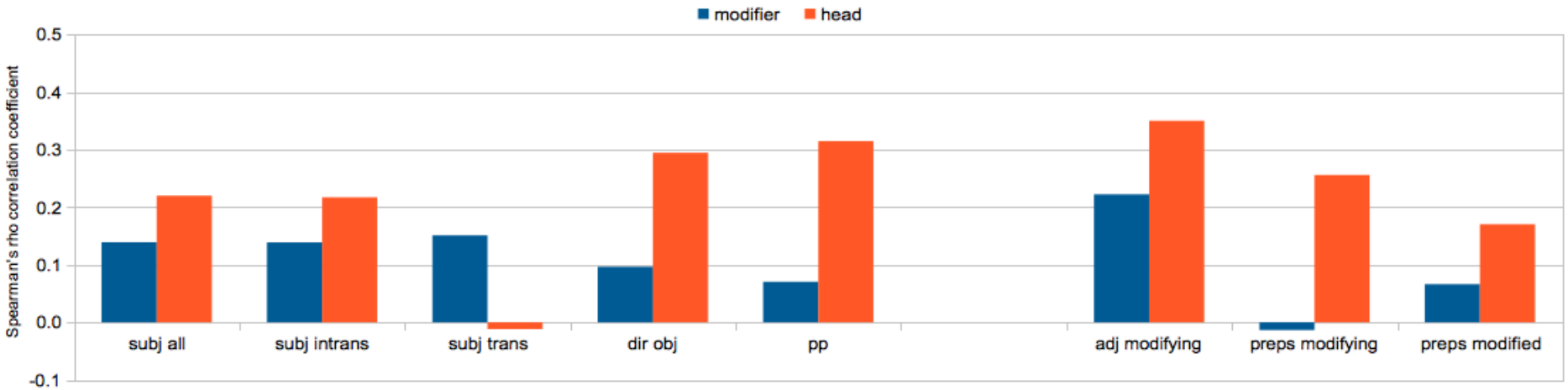

window-based:

- NN > NN+ADJ+VV > VV > ADJ (same as before)
- window sizes: 20 > 10 > 5 > 2 > 1 (same as before)
- small windows: compound–head > compound–modifier predictions
- larger windows: difference vanishes

