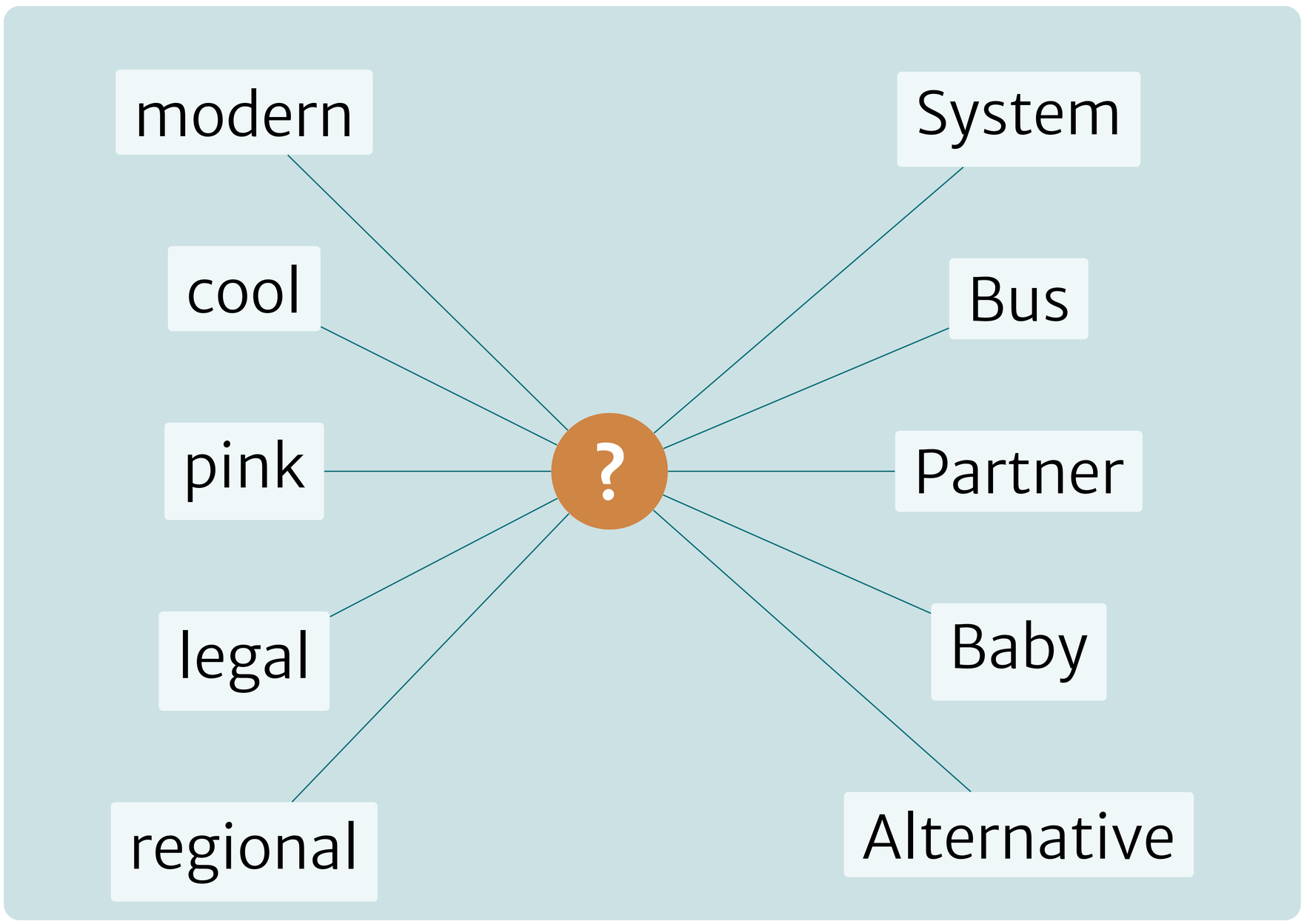
# AET: Web-based adjective exploration tool for German

*Tatiana Bladier, Esther Seyffarth, Oliver Hellwig, Wiebke Petersen*
Heinrich Heine University, Düsseldorf

aet.phil.hhu.de



## Adjectives as modifiers

In German, adjectives can modify nouns, verbs, or other adjectives. Corpus evidence for **modifier-modifiee** patterns is useful for linguistic research in various areas, including:

- semantic compositionality
- synonymy and polysemy
- semantic clustering
- distributional properties of adjectives and adverbials
- ...

