From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions

Jörg Didakowski and Alexander Geyken

1. Introduction

In this article Wortprofil 2012 is presented, the current version of a lexical profiling tool for German based on grammatical co-occurrences. Wortprofil 2012 can be used twofold: it assists lexicographers in their work to compile collocations and it provides useful corpus-based syntagmatic information for users interested in improving their language production skills. It is implemented as an additional functionality of the lexical information system of the Digital Dictionary of the German Language (‘Digitales Wörterbuch der deutschen Sprache’, DWDS) which is accessible to all users via the internet (www.dwds.de). Wortprofil 2012 is a further development of the word profile system presented in Geyken et al. (2009).

Wortprofil 2012 provides separated co-occurrence lists for twelve different grammatical relations and links them to their corpus contexts where the node word and it’s collocate co-occur. The co-occurrence lists and their ordering are based on statistical computations over a fully-automatic annotated German corpus containing about 1.8 billion tokens.

Wortprofil 2012 can help to answer questions like 'which attributive adjectives are typically used for the noun Vorschlag ‘proposal’' or 'which active subject does a verb like ausstoßen ‘emanate’ usually take' and it can be a good starting point for looking at a specific word.

The remainder of this article is structured as follows. In section 2 we shortly describe the specific challenges that the German language provides for the extraction of syntactically relevant co-occurrences. In section 3 we comment on related work in this field. In section 4 we describe the corpus basis used for the extraction of the co-occurrence. This process together with the statistical computations are described in section 5 and 6. Finally section 7 provides some concluding remarks as well as an outlook on future work.

2. Challenges of the German Language

There are several grammatical characteristics of German which make the extraction of grammatical co-occurrences particularly challenging.

German has a variable placement of the finite verb (verb-first, verb-second and verb-final) which depends on the clause type. Additionally, German has a relatively rich morphology with case marking which allows a relatively flexible phrase ordering. These flexible orderings can cause discontinuous verb chains and they can cause separable prefix verbs where the prefix is far away from its target. Furthermore, in newspaper text it can be shown that the ambiguity rate with regard to the morphological case information is very high (Evert 2004).

These features of German make it very difficult to extract grammatical co-occurrences by sequence based formalisms if one is interested in syntactic relations on sentence level.

3. Related Work

Church/Hanks (1991) show that lexical statistics are very useful for summarizing concordance data of a corpus by representing a sorted list of the statistically most salient collocates. In order to extract collocates they use natural language processing tools like a part-of-speech tagger and a partial phrasal parser. Their statistical calculation is based on mutual information.

Kilgarriff/Tugwell (2002) build on this idea in their word sketch approach. Word sketches are “summaries of a word’s grammatical and collocational behaviour” (Kilgarriff et al., 2004) and are provided by a so called Sketch Engine, a corpus query system. For languages with fixed word order the Sketch Engine uses patterns over part-of-speech sequences to detect grammatical relations in form of a sketch grammar.

Ivanova et al. (2008) develop a sketch grammar for German. Their main problem is to achieve high precision and high recall at the same time. Their results show that richer linguistic analysis is necessary to obtain a better overall performance.

Horák/Rychlý (2009) use a robust syntactic parser in order to extract grammatical relations for the Czech Language, which has a rich morphology and a relatively free word order. The result of the deeper analysis is used as input to the Sketch Engine.

4. Corpus Selection and Corpus Size

The base of Wortprofil 2012 is the corpus collection shown in table 1. It consists of renowned German daily and weekly newspapers for which legal arrangements are obtained for the use within the DWDS project together with the balanced reference corpus 'DWDS Kernkorpus', the core of our corpus collection. The corpora range from 1900 up to now.

corpus	tokens	sentences	documents
Süddeutsche Zeitung	453,945,194	29,125,790	1,099,920
DIE ZEIT	417,422,714	23,631,230	499,520
Berliner Zeitung	242,046,373	15,951,701	869,023
DIE WELT	238,403,711	15,787,624	600,007
Der Tagesspiegel	184,202,717	10,392,257	394,465
DWDS-Kernkorpus	125,990,080	7,046,937	79,312
Bild	121,520,037	12,629,828	548,181
total	**1,783,530,826**	**114,565,367**	**4,090,428**