# Role of Modifiers vs. Heads (1)



# Role of Modifiers vs. Heads (1)



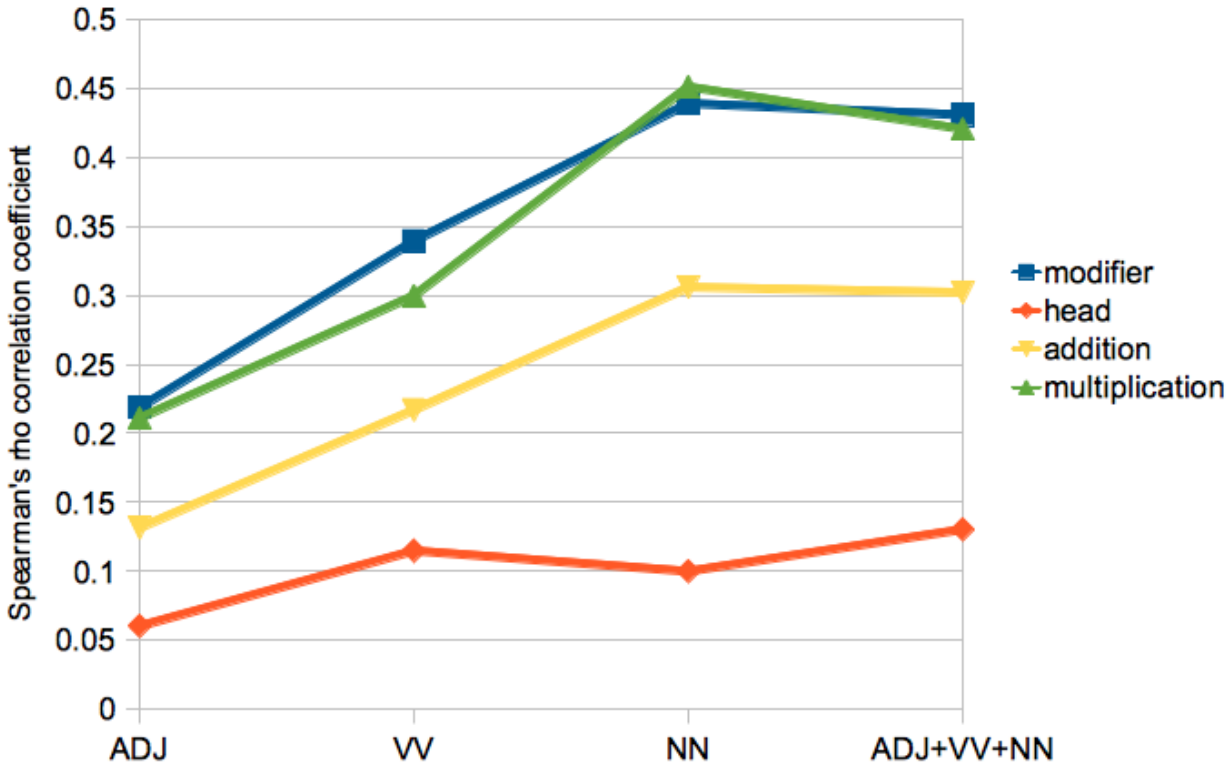

syntax-based:

- window-based > syntax-based (as before)
- compound–head > compound–modifier predictions (excp: trans. subjects)
- patterns with regard to function types vary
  (in comparison to previous models, and for modifiers vs. heads)

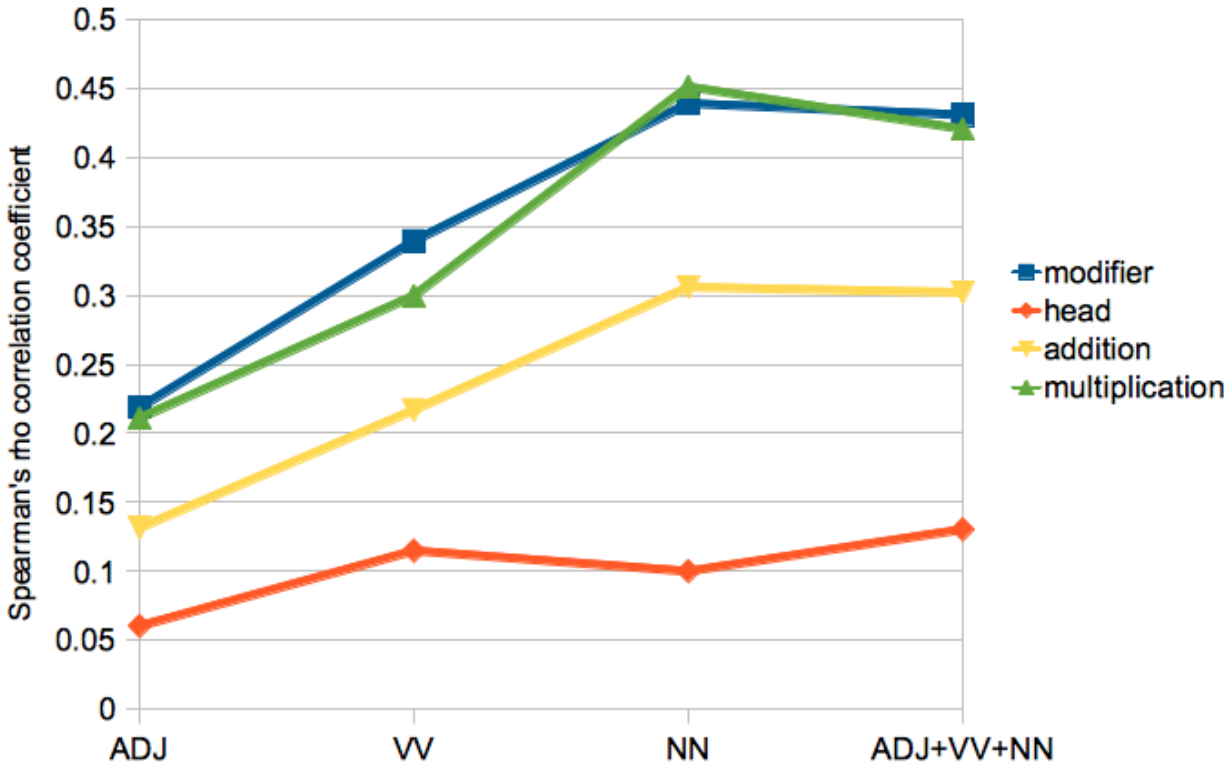
# Role of Modifiers vs. Heads (2)