**AET** is a new web-based query tool that can be used to collect corpus evidence according to complex syntactic or morphological criteria.

Use **AET** to answer questions like:

- Which denominal adjectives occur in the corpus?
- Which adjectives never occur in inflected forms?
- Which adjectives with a particular suffix occur at least 50 times in the corpus?

## Our contribution: AET

**AET** is designed to provide information on modifier-modifiee pairs that goes beyond the mere surface forms found in a corpus. The types of filter criteria currently supported include:

- Morphological: How is a modifier or modifiee derived morphologically?
- Syntactic: Which part-of-speech tag does the modifier or modifiee have?
- Lexical: In which surface forms does the lemma of the modifier or modifiee occur?
- Character-based: Which character sequence occurs at the start or end of a modifier or modifiee?
- Statistical: How often does the modifier or modifiee occur in the corpus?

Pairs are retrieved from a dependency-parsed corpus, so that modifiers and modifiees do not need to be adjacent in the text:

Der Kuchen **schmeckt**$_{V}$ wegen der verwendeten
the cake **tastes**$_{V}$ due to the used
Zuckerart **gut**$_{ADJ}$.
sugar type **good**$_{ADJ}$

## AET facts

- Relational MySQL database
- Website written in CakePHP
- Search language parser written in Python
- Corpus from the Workshop on Machine Translation 2014 (WMT14) shared tasks
- Sentences parsed with MATE
- Morphological analyses added with DerivBase and CELEX
- Demo version: 1,248,869 modifier-modifiee token pairs, 616,447 type pairs, 1,250,284 sentences
- Full database: 13,274,424 token pairs, 3,738,414 type pairs, 8,206,827 sentences

## References


**Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX Lexical Database.**

**Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A High-performance Syntactic and Semantic Dependency Parser.**

**Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation.**

**Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German.**

See proceedings for full bibliography.

## How to use AET

- Form complex queries using the boolean operators AND, OR and NOT.
- More specific queries execute faster.
- Results are sorted by lemma pair frequency.
- Select a lemma pair to view surface forms in individual sentences.
- Export results to `*.csv` to work with data offline.


## Properties of word forms

| | |
|---|---|
| word form | vertretbaren |
| lemma | vertretbar |
| word type | adjective |
| derivation type | deverbal |
| derivation scheme | Vx |
| prefix | ver |
| suffix | bar |
| derivation tree | (((ver)[V\|.V],(tret)[V])[V], (bar)[A\|V.])[A] |
| prev. derivation step | vertret+bar |
| composition | non-compositum |
| gradation | positive |
| frequency in AET | 62 |
| number | plural |
| case | dative |
| gender | feminine |
| never inflected | no |
| occurs sentence-final | yes |

Tab. 1: Searchable DB fields in AET for word forms, here: *vertretbaren*

## Properties of modifier–modifiee pairs

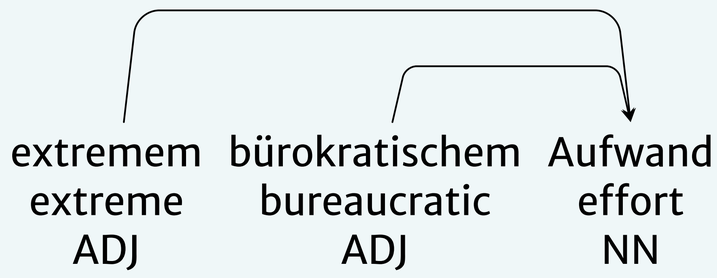
| | |
|---|---|
| modifier word form | benötigten Mittel |
| modifiee word form | Mittel |
| modifier lemma | benötigt |
| modifiee lemma | Mittel |
| modifier POS tag | ADJA |
| modifiee POS tag | NN |
| POS category pair | AN |
| pair frequency | 56 |
| precedence | yes |

Tab. 2: Searchable DB fields in AET for modifier–modifiee pairs, here: *benötigten Mittel*

## Future work

Adjectives sometimes occur in chains, where two or more adjectives modify the same token.

So far, AET represents each instance of modification in such cases as an individual modifier–modifiee pair. We plan to add search terms for filters on the chain level.




AET separates functionality from data: Other languages or other corpora can be introduced with minimal effort.

The architecture can also accomodate other types of modification, such as nominal compounds or prepositions and their complements.

Additional fields can be included in the database with little effort. The search language can easily be updated or extended as necessary.

## Selection of search terms

| | |
|---|---|
| modifier_lemma(blau) | Find all pairs whose modifier is a form of the lemma *blau*. |
| modifiee_lemma(Haus) | Find all pairs whose modifiee is a form of the lemma *Haus*. |
| modifier_wordform(rotem) | Find all pairs whose modifier is the word form *rotem*. |
| modifiee_wordform(Kreuzen) | Find all pairs whose modifiee is the word form *Kreuzen*. |
| modifier_lemma_starts(ab) | Find all pairs whose modifier lemma starts with the character sequence *ab*. |
| modifiee_lemma_starts(be) | Find all pairs whose modifiee lemma starts with the character sequence *be*. |
| modifier_wordform_starts(ver) | Find all pairs whose modifier word form starts with the character sequence *ver*. |
| modifiee_wordform_starts(ent) | Find all pairs whose modifiee word form starts with the character sequence *ent*. |
| modifier_lemma_ends(lich) | Find all pairs whose modifier lemma ends with the character sequence *lich*. |
| modifiee_lemma_ends(ung) | Find all pairs whose modifiee lemma ends with the character sequence *ung*. |
| modifier_wordform_ends(ende) | Find all pairs whose modifier is a word form that ends with the character sequence *ende*. |
| modifiee_wordform_ends(ungen) | Find all pairs whose modifiee is a word form that ends with the character sequence *ungen*. |
| modifier_pos(a) | Find all pairs whose modifier is an adjective. |
| modifiee_pos(v) | Find all pairs whose modifiee is a verb. |
| modifier_pos_stts(ADJA) | Find all pairs whose modifier is tagged with the ADJA tag. |
| modifiee_pos_stts(VVINF) | Find all pairs whose modifiee is tagged with the VVINF tag. |
| modifier_derivationtype(deverbal) | Find all pairs whose modifiers is known to be morphologically derived from a verb. |
| modifiee_derivationtype(denominal) | Find all pairs whose modifiee is known to be morphologically derived from a noun. |
| modifier_never_inflects(true) | Find all pairs whose modifier takes only one surface form in the corpus. |
| modifier_always_positive(true) | Find all pairs whose modifier always takes the positive form in the corpus. |
| pair_type_frequency_greater(100) | Find all pairs whose modifier and modifiee lemma occur together more than 100 times in the corpus. |
| pair_type_frequency_lower(100) | Find all pairs whose modifier and modifiee lemma occur together less than 100 times in the corpus. |


aet.phil.hhu.de

petersen@phil.uni-duesseldorf.de

http://www.sfb991.uni-duesseldorf.de/en/


Poster presenter: Esther Seyffarth


DFG Deutsche Forschungsgemeinschaft

This work is funded by the German Research Foundation (DFG) as part of the collaborative research center SFB 991 *The Structure of Representations in Language, Cognition, and Science*. The authors wish to thank Sebastian Löbner, Curtis Anderson, Ekaterina Gabrovska, Kai Koch, and three anonymous reviewers.