Table 1

In table 2 a more compact listing of the collection is presented. It shows that the major part of the collection consists of daily newspapers. The total size of the corpus bases amounts to 1.78 billion tokens in 4.09 million documents.

corpus	tokens	sentences	documents
daily newspaper	1,240,118,032	83,887,200	3,511,596
weekly newspaper	417,422,714	23,631,230	499,520
balanced corpus	125,990,080	7,046,937	79,312
total	**1,783,530,826**	**114,565,367**	**4,090,428**

Table 2

The newspaper archives contain a substantial number of duplicates, thus leaving the user with a distorted view of the Wortprofil 2012 statistics. Therefore, we decided to remove all duplicates from the whole corpus collection. After the removal our collection comprises 90,806,646 sentences and 1,594,223,632 tokens.

Linguistic preprocessing for the relation extraction comprises the following steps: the corpus texts are automatically tokenized, split up into sentences and annotated with part-of-speech tags by the moot tagger (Jurish 2003). This is the standard preprocessing of the DWDS corpora.

5. Relation Extraction

In the approach presented in this article syntactic parsing is used as backbone for the extraction of grammatical co-occurrences. That is, the extraction is divided into two tasks: first the sentences of a corpus are syntactically analysed via a robust dependency parser; second the syntactic relations of interest are extracted from the parsing results. An example of this procedure is given below.

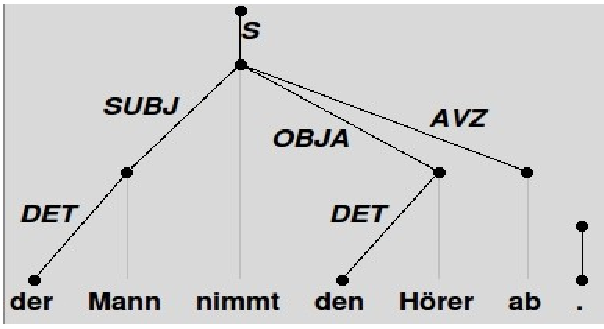


Figure 1

In figure 1 the parsing result of the sentence ‘Der Mann nimmt den Hörer ab’ (‘The man picks up the earphone.’) is shown. Via the dependency links and the dependency labels (SUBJ, OBJA, AVZ) three binary grammatical relations relevant for collocation extraction can be extracted:

- verb prefix: <nehmen ‘to pick’, ab ‘up’>
- active subject: <abnehmen ‘to pick up’, Mann ‘man’>
- accusative object: <abnehmen ‘to pick up’, Hörer ‘earphone’>

For this task, it is not important which kind of syntactic parser is chosen. It could be a dependency parser or a phrase structure parser or any other parser providing sufficient information for the relation extraction task.

5.1 The Syntax Parser

For syntactic annotation the rule based dependency parser SynCoP (Syntactic Constraint Parser, Didakowski 2008a/2008b) is used. The parser is implemented with the help of finite state techniques and is based on the high-coverage morphology TAGH (Geyken/Hanneforth 2006).

A grammar for the SynCoP parser is developed which is designed for the specific relation extraction task. Therefore, issues like the attachment of sub-clauses or specific rare syntactic phenomena are not dealt with in this grammar. Also, it does not cover long distance dependencies, it provides only weak lexicalisation information and it does not use subcategorisation information. The parser for this grammar is similar to what Grefenstette (1998) called an approximation to full parsing.

In the following a few features of the SynCoP parser are mentioned which are essential for the quality and quantity of extracted relations and for the parsing time.

The parser only assigns dependency structures which are allowed by its grammar. There is a distinction between grammatical and ungrammatical in contrast to grammarless data-driven parsers (cf. McDonald/Nivre 2011). We attempt to avoid evident annotation mistakes which emerge from phenomena such as agreement errors or uniqueness violations.

It is unrealistic to require a parser grammar to fully cover the syntactic structure of all sentences of a corpus because corpora generally contain fragments and ungrammatical sentences and because a grammar does not cover all grammatical phenomena. Therefore the parser allows partial analyses in order to annotate as much as possible. Thus the relevant phenomena for our extraction tasks can be annotated separately, including subordinate clauses, noun groups, and prepositional groups.