- Hypothesis 2:
  distributional properties of heads are more salient than distributional properties of modifiers
- Perspective (ii): contribution of modifiers vs. heads to compound meaning
- Setup:
  - window-based, window 20, across parts-of-speech
  - correlate only one type of compound–constituent predictions with the compound whole ratings
  - apply addition/multiplication
  - correspondence to upper bound

# Role of Modifiers vs. Heads (2): Results



# Role of Modifiers vs. Heads (2): Results




- impact of distributional semantics: modifiers > heads
- multiplication ∼ modifiers only
- multiplication > addition

# Summary: Lexical Model

- Hypothesis 1 (i): against our intuition, not adjectives or verbs but nouns provided the most salient distributional information.
- Hypothesis 1 (ii): syntax-based predictions by adjective and preposition modification and by verb subcategorisation (and various concatenations) were all worse than predictions by the respective window-based parts-of-speech.
- Best Model: nouns within a 20-word window ($\rho = 0.6497$)

# Summary: Lexical Model

- Hypothesis 2 (i):
  - salient features to predict similarities between compound–modifier vs. compound–head pairs are different
  - small windows: distributional similarity between compounds and heads > compounds and modifiers; larger contexts: difference vanishes
- Hypothesis 2 (ii): influence of modifier meaning on compound meaning is stronger than influence of head meaning (a) for human ratings, and (b) according to the vector space models.
- Task: learn more about the semantic role of modifiers vs. heads in noun-noun compounds (as do Gagné and Spalding, 2009; 2011, among others).

# Multi-Modal LDA Model

- Extension of Latent Dirichlet Allocation (LDA) model relying on two-dimensional topics (Andrews et al., 2009)
- Multi-modal features:
  - Textual Modality: word–document topics relying on WebKo
  - Psycholinguistic Modality:
    1. association norms
    2. feature norms
  - Visual Modality: *BilderNetle*, German noun–ImageNet mappings for compounds and constituents
    1. SURF (clusters): selects key-points within an image
    2. GIST (clusters): computes a high-level vector for an image

# Multi-Modal LDA Model: Results

| Modality | $k$ | $\rho$ |
|---|---|---|
| **Text Only** | | |
| Text | 200 | .204 |
| **Bimodal mLDA** | | |
| Text + Feature Norms | 150 | ***.310 |
| Text + Association Norms | 200 | **.328 |
| Text + SURF | 50 | .251 |
| Text + GIST | 100 | .204 |
| Text + SURF Clusters | 200 | .159 |
| Text + GIST Clusters | 150 | .233 |
| **3D mLDA** | | |
| Text + FN + AN | 250 | .259 |
| **Hybrid Models** | | |
| (Text + FN) & (Text + AN) | 150+200 | ***.**390** |
| (Text + FN) & (Text + SURF) | 150+50 | ***.350 |
| (Text + FN) & (Text + GC) | 150+150 | ***.340 |

# Summary: German Noun-Noun Compounds

- Both experiential and distributional data provide strong evidence for the compositionality of German noun-noun compounds.
- Simple lexical co-occurrence features are impressively strong.
- Domain knowledge (provided by nominal co-occurrence) represents the overall most salient contextual knowledge.
- What are the conditions and contributions of modifier vs. head constituents with regard to compound meaning?

# German Particle Verbs

# Phenomenon

- Composition:
  - composition of base verbs (BVs) and prefix particles
  - focus: preposition particles
- Examples:
  - *abholen* 'fetch': *ab* + *holen* 'fetch'
  - *anfangen* 'begin': *an* + *fangen* 'catch'
  - *einsetzen* 'insert'/'begin': *ein* + *setzen* 'put/sit (down)'
- References:
  - Stiebels (1996); Lüdeling (2001); Dehe et al. (2002)
  - Lechler & Roßdeutscher (2009); Kliche (2011); Springorum (2011)

