Generation of Word Profiles on the basis of a large and balanced German corpus

Alexander Geyken, Jörg Didakowski und Alexander Siebert

- submitted to Euralex 2008 – do not cite

1. Background

Electronic corpora have been used in lexicography and the domain of language learning for more than two decades (cf. Braun et al. 2006, Sinclair 1991). Traditionally, computer platforms exploiting these corpora were based on concordances that present a word in its different contexts. However, concordances hit their limits for very large corpora where the result sets are generally too large for manual evaluation. To answer questions like 'which attributive adjectives are used for the noun *book*' or 'is the adjective *groundbreaking* more typical for *book* than *pioneering*', would require one to look at several thousand concordance lines, a quite impracticable task to do by hand. Likewise, the exclusive use of concordance lines in an attempt to answer a question like 'which objects does a verb like *hit* typically take' would be unsuitable, since one would not only have to find all the different objects of *hit* but it would also be necessary to discard all the false positives. These types of questions involve counting of co-occurrences, and, if they are linguistically motivated, collocations. The cases above are examples for collocations of a certain syntactic type, i.e. adjective-noun and verb-object collocations. The importance of describing collocations has long been acknowledged both for language learning (e.g. Hausmann 1984) as well as for lexicographic purposes (e.g. Harris 1968). Church & Hanks (1989) were the first to show that lexical statistics are useful to summarize concordance data by presenting a list of the statistically most salient collocates. More recently, databases have been built for large corpora that make use of this abstraction of concordance lines. Examples are *Lexiview*, an interactive platform for German supporting the manual work of the lexicographer (Evert et al. 2004), or the *Sketch Engine* (Kilgarriff 2004) that produces so called 'word-sketches' for languages as different as Czech, Italian or Chinese. Both approaches provide lists of the statistically most salient collocates for each grammatical relation in which the word participates.

In this paper we present the DWDS word profile system, a unified approach to the extraction of collocations for German based entirely on finite state transducers. The DWDS word profile which is described in section 2 consists of two parts: a language specific part which consists of a complete German morphology and an efficient syntax parser for German, and a language independent part that comprises a database management system for collocations and a corpus query engine together with a web interface. In section 3, we apply the DWDS word profile to a balanced German corpus of the 20th century and present some technicalities. Another experiment using the DWDS word profile with a tabloid newspaper shows that there may be significant differences between corpora, which underlines the importance of the corpus choice for language learning as well as for the construction of lexical resources.

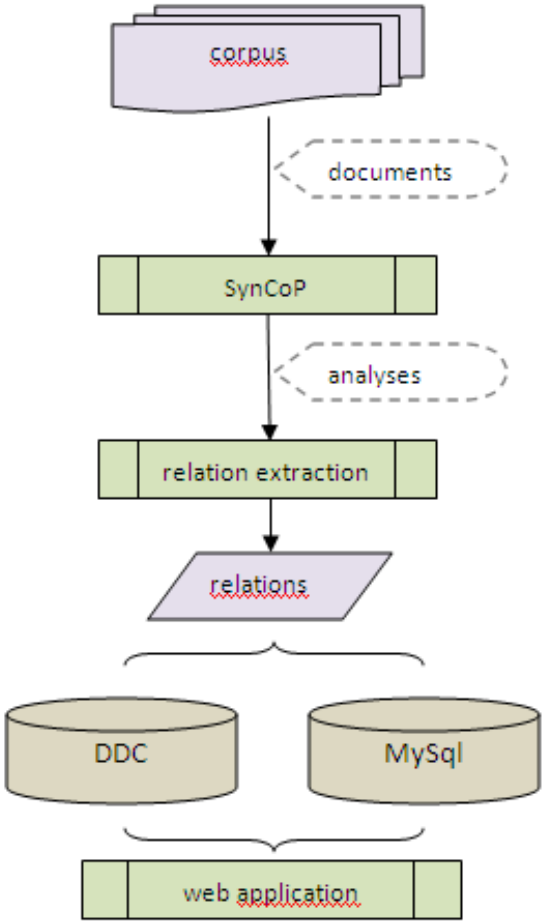
For languages with fixed word order, the Sketch Engine uses patterns over part-of-speech sequences to detect grammatical relations. For example, in order to detect verb-object pairs for English, at least for active sentences, patterns are formulated that capture a verb followed by the head noun of a noun phrase that occurs post-verbally. For languages with relatively free word order such as German, these sequence-based extraction methods to word sketches are less well suited. Kilgarriff et al. (2004) describe a Sketch Engine for Czech based on a robust deep parser for Czech. Even though the results of the parser were very precise, the parser had a problem of 'silence', i.e. it missed many of the correct relations, which resulted in word-sketches that were not very informative. The relaxation of grammar rules ended in an