It is possible to assign weightings to the grammar rules. In this way lexical and structural preferences can be defined. Note that only one dependency structure for a sentence is used for the relation extraction step and that ambiguities can only be resolved by the parser with the help of heuristics. With the weighted rules it can for example be expressed that a subject of a sentence is expected to precede its corresponding finite verb. It is also possible to model more general preferences like a longest match strategy implementing the late closure parsing principle (see Didakowski 2008b).

The parser does not rely completely on results of the part-of-speech tagger but uses them as preferences only. This approach is compatible with the observation that part-of-speech tagging is a considerable source of errors (Kilgarriff et al. 2004).

The corpus collection used in Wortprofil 2012 comprises a large amount of data and has to be annotated in adequate time. Therefore, the parser implements some mechanisms that speed up parsing. Left and right embeddings of sentences are replaced by iteration and centre embeddings of sentences are restricted to a depth of one. This approach follows the observation that there exists an absolute limit on centre-embeddings in written and spoken language (Karlsson 2010).[1] Furthermore, pruning is performed if the search space is too large to parse a sentence completely.

5.2 The Covered Syntactic Relations

The syntactic annotation of our corpus collection is followed by the extraction of binary and ternary syntactic relations from the parsing analyses. Ten different syntactic binary relations are extracted. These relations are listed in table 3. In the second column of table 3 tuples of

[1] Karlsson (2010) studies different types of recursion and iteration in written and spoken language empirically and examines empirical determinable constraints on the number of recursive and iterative cycles.

parts-of-speech are shown which are included by a syntactic relation. The coordination is the sole symmetric relation and the verb prefix relation emerges from morphological analysis and syntax analysis.

syntactic relation	**part-of-speech tuples**
accusative object	{<verb,noun>}
active subject	{<verb,noun>}
adjective attribute	{<noun,adjective>}
coordination	{<verb,verb>,<noun,noun>,<adjective,adjective>}
dative object	{<verb,noun>}
genitive attribute	{<noun,noun>}
modifying adverbial	{<verb,adverb>,<adjective,adverb>}
passive subject	{<verb,noun>}
predicative complement	{<noun,noun>,<noun,adjective>}
verb prefix	{<verb,prefix>}

Table 3

In addition two ternary relations are covered. They are listed in table 4.

syntactic relation	**part-of-speech tuples**
comparative conjunction	{<noun,conjunction,noun>,<verb,conjunction,noun>}
prepositional group	{<noun,preposition,noun>,<verb,preposition,noun>}

Table 4

To parse one corpus of our corpus collection and to extract the syntactic relations took five days in average on five processors working in parallel. The overall extraction took about one month.

Frequency information about the extracted relations is given in table 5 in the second column. In total 381 million relations are extracted from the annotated corpora.

5.3 Filtering of Extracted Syntactic Relations

In some cases the parser produces systematic errors. In order to increase the quality of the extracted relations a filter is applied taking the context into account in which an extracted relation is found. In this approach we distinguish between safe and unsafe relations depending on the context. That is, safe syntactic relations must be contained in a safe context. On the basis of this assumption a threshold calculation is applied. All relations are removed if they do not occur at least twice in a safe context.

In the following an example for an unsafe context is given. The parser has a problem with the differentiation of active perfect and static passive. This is shown by the sentences 1 and 2:

1 Der Mann ist gerudert. (active perfect)
’The man had paddled’

2 Der Baum ist gefällt. (static passive)
’The tree is felled’

Because no subcategorisation information is used in the grammar, the parser cannot distinguish between the two different analyses because the parser has to know whether the verb is transitive or intransitive. Therefore, both sentences would get the same annotation and this would cause the problem of quality in the extraction of the syntactic relations 'active subject' and 'passive subject'.

There are also some more general criteria the context of safe relations has to meet:

- Nouns within a requested syntactic relation must have a determiner.
- All words involved in a requested syntactic relation have to be within an analysable subordinate clause or main clause.
- A sentence must start with an uppercase letter and must end with a final punctuation mark.
- A sentence must not contain any unknown word.