# Past and Ongoing Research

- Empirical subcategorisation transfer patterns at the syntax-semantics interface (Hartmann et al., KONVENS Workshop 2008)
- Particle verb clusters: distributional clusters of particle verbs and base verbs (Kühner & Schulte im Walde, KONVENS 2010)
- Particle (verb) clusters: distributional clusters of the German verb particle *an* (Springorum et al., LREC 2012)
- Systematic neologisms of particle verbs: empirically identify regularities of PV composition, based on a collection of example sentences
- Metaphorical shifts of particle verbs: identify regularities at the syntax-semantics interface that indicate metaphorical uses of particles or particle verbs (Springorum et al., IWCS 2013)

# Clustering Experiments

1. Distributional clusters of particle verbs and base verbs
   - Task: predict the degree of compositionality of the compounds

2. Distributional clusters of the German verb particle *an*
   - Task: classify the verb particle according to its senses

# (1) Particle Verb Clusters

- Hypothesis: The more compositional a particle verb is, the more often it appears in the same cluster with its base verb.
  - compositionality is restricted to the relationship between particle verbs and base verbs
  - contribution of particle is ignored
- Dataset: 99 German particle verbs across 11 particles and 8 frequency ranges plus 1 deliberately ambiguous particle verb

# Clustering

- Soft clustering:
  - cluster membership is represented by a probability
  - probabilistic membership is turned into binary membership by establishing a membership cut-off
- Clustering approaches:
  - Latent Semantic Classes (LSC) (Rooth, 1998):
    - two-dimensional soft clusters that generalise over hidden data
    - Expectation-Maximisation (EM) algorithm for unsupervised training on un-annotated data
    - model selectional dependencies between two sets of words participating in a grammatical relationship
  - Predicate-Argument Clustering (PAC) (Schulte im Walde et al., 2008):
    - extension of LSC to incorporate selectional preferences
    - combination of EM algorithm and Minimum Description Length (MDL) principle

# LSC: Example Cluster

| *dimension 1: verbs* | | *dimension 2: direct object nouns* | |
|---|---|---|---|
| schicken | 'send' | Artikel | 'article' |
| verschicken | 'send' | Nachricht | 'message' |
| versenden | 'send' | E-Mail | 'email' |
| nachweisen | 'prove' | Brief | 'letter' |
| überbringen | 'deliver' | Kind | 'child' |
| abonnieren | 'subscribe to' | Kommentar | 'comment' |
| zusenden | 'send' | Newsletter | 'newsletter' |
| downloaden | 'download' | Bild | 'picture' |
| bescheinigen | 'attest' | Gruß | 'greeting' |
| zustellen | 'send' | Soldat | 'soldier' |
| abschicken | 'send off' | Foto | 'photo' |
| zuschicken | 'send' | Information | 'information' |

# PAC: Example Cluster

| *dimension 1: verbs* | | *dimension 2: WN concepts over PP arguments* | |
|---|---|---|---|
| steigen | 'increase' | Maßeinheit | 'measuring unit' |
| zurückgehen | 'decrease' | e.g., Jahresende | 'end of year' |
| geben | 'give' | Geldeinheit | 'monetary unit' |
| rechnen | 'calculate' | e.g., Euro | 'Euro' |
| wachsen | 'grow' | Transportmittel | 'means of transportation' |
| ansteigen | 'increase' | e.g., Fahrzeug | 'automobile' |
| belaufen | 'amount to' | Gebäudeteil | 'part of building' |
| gehen | 'go' | e.g., Dach | 'roof' |
| zulegen | 'add' | materieller Besitz | 'material property' |
| anheben | 'increase' | e.g., Haushalt | 'budget' |
| kürzen | 'reduce' | Besitzwechsel | 'transfer of property' |
| stehen | 'stagnate' | e.g., Zuschuss | 'subsidy' |

# Clustering Setup

- Corpus:
  - data: *SdeWaC*, parsed with *FSPar* (Schiehlen, 2003)
  - 2,152 verb types with $1,000 < \text{freq} < 100,000$, plus targets
- Distributional features:
  - nominal features
  - syntactic functions: subjects, objects, pp objects
  - incorporating vs. excluding the notion of syntax
- Clustering parameters:
  - number of clusters: 20, 50, 100, 200
  - probability thresholds: 0.01, 0.001, 0.0005, 0.0001