approximation of syntax rules by regular patterns. The extraction of collocations in the *Lexiview* platform is performed in a hybrid way: fast chunking techniques are used for most grammatical relations; only for verb-complement extractions is a slower full probabilistic syntactic analyzer employed.

2. DWDS Word Profile System

The DWDS word profile system was implemented as an additional functionality of the DWDS website (www.dwds.de). The DWDS website – with 2.5 million p.i. (page impressions) per month – is a widely used internet platform that provides a word information system on the basis of a large monolingual German dictionary and the DWDS-Kerncorpus, a balanced corpus of German texts of the 20th century.

Input to the DWDS word profile is lemmatized text. All the texts used by the DWDS word profile have been annotated using the *TAGH* morphology (Geyken & Hanneforth 2005), a system for automatic morphological analysis of German word forms with a recognition rate of more than 99% for modern texts. A set of syntactic relations is predefined, including the aforementioned adjective-noun and the verb-object relation, as well as the genitive attribute or the separable prefix relation for complex verbs. Relations may be binary (such as adjective-noun) or ternary. An example for a ternary relation is the sequence preposition-verb-object that contains support verb constructions like *zur Verantwortung ziehen* (to hold s.o. liable) or *zur Anwendung bringen* (to apply). The extraction of the syntactic relations is realized by the *SynCoP* (Syntactic Constraint Parser), a parser which performs the syntactic dependency annotation of the corpora (Didakowski 2007) on the basis on *weighted finite state transducers* (WFSTs). SynCoP is a grammer-driven parser that combines syntactic tagging (Karlsson 1995) with chunking (Abney 1991). SynCoP is a compromise between deep and shallow parsing: on the one hand, shallow parsing is not sufficient to cope with German free word order, on the other hand deep parsing is very time consuming and not robust in the sense that sentences can't be analysed partially. SynCoP is required for a variety of different phenomena such as the resolution of case/number/gender agreement which are important to determine subject-verb relations or the recognition of verb particles which are used for the correct lemmatization of complex verbs.



The syntactic relations extracted by SynCoP are stored as tuples containing the relation name, and the collocating word forms, as well as their offsets in the text documents. For each collocation, the frequency as well as the statistic salience is computed. We use the enhanced MI statistics suggested by Kilgarriff (2004). The collocation's tuples together with its statistic information are imported into a relational database (MySql), indexed and related to the corpus sentences by their offsets. The corpus is indexed via *DDC* (Dialing DWDS Concordancer), a linguistic search engine that is used for querying the corpora on the DWDS website. Figure 1 illustrates the DWDS word profile generation process.

Figure 1: Illustration of the DWDS word profile system

The DWDS word profile system is intended to serve as an additional resource for the DWDS web-platform. Therefore, a web front-end has been implemented that visualizes the results in an intuitive way. The user can query a word form and gets back all the collocations sorted by their syntactic relations. The default view for each syntactic relation is a word-cloud where higher statistical salience is represented by larger font size. This has not only the advantage that the reader's attention is focused on the word and not on the salience values, but also that it is possible to place more syntactic relations for one word than within a tabular view. Figure 1 gives an example of the result of the verb-object and the preposition-noun-verb relation for the verb *essen* (to eat).