The frequency information about the extracted relations after the application of the filter is given in table 5 in the third column. In total 257 million relations remain.

syntactic relation	frequency	frequency (applied filter)
prepositional group	93,640,099	45,165,932
adjective attribute	68,658,904	58,297,991
modifying adverbial	63,341,241	45,392,107
active subject	51,968,759	37,824,884
accusative object	29,458,909	19,251,695
Coordination	21,886,952	21,685,018
genitive attribute	22,051,327	10,975,398
verb prefix	8,488,938	8,142,960
predicative complement	7,531,807	3,895,181
dative object	5,263,222	2,685,383
passive subject	5,017,454	2,784,627
comparative conjunction	3,806,860	1,300,991
total	**381,114,472**	**257,402,167**

Table 5

6. Statistical Computations

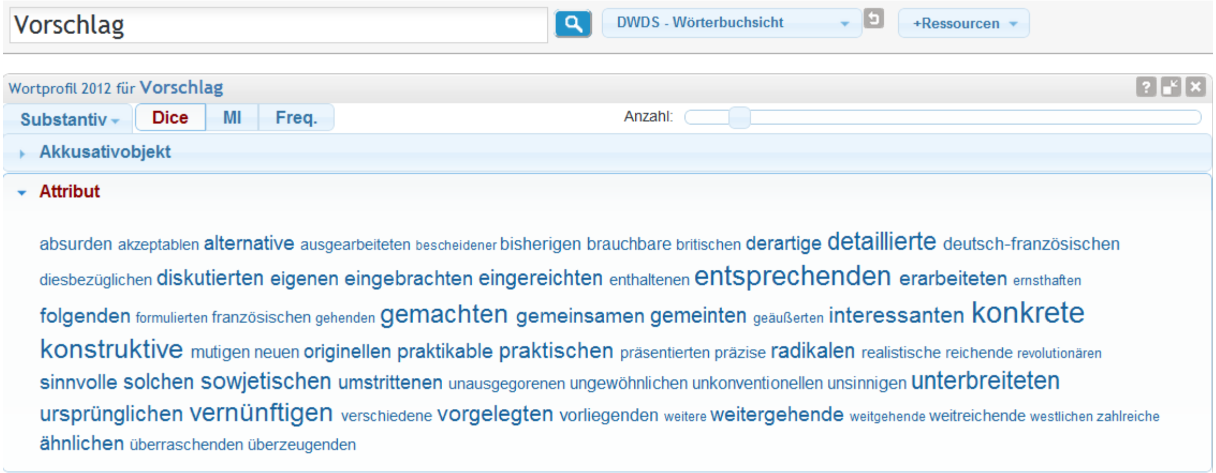
After the relation extraction a statistical computation is applied in order to determine the attraction between words of a specific syntactic relation.

Two statistics are used in Wortprofil 2012: logDice which is based on the Dice coefficient (see Rychlý 2008) and MI-log-Freq which is based on mutual information (see Kilgarriff/Tugwell 2002). These statistics can be used as a quantitative measure of the attraction between words where high scores indicate high correlation and where a distinction between negative (<0) and positive (>0) association is made (see Evert 2008).

In order to generate a sorted candidate set of collocations a ranking approach and a threshold approach are combined. First the candidate set is determined by using a threshold value of zero for logDice and MI-log-Freq and the threshold 5 for absolute frequency of word-occurrence is set. Then the candidate set is sorted by their association score, thus providing an ordering by collocational strength.

As a result of the statistical computations the database contains 11,980,910 different collocation candidates of specific syntactic types where it is possible to query 104,704 different lemma/part-of-speech pairs.

An example for a query in Wortprofil 2012 is shown in figure 2. It shows the query results for the noun ‘Vorschlag’ (‘recommendation’) with the syntactic relation ‘adjective attribute’. The result set is presented as a word cloud where the font size of the adjectives occurring with the query word correlates with the score of the statistical computations.



absurden akzeptablen alternative ausgearbeiteten bescheidener bisherigen brauchbare britischen derartige detaillierte deutsch-französischen diesbezüglichen diskutierten eigenen eingebrachten eingereichten enthaltenen entsprechenden erarbeiteten ernsthaften folgenden formulierten französischen gehenden gemachten gemeinsamen gemeinten geäußerten interessanten konkrete konstruktive mutigen neuen originellen praktikable praktischen präsentierten präzise radikalen realistische reichende revolutionären sinnvolle solchen sowjetischen umstrittenen unausgegorenen ungewöhnlichen unkonventionellen unsinnigen unterbreiteten ursprünglichen vernünftigen verschiedene vorgelegten vorliegenden weitere weitergehende weitgehende weitreichende westlichen zahlreiche ähnlichen überraschenden überzeugenden

Figure 2

7. Conclusion

In this paper we have presented Wortprofil 2012, a lexical profiling tool for the extraction of statistically salient and syntactically relevant co-occurrences. We have presented the corpus base as well as the extraction method that relies on syntactic and statistical analysis. We have also shown that additional heuristics to reduce syntactic complexity as well as computational optimizations such as the introduction of thresholds are necessary in order to obtain satisfying results in a reasonable time.

Further work will focus on the evaluation of the results and on the evaluation of the heuristics and thresholds. In addition we will implement a more intuitive user interface in order to make the Wortprofil 2012 useful for a broader public. Finally, we will apply the Wortprofil method to a diachronic corpus of German.

8. References

Didakowski, J. 2008a. ‘Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers’. In A. Storrer et al. (eds.), *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing*. KONVENS 2008. Berlin: Mouton de Gruyter, 65-78.

Didakowski, J. 2008b. ‘SynCoP – Combining Syntactic Tagging with Chunking Using Weighted Finite State Transducers’. In T. Hanneforth and K.-M. Würzner (eds.), *Finite-State Methods and Natural language Processing*. 6th International Workshop. FSMNLP 2007. Universitätsverlag Potsdam, 107-118.

Church, K. and Hanks, P. 1991. ‘Word Association Norms, Mutual Information and Lexicography’. *Computational Linguistics*. vol. 16, no. 1, 22-29.

Evert, S. 2008. ‘Corpora and collocations’. In A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook of the Science of Language and Society*. article 58. Berlin/New York: Mouton de Gruyter.

Evert, S. 2004. ‘The Statistical Analysis of Morphosyntactic Distributions’. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. LREC 2004. Lisbon, Portugal, 1539-1542.

Kilgarriff, A. et al. 2004. ‘The Sketch Engine’. In *EURALEX 2004 Proceedings.* Lorient, France, 105-116.

Kilgarriff, A. and Tugwell, D. 2002. ‘Sketching words’. In M.-H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* EURALEX, 125-137.

Geyken, A. et al. 2009. ‘Generation of word profiles for large German corpora’. In Y. Kawaguchi et al. (eds.) *Corpus Analysis and Variation in Linguistics*. Tokyo University of Foreign Studies, Studies in Linguistics 1. John Benjamins Publishing Company, 141-157.

Geyken, A. and T. Hanneforth 2006. ‘TAGH: a Complete Morphology for German Based on Weighted Finite State Automata’. In A. Yli-Jyrä et al. (eds.), *Finite-state methods and natural language processing, 5th international workshop, FSMNLP 2005, Helsinki, Finland, Revised Papers*. Lecture Notes in Artificial Intelligence 4002. Berlin/Heidelberg: Springer, 55-66.

Grefenstette, G. 1998. ‘The future of linguistics and lexicographers: will there be lexicographers in the year 3000?’. In T. Fontenelle et al. (eds.), *EURALEX 1998 Proceedings.* Liège, Belgium, 25-41.

Horák, A. and Rychlý, P. 2009. ‘Discovering Grammatical Relations in Czech Sentences’. In *Proceedings of the RASLAN Workshop 2009.* Vyd. první. Brno : Masaryk University.

Ivanova, K. et al. 2008. ‘Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case’. In *Proceedings of the 6th Conference on Language Resources and Evaluation*. LREC 2008. Marrakech, Morocco.

Jurish, B. 2003. *A Hybrid Approach to Part-of-Speech Tagging*. Final report. Project ‘Kollokationen im Wörterbuch’. Berlin-Brandenburgische Akademie der Wissenschaften. 16 Sept. 2012. http://www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf.

Karlsson, F. 2010. ‘Recursion and Iteration’. In H. Hulst (ed.), *Recursion and Human Language*. Studies in Generative Grammar 104. Berlin/New York: De Gruyter Mouton, 43-47.

McDonald, R. and Nivre J. 2011. ‘Analyzing and Integrating Dependency Parsers’. In *Computational Linguistics*. vol. 37, no. 1, 197-230.

Rychlý, P. 2008. ‘A lexicographer-friendly association score’. In P. Sojka and A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing*. RASLAN 2008, 6–9.