# Experiential Data and Evaluation

- Human ratings on the degree of compositionality:
  - scale: 1 (definitely opaque) to 10 (definitely transparent)
  - data: rating means
  - examples:
    - *nachdrucken* 'reprint': 9.250
    - *aufhängen* 'hang up': 8.500
    - *nachweisen* 'prove': 5.000
    - *zutrauen* 'feel confident': 3.250
    - *umbringen* 'kill': 1.625
- Evaluation: proportion of PV–BV cluster co-occurrence $comp_{pv}$ against human ratings using Spearman's $\rho$:

$$comp_{pv} = \frac{\sum_c p(pv, c) \geq t \ \wedge \ p(bv, c) \geq t}{\sum_c p(pv, c) \geq t} \quad (1)$$

# Results

**LSC**:

| | best result | | | analysis | | membership |
|---|---|---|---|---|---|---|
| input | corr | cov | f-score | clusters | iter | threshold |
| obj | **.433** | .59 | **.499** | 100 | 200 | .0005 |
| subj | .205 | .76 | .323 | 50 | 200 | .0001 |
| pp | **.498** | .40 | **.444** | 20 | 200 | .0005 |
| n+syntax | .303 | .54 | .388 | 50 | 200 | .0005 |
| n-syntax | .336 | .56 | .420 | 100 | 200 | .001 |

**PAC**:

| | best result | | | analysis | | membership |
|---|---|---|---|---|---|---|
| input | corr | cov | f-score | clusters | iter | threshold |
| obj | .100 | .53 | .168 | 100 | 50 | .0005 |
| subj | .783 | .05 | .094 | 20 | 50 | .01 |
| pp | .275 | .21 | .238 | 200 | 100 | .01 |
| n+syntax | **.213** | .61 | **.316** | 20 | 100 | .0001 |
| n-syntax | **.236** | .53 | **.327** | 200 | 100 | .001 |

# (2) Particle Clusters

- Theoretical classification: *an* belongs to 11 semantic classes (Springorum, 2009; 2011)
- Gold standard classification: subset of four semantic classes
  1. Topological verbs: contact situation between a direct object of the *an* particle verb and an implicit background.
     *Maria kettet den Hund an*. 'Maria chains the dog.'
  2. Directional verbs: verb event points from the subject to the direct object of the *an* particle verb.
     *Maria lächelt ihre Mutter an.* 'Maria smiles at her mother.'
  3. Event initiation verbs: the *an* particle contributes a change from a non-progressive state to a progressive state.
     *Opa heizt den Ofen an.* 'Grandfather heats up the oven.'
  4. Partitive verbs: event is performed only on parts of the direct object.
     *Der Dachdecker sägt das Brett an.* 'The roofer saws at the plank.'

# Classification Setup

- Corpus and verbs:
  - data: part of *SdeWaC*, parsed with *FSPar*
  - 40 *an* particle verbs (10 from each class)
- Distributional features:
  - prepositional heads of prepositional phrases
  - direct objects and their GermaNet generalisations
  - subjects (baseline)
- Classification approach:
  - WEKA J48 decision tree algorithm with pruned trees

# Results

| Experiment | Feature | Accuracy | | Top. | Ev.I. | Dir. | Par. |
|---|---|---|---|---|---|---|---|
| Baseline | Subject | 13 | 32.50% | 0 | 3 | 1 | 9 |
| Judgements | | | 79.06% | | | | |
| Exp. 1 | PPs | 25 | 62.50% | 6 | 5 | 5 | 9 |
| Exp. 2 | Objects | 11 | 27.50% | 0 | 0 | 2 | 9 |
| Exp. 3 | Object Classes | 27 | 67.50% | 1 | 8 | 8 | 10 |
| Exp. 4 | *an*+Object Classes | 28 | 70.00% | 4 | 7 | 7 | 10 |

# Systematic Neologisms: Goals and Data

- Research questions:
  - Are German particle verbs compositional?
  - Are there any (prototypical) particle readings?
  - What is the meaning contribution of the base verbs?
- Dataset: 125 German particle verbs across 5 particles and 5 semantic base verb classes
  - particles: *ab*, *an*, *auf*, *aus*, *nach*
  - semantic verb classes:
    1. DE-ADJECTIVAL e.g. *kürzen* 'shorten'
    2. ACHIEVEMENT/ACCOMPLISHMENT e.g. *finden* 'find'
    3. PHYSICAL PROCESS e.g. *stricken* 'knit'
    4. MENTAL PROCESS e.g. *denken* 'think'
    5. STATE e.g. *lieben* 'love'

# Experiment: Task and Example Sentences

- Task: generation of sentences with attested PVs and with systematic neologisms of German particle verbs

# Experiment: Task and Example Sentences

- Task: generation of sentences with attested PVs and with systematic neologisms of German particle verbs

- Examples:

  *Er hatte an der Wand angelauscht und wusste Bescheid.*
  'He had listened at the wall and knew it all.'

  *Ich musste mich noch lange Zeit nachwundern.*
  'I was wondering about it for a long time.'

  *Ich muss meine Mülltonne anleeren.*
  'I have to start emptying my bin.'

  *Ich werde den Zombie schon mal antöten, damit du ihn erledigen kannst.*
  'I will kill at the Zombie, so that you can execute him.'

# Metaphorical Shifts: Hypothesis and Examples

Hypothesis: There are regular mechanisms wrt the syntax-semantic interface

- that trigger meaning shifts of a base verb in combination with a particle meaning and
- that apply across a semantically coherent set of verbs.

$$\texttt{BV}\ \{pBV_1, pBV_2, \ldots, pBV_n\}\ \texttt{+}\ \texttt{PM} \rightarrow \texttt{PV}\ \{pPV_1, pPV_2, \ldots, pPV_m\}$$

Meaning shift classes (examples):

1. AN: **"positive directed communication"**
   BV {pleasing, emission} + PM {dir+com} → PV {pos. dir. communic.}
   with BVs *funkeln, grinsen, lächeln, strahlen*

2. AUF: **"negative social pressure"**
   BV {loud/heavy pressure} + PM {vert. contact} → PV {neg. soc. pressure}
   with BVs *brummen, bürden, donnern, lasten, zwängen*

# Metaphorical Shifts: Data Basis

- Task: identify regularities in distributional features that indicate metaphorical uses of particles or particle verbs

- Basis: corpus information on subcategorisation frames and nominal complements

$$\texttt{BV}\ \{pBV_1, pBV_2, \ldots, pBV_n\}\ \texttt{+}\ \texttt{PM} \rightarrow \texttt{PV}\ \{pPV_1, pPV_2, \ldots, pPV_m\}$$

# Metaphorical Shifts: Data Basis (Example)

| base verbs | frames | complements | connotations | properties |
|---|---|---|---|---|
| *strahlen* 'beam' | intrans | *Sonne* 'sun'<br>*Auge* 'eye' | bright, warm | light emission |
| *funkeln* 'twinkle' | intrans | *Sternlein* 'little star'<br>*Auge* 'eye' | pleasing, valuable | |
| *lächeln* 'smile' | intrans | *Mädchen* 'girl' | happy, friendly | positive emotion |
| *grinsen* 'grin' | intrans | *Freund* 'friend' | expression | |

| particle verbs | frames | complements | connotations | properties |
|---|---|---|---|---|
| *anstrahlen* 'beam at' | trans | *Decke* 'ceiling' | | pos. directed communication |
| *anfunkeln* 'beam at' | trans | *Muffel* 'grumpy person'<br>*Großmaul* 'loudmouth' | pleasing, positive | |
| *anlächeln* 'smile at' | trans | *Mädchen* 'girl' | communication | |
| *angrinsen* 'grin at' | trans | *Mädchen* 'girl' | | |

# Summary: German Particle Verbs

- Up to now:
  - various small-scale experiments to explore particle (verb) meaning
  - successful distributional models
- *SemRel* + SFB: systematic analyses
  - clusters of PVs, BVs and particles
  - meaning aspects and distributional features of PVs, BVs and particles
  - regularities and irregularities in syntax-semantics subcategorisation transfer

# Summary

1. Degree of compositionality of compounds → various models
2. Contribution of meaning aspects of constituents to compound meaning → various models, to be continued
3. Role of ambiguity → future work
4. Specification of gold standards → identify suitable models

5. Role of modifier vs. head constituents (in noun-noun compounds) → first study done; to be continued
6. Metaphorical shifts of particle (verb) meaning aspects → just started

# *SemRel/IMS* Team working on Compounds

- Natalie Kühner (Studienarbeit)
- Stefan Müller (Studienarbeit)
- Stephen Roller (PhD)
- Sylvia Springorum (PhD)

- Antje Roßdeutscher (Senior Researcher)
- Jason Utt (PhD)