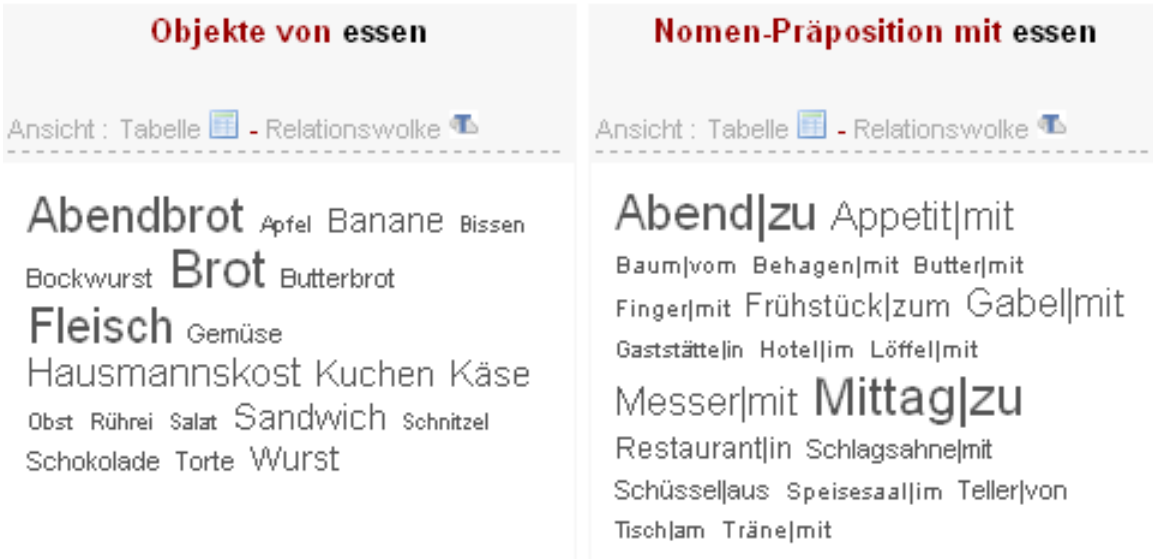


Figure 2: word-cloud for the object and the prep-noun relation for *essen* (to eat) in the DWDS/ZEIT-corpus

Additionally, the extracted relations are stored as a special index in the DDC search engine. This enables the user of the word profile system to search the entire corpus for specific patterns and filter them by syntactic functions.

3. Word Profiles for a large German corpus

As a first test of the DWDS word profile tool it was applied to two different corpora: the balanced DWDS-Kerncorpus (Geyken 2007), a 100 million token corpus of German texts of the 20th century which is equally distributed over time and over the following text types: journalism (27%), literary texts (26%), scientific literature (22%), other non-fiction (20%), as well as transcripts of spoken language (5%). The second corpus to which the word profile system was applied is a 60 million corpus of the weekly newspaper *Die ZEIT* (1997-2007). For both corpora, 34 syntactic relations using the part-of-speech (pos) categories proposed by the STTS (Schiller et al. 1999) are taken into consideration. It took 2 days on a 8-processor computer to extract 68 million syntactic relations corresponding to 1.26 million lemma, part-of-speech pairs. Only 171.000 (42.929, 8.500) lemma-pos pairs occur 10 (100, 1.000) times or more in the corpus.

