AET: Web-based Adjective Exploration Tool for German

Tatiana Bladier, Esther Seyffarth, Oliver Hellwig, Wiebke Petersen
Heinrich Heine University of Düsseldorf, Germany
bladier@phil.hhu.de, seyffarth@phil.hhu.de, hellwig7@gmx.de, petersen@phil.hhu.de

Abstract

We present a new web-based corpus query tool, the Adjective Exploration Tool (AET), which enables research on the modificational behavior of German adjectives and adverbs. The tool can also be transferred to other languages and modification phenomena. The underlying database is derived from a corpus of German print media texts, which we annnotated with dependency parses and several morphological, lexical, and statistical properties of the tokens. We extracted pairs of adjectives and adverbs (*modifiers*) as well as the tokens modified by them (*modifiees*) from the corpus and stored them in a way that makes the modifier-modifiee pairs easily searchable. With AET, linguists from different research areas can access corpus samples using an intuitive query language and user-friendly web interface. AET has been developed as a part of a collaborative research project that focuses on the compositional interaction of attributive adjectives with nouns and the interplay of events and adverbial modifiers. The tool is easy to extend and update and is free to use online without registration: http://aet.phil.hhu.de

Keywords: Corpus-query tool, adjectives, adjective modification, German, syntactic dependencies

1. Introduction

Current research on semantic compositionality, relatedness, clustering and distributional properties of adjectives and adverbials in German (Eroms, 2011; Hartung and Frank, 2011; Dalmas et al., 2015; Petersen and Hellwig, 2016) requires access to corpus tools which enable linguists to answer complex questions about the behavior of the adjectives in their context. These questions may, for instance, relate to the syntactic relationships or morphological properties of the adjectives at hand. Currently available online corpus query tools for German include the Stuttgarter IMS WORKBENCH tool (Evert and Hardie, 2011), COSMAS II (Kupietz and Keibel, 2009) and DWDS (Klein and Geyken, 2010). These and other resources provide data that is annotated on several linguistic levels, but do not lend themselves specifically to research on adjectives. In particular, we are not aware of any freely accessible corpus query tools that offer the possibility to easily filter the corpora according to complex syntactic or morphological (e. g. derivational) criteria.

In this paper, we present a user-friendly, web-based corpus query tool designed for linguists researching different aspects of usage patterns of adjectives and adverbs in German. The underlying corpus of our Adjective Exploration Tool (AET) is based on a subset of the corpora from the Workshop on Machine Translation 2014 (WMT14) shared tasks (Bojar et al., 2014)[1] and contains texts from German print media. In addition to the corpus text itself, the AET database also contains data on syntactic dependencies, word co-occurrences, morphological properties of the occurring tokens, and frequency statistics for the adjectives and adverbs. We store dependency-parsed sentences in a relational MySQL database, capturing direct modification[2] relations between the *modifiers* (i.e., adjectives or adverbs) and the *modifiees* (i.e., nouns, verbs, or adjectives modified by adjectives or adverbs). AET currently contains over 28 million tokens in roughly 8 million sentences[3], from which we extracted and stored around 13 million adjective-modifiee pairs.

Users of AET can query the underlying database using a web interface and a query language. This lets them define morphological, syntactic, lexical, character-based, or statistical criteria to retrieve all samples of particular modification or co-occurence patterns, together with the sentences in which they occur in the corpus. The results from such a query can be downloaded as a `.csv` (comma-separated values) file to enable offline work and further processing of the retrieved items.

AET was originally designed in a research project concerned with modeling the compositional interaction of attributive adjectives with nouns[4] and a project concerned with the interplay of events and adverbial modifiers[5]. However, the structure of the AET database makes it easy to extend the tool to different research areas and languages or to add more corpora. AET is available freely without registration at http://aet.phil.hhu.de.

1.1. Why AET?

In order to analyze the modification patterns of German adjectives and adverbs, it does not suffice to look for simple surface-level co-occurrences in a corpus. While adjectives and the tokens they modify may occur directly adjacent to each other in German sentences, they are also often found far apart and in different orders, as shown in examples (1) to (3). This is why information on syntactic dependencies is necessary in order to identify the modification behavior independently of the order or relative distance in which the participants of that modification are observed.

[1] The file is available at http://www.statmt.org/wmt14/training-monolingual-news-crawl/news.2013.de.shuffled.gz.

[2] In this paper, we use the term *modification* as an umbrella term for the syntactic relationships in which adjectives or adverbs are involved, including *attribution* and *predication*. For more discussion on this, see (McNally and Boleda, 2004).

[3] When processing the original corpus contents, we removed the sentences that did not contain any modifier-modifiee pairs, since these sentences are not of interest in the context of AET.

[4] http://www.sfb991.uni-duesseldorf.de/en/c10/

[5] http://www.sfb991.uni-duesseldorf.de/en/b09/

(1) Der Kuchen **schmeckt$_{V}$** **gut$_{ADJ}$**.
the cake tastes$_{V}$ good$_{ADJ}$

(2) Er sagt, dass der Kuchen **gut$_{ADJ}$** **schmeckt$_{V}$**.
he says that the cake good$_{ADJ}$ tastes$_{V}$

(3) Der Kuchen **schmeckt$_{V}$** wegen der verwendeten
the cake tastes$_{V}$ due_to the used
Zuckerart **gut$_{ADJ}$**.
sugar_type good$_{ADJ}$

In all three examples above, the adjective *gut* (Engl. *good*) modifies the verb *schmeckt* (Engl. *tastes*). The search language provided by AET can be used to retrieve all modification pairs that include the lemma *gut* as the modifier, or the lemma *schmeckt* as the modifiee; but it can also answer other types of questions, for example:

- 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?

Specific questions like these can only be answered with a corpus query tool if that tool enables user-defined filtering with regard to several different kinds of linguistic information. AET provides this possibility in a fast and user-friendly way.

1.2. Related Work

In recent years, several corpus query tools have been made available with the aim of studying the collocations of certain lemmas, the most popular of which are COSMAS II (Kupietz and Keibel, 2009), DWDS (Klein and Geyken, 2010), and the IMS OPEN CORPUS WORKBENCH (CWB) (Evert and Hardie, 2011). In this section we will briefly compare AET with the mentioned tools.

COSMAS II and DWDS contain large amounts of data for German from different genres and centuries, annotated with several tagsets. While these tools can display a number of different morphological or syntactic observations about tokens in the corpora along with the search results, they do not provide an easy way for the user to include these facts in the search query.

CWB is a corpus analysis architecture developed at the University of Stuttgart. It can be used in combination with any corpus and uses CQL (Corpus Query Language) for user queries. CWB allows users to search for words of a certain part of speech which occur in a context of certain window size or within a single syntactic constituent. However, it cannot be used to query the data with regard to morphological information on the derivation processes of lemmata (i.e., deverbal, denominal, deadjectival, etc.) or to the possible syntactic positions of the adjectives, and CQL queries quickly become long and complex.

Since none of the tools described above provide the possibility to query for the properties we have mentioned, the example questions given in Section 1.1. cannot be answered using the already existing tools. AET aims at filling this gap in the available resources.

2. Data Sources and Processing

The AET corpus is derived from a German subset of the WMT14 corpus (Bojar et al., 2014) that contains samples from German online newspapers. We used the MATE parser (Björkelund et al., 2010) to annotate the corpus with syntactic dependencies, part-of-speech (POS) tags, and lemma identification. MATE was chosen because it is among the best-performing tools for German with respect to accuracy (Choi et al., 2015). Figure 1 shows an example sentence parsed with MATE[6].

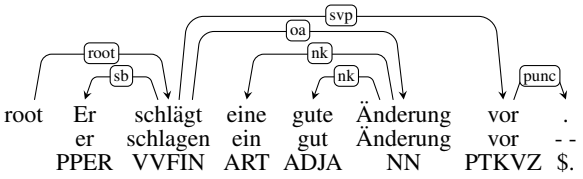

Figure 1: Dependency tree visualizing the output of MATE

We added morphological data for the tokens in the corpus from the two databases CELEX (Baayen et al., 1995) and DERIVBASE (Zeller et al., 2013) to enhance the morphological output from MATE. These databases contain detailed information on the affixation and derivation of the lemmata and word forms; CELEX additionally contains frequency counts[7]. Since CELEX provides no options to analyse out-of-vocabulary words, we use DERIVBASE as a fallback solution for those cases.

2.1. Qualitative Evaluation of the Input Data

The architecture of the database behind AET was designed to store data output from different dependency parsers. This makes the recognition of modification pairs dependent on the parser performance.

Although the MATE parser is known to be one of the best state-of-the-art parsers for German, we encountered several cases of erroneous analyses provided by it. The main source of incorrect parsing analyses are sentences in which several adjectives modifying the same token are interrupted by punctuation marks or conjunctions. Figure 2 shows a case of an erroneous analysis resulting from the punctuation mark between the adjectives in the adjective chain. Long chains of adjectives with different modification relations (for example, with both adverbial and adjectival use of adjectives) also lead to erroneous parses.

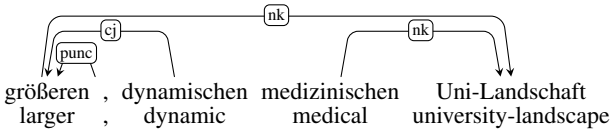

Figure 2: Erroneously recognized dependencies in MATE

[6]Morphological analyses and semantic role labeling are part of the MATE output, but are not shown here.

[7]The frequencies provided in CELEX are those observed in the Mannheim corpus of the Institut für deutsche Sprache. For more information, see (Gulikers et al., 1995).

We also encountered cases in which dependency relations provided by MATE systematically do not correspond to the correct modification relations. Among such cases are predicative uses of adjectives ("*They found the cat well-fed*") and the cases in which an adjective follows after a copula verb ("*The cat is well-fed*"). In both cases MATE recognizes the dependency relation between the verb and the adjective, but does not recognize the modification relation between the noun and the adjective. Elimination of such systematic errors requires improving the performance of the parser, a task which goes beyond the scope of the AET tool.

3. How AET Works

The main goal of AET is to let users retrieve specific types of corpus samples without needing any programming skills. A MySQL database stores the processed corpus contents, and a web interface developed with CakePHP lets the user enter queries in an easy-to-understand query language that handles the interaction with the database. A query translation module handles the generation of SQL expressions based on the user input. The following subsections describe the architecture of the database, the query language and its translation to SQL, and how the tool can be used to retrieve modifier-modifiee pairs according to user-specified criteria.

3.1. Relational Database

We created a MySQL database to store the processed data from the original corpus. The database contains information on individual tokens and lemmata in the corpus (such as morphology, word type, gradation, frequency, etc.), as well as information on the adjective-modifiee pairs that are found in the sentences. Table 1 shows the searchable fields stored for the adjective *vertretbaren* (Engl. *justifiable*).

Searchable field	Value
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]
previous 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

Table 1: Searchable fields in AET for the word form *vertretbaren*. The user can search for either the word form or the lemma.

Table 2 shows the searchable fields in the database for adjective-modifiee pairs, shown by the example pair *benötigten Mittel* (Engl. *necessary funds*).

Searchable field	Value
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

Table 2: Searchable information for the adjective-modifiee pair *benötigten Mittel*. The user can search for either surface form or lemma pairs.

Our intuitive query language enables users to interact with this (relatively complex) database structure. We have optimized the database structure in order to avoid long waiting times while the results are being collected from the server.

3.2. The Query Language of AET

One of the main criteria guided the design of AET was that the tool should be easy to use, particularly for researchers with no background in programming or computer science. The user input is parsed and turned into the corresponding SQL expression by the query translation module, which consists of a script written in Python 3.6. The relations between the search terms of the query language on the one hand and the database structure on the other hand are defined in an easy-to-edit configuration file in `.yaml` format. Therefore, search terms can be added or changed by the administrator by simply editing that configuration file.
Some of the available search terms are presented in Table 3 to give an overview of the types of queries that can be formed. Since the intended function of AET is to retrieve information about modifier-modifiee pairs, the results of each search query are always sorted with respect to pairs. As an example, consider the query in (4). It will return all modifier-modifiee pairs from the database that contain a modifier that is some form of the lemma *essbar* (Engl. *edible*).

(4) `modifier_lemma(essbar)`

After the query is submitted, the query translation component generates the corresponding SQL query for the database search. The resulting expression is this:

```
SELECT DISTINCT amtokenpair.id
FROM token_pairs amtokenpair
JOIN lemmas l1
ON amtokenpair.mer_lemma_id = l1.id
WHERE l1.lemma = 'essbar';
```

The Boolean operators AND, OR, and NOT can be used to build complex queries using the search terms that are available. Brackets can be used to indicate operator precedence between subqueries; where they are absent, natural precedence is assumed. The query in (5) is an example of the way search terms can be combined to build more specific search queries. It will retrieve all pairs from the database which

Search term	Explanation
`modifier_lemma(blau)`	Single-word modifier lemma to look for, e.g. *blau*.
`modifiee_lemma(Haus)`	Single-word modifiee lemma to look for, e.g. *Haus*.
`modifier_wordform(rotem)`	Single-word modifier word form to look for, e.g. *blaues*.
`modifiee_wordform(Kreuzen)`	Single-word modifiee word form to look for, e.g. *Kreuzen*.
`modifiee_pos(v)`	Search for modifiees of the specified part of speech, e.g. *verb*.
`modifier_lemma_starts(ab)`	(Character) prefix of a modifier lemma to look for.
`modifier_wordform_starts(ver)`	(Character) prefix of a modifier word form to look for.
`modifiee_wordform_starts(ent)`	(Character) prefix of a modifiee word form to look for.
`modifier_lemma_ends(lich)`	(Character) suffix of a modifier lemma to look for.
`modifier_wordform_ends(ende)`	(Character) suffix of a modifier word form to look for.
`modifier_derivationtype(deverbal)`	Derivation type of the modifier as analysed by the parser.
`modifier_never_inflects(true)`	Only show pairs in which the modifier is never inflected.
`pair_type_frequency_greater(100)`	Frequency filter for the pair types.

Table 3: Selection of search terms available in AET (for the full list, see documentation on website)

fulfill the constraint of containing both a non-deverbal modifier and a modifiee that is a form of the lemma *Mann*:

(5) `(NOT modifier_derivationtype(deverbal)) AND modifiee_lemma(Mann)`

Many search terms, including those listed in Table 3, have several possible spellings, in order to make it easier to remember the terms. For instance, the alternative spellings for the search term `modifiee_wordform_starts(ver)` include `mee_wf_starts(ver)` and `mee_wordform_starts(ver)`. A more detailed documentation of the query language is available on the AET website.

3.3. Searching With AET

Figure 3 shows a screenshot of the AET query input field, including a sample query. The tab labeled `Documentation` lists all available search terms when clicked. The `Submit` button sends the query to the server, which handles the translation to SQL and the collection of results from the database.

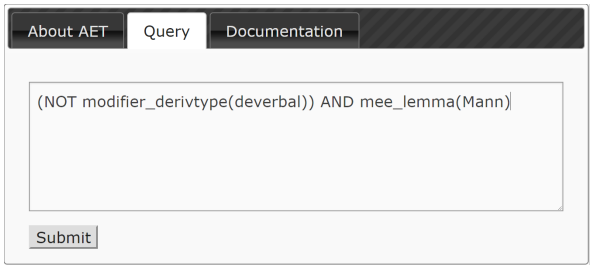

Figure 3: Query input field

Figure 4 shows an extract of the results as displayed on the website. The results are grouped by the modifier and modifiee lemmata that occur in each pair. Clicking on a lemma pair expands that section to show all sentences in the corpus that contain the given lemmata.
Clicking the `Export results` button lets the user download a `.csv` file that contains all results from the query and can be used for offline work and further processing steps.

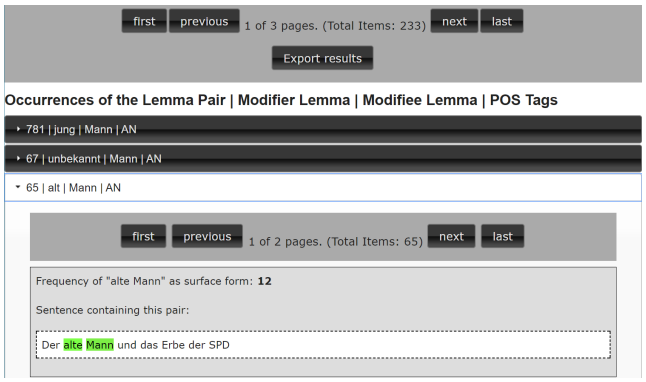

Figure 4: Representation of the search results

4. Conclusion and Future Work

AET was developed to aid linguistic researchers who are interested in the behavior and co-occurrence patterns of German adjectives, adverbs, and the words they modify. The current version of the tool provides a number of ways to filter and extract modifier-modifiee pairs from the corpus. We now outline some directions for further development.
One major aspect of the behavior of adjectives with regard to the adjective ordering has not been touched upon yet by our representation – namely, the analysis of *adjective chains*. Adjective chains are sequences of two or more adjectives modifying the same token in the sentence or each other (Dye et al., 2017). For an example, see Figure 5. Regarding the inclusion of such chains in AET, the central challenge is to find an appropriate mode of presentation for chains of different structures, lengths, and argument orders. We have explored the option of extending the database by adding more corpora in order to increase the diversity of the

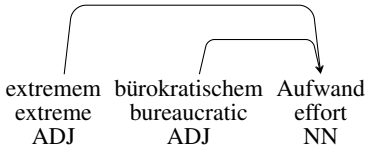

Figure 5: Example of an adjective chain

genres covered by the tool. The structure of AET makes this easy in theory, with the main practical limitations being the availability of large corpus files and the execution time of individual queries as the database grows.

We are also interested in creating a version of the tool that lets the user search for modifier-modifiee pairs in other languages. As long as inflected languages are being processed, our database structure is independent of the language of the corpus and can be reused with minimal changes to the architecture.

The tool we have developed is not restricted to adjectives and adverbs. It can also easily be used to store and query other modifications, such as nominal compounds or preposition collocations.

Regarding the translation of the user input to an SQL expression, it is simple to update only the parts of the configuration file that define the search terms which are directly affected by the changes when tables are being added or restructured; there is no need to modify the scripts which parse the user input and manage the SQL translation.

Since parsing tools for natural languages are based on different algorithms, it may be advantageous to use one or more additional parsing systems for processing the corpora in AET and to compare the parser outputs in order to decrease the number of erroneous syntactic analyses.

The methods and structures we used when designing AET were chosen specifically to enable quick and effortless extensions, updates and changes. The tool in its current state is already well suited for linguistic corpus research concerning adjective modification and co-occurence patterns of adjectives and adverbs in German. Many functionalities could still be added to AET, and the design of the tool makes it easy to do so.

Acknowledgements

This work was carried out as a part of the DFG Collaborative Research Centre 991 “The Structure of Representations in Language, Cognition, and Science” in Düsseldorf, Germany. The authors of the paper would like to thank Kai Koch for the implementation of the user interface prototype as well as Sebastian Löbner, Ekaterina Gabrovska, and Curtis Anderson for their valuable advice on the semantics of adjectives and adverbs. We would also like to thank three anonymous reviewers for their suggestions and constructive comments.

5. Bibliographical References

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database*. Linguistic Data Consortium, Univer- sity of Pennsylvania, Philadelphia, Pennsylvania.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING ’10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Choi, J. D., Tetreault, J. R., and Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *ACL (1)*, pages 387–396. The Association for Computer Linguistics.

Dalmas, M., Dobrovol’skij, D., Goldhahn, D., and Quasthoff, U. (2015). Evaluation with adjectives. towards a corpus-based approach to synonymy. *Lili - Zeitschrift fur Literaturwissenschaft und Linguistik*, 45(177):12–29.

Dye, M., Milin, P., Futrell, R., and Ramscar, M. (2017). Cute little puppies and nice cold beers: An information theoretic analysis of prenominal adjectives. CogSci 2017, 39th Annual Meeting of the Cognitive Science Society.

Eroms, H.-W. (2011). Attributive adjective clusters [Attributive Adjektivcluster]. *Deutsche Sprache*, 39(2):113–136.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium.

Gulikers, L., Rattink, G., and Piepenbrock, R. (1995). German linguistic guide.

Hartung, M. and Frank, A. (2011). Exploring supervised LDA models for assigning attributes to adjective-noun phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP’11, pages 540–551, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klein, W. and Geyken, A. (2010). Das digitale Wörterbuch der deutschen Sprache (DWDS). In *Lexicographica: International annual for lexicography*, pages 79–96. De Gruyter.

Kupietz, M. and Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. *Working papers in corpus-based linguistics and language education*, 3:53–59.

McNally, L. and Boleda, G. (2004). Relational adjectives as properties of kinds. *Empirical Issues in Formal Syntax and Semantics*, 5:179–196.

Petersen, W. and Hellwig, O. (2016). Exploring the value space of attributes: Unsupervised bidirectional clustering of adjectives in German. In *COLING*, pages 2839–2848. ACL.

Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL 2013*, pages 1201–1211, Sofia, Bulgaria.