4. The role of corpora

We have also generated a word profile for a 100 million tokens corpus compiled from the electronic archive of *BILD* (1997-2006), a tabloid daily newspaper that has the highest circulation of any daily German-language newspaper with more than 3.5 million copies sold daily. It can easily be verified that collocations vary between both corpora. For example, the balanced corpus has a much larger variety of collocating direct objects of the verb *übertragen* ('to transmit'), many of them corresponding to support verb constructions and hence a formal language: *Ermächtigung*, *Befugnis* (both authorization), *Aufgabe* (task), *Daten* (data), *Verantwortung* (responsibility), *Zuständigkeit* (compentency), *Eigentum* (belongings), *Vollmacht* (authority), *Kompetenz* (compentency), *Rechte* (rights) (ordered by salience, frequency >= 5). On the other hand the BILD mentions primarily concrete direct objects which are more likely to refer to events : *Spiel* (match), *Nummer* (number), *Krankheit* (disease), *Daten* (data), *Virus* (virus), *Erreger* (germ), *Verantwortung* (responsibility), *Kampf* (fight), *Veranstaltung* (event), (ordered by salience, frequency >= 3).

This variation in word profiles indicates that word profiles obtained from different corpora could be applied in different user scenarios : the comparatively balanced DWDS/ZEIT corpus is more appropriate for native speakers or professional writers whereas the *BILD* corpus is useful for foreign language learners or learners who want to be familiar with colloquial German. Indeed a preliminary study shows that collocations extracted from the *BILD* have been proved to be useful for language teaching in class courses in Italy (Bolla and Drumbl in press).

5. Conclusion and Future work

We have presented the DWDS word profile system, a collocation extraction tool for German that represents a unified approach to the extraction of collocations for German based entirely on Weighted Finite State Transducers. The system is intended as an additional information source for the DWDS web-platform. Future work will focus on language learning; in particular, we will use a simplified tag set and a more systematic description of the word profile differences between corpora. We also plan to create word profiles for the DWDS-extended corpus, a 2 billion token corpus.

References:

Abney, S (1991): *Parsing by chunks*. In: Principle-Based Parsing. S. Berwick, A. & Tenny (ed). Kluwer Academic Publishers

Bolla, E. and Drumbl, J.: *Theoretische und praktische Aspekte der Wortschatzarbeit mit Korpusinstrumenten: ein Werkstattbericht.* In Press.

Braun, S.; Kohn, K. and J. Mukherjee (2006): *Corpus Technology and Language Pedagogy*. Peter Lang. Frankfurt.

Church, K; Hanks, P. (1989): *Word Association Norms, Mutual Information, and Lexicography*. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics; reprinted in Computational Linguistics, Spring 1990.

Didakowski, J. (2007): *SynCoP - Combining syntactic tagging with chunking using WFSTs.* Linguistik in Potsdam, In: Proceedings of FSMNLP 2007.

Evert, S.; Heid, U.; Säuberlich, B.; Debus-Gregor, E.; Scholze-Stubenrecht, W. (2004): *Supporting corpus-based dictionary updating*. In Proceedings of the 11th Euralex International Congress Lorient, France.

Geyken, A.; Hanneforth, Th. (2006): *TAGH - A Complete Morphology for German based on Weighted Finite State Automata.* In: Proceedings of FSMNLP 2005, Lecture Notes in Artificial Intelligence. Springer.

Geyken, A. (2007): *A reference corpus for the German language of the 20th century.* In Fellbaum C. (ed.) Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London (Continuum Press).

Hausmann, F.-J. (1984): *Wortschatzlernen ist Kollokationslernen.* In: Praxis des neusprachlichen Unterrichts. 31. Jg. (1984), S. 395-406.

Harris, Z. (1968): *Distributional Structure*. In Jerold J. Kart, editor, The Philosophy of Linguistics, Oxford Readings in Philosophy, pages 26-47. Oxford University Press.

Karlsson, F.; Voutilainen,A.; Heikkilä J.; Antilla A. (1995): *language independent system for parsing unrestricted text* Mouton de Gruyer. Berlin/New York

Kilgarriff, A.; Rychly, P., Smrz, P., and D.Tugwell (2004): *The Sketch Engine.* In Proceedings Euralex 2004. Lorient, France, July: 105-116.

Schiller, A., S. Teufel und C. Stöckert (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Forschungsbericht, Universität Stuttgart und Universität Tübingen.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford.