Generation of Word Profiles on the Basis of a Large and Balanced German Corpus

Alexander Geyken
Berlin-Brandenburgische Akademie der Wissenschaften

Jörg Didakowski
Universität Potsdam

Alexander Siebert
Berlin-Brandenburgische Akademie der Wissenschaften

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 system is intended as an additional informational source for the DWDS web-platform (www.dwds.de). The DWDS website-with 2.5 million page impressions per month-is a widely used internet platform that provides a word-information system based on a large monolingual German dictionary and the DWDS-Kerncorpus, a balanced corpus of German texts of the 20th century. The DWDS word profile 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 comprised of a database management system for collocations and a corpus query engine, together with a web interface. We have applied the DWDS word profile to a balanced German corpus of the 20th century and subsequently present some technicalities. Another experiment using the DWDS word profile in conjunction with a tabloid newspaper shows that there may be significant differences between corpora, underlining the importance of the corpus choice for language learning as well as for the construction of lexical resources. 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.

1. Background

Electronic corpora have been used in lexicography and the domain of language learning for more than two decades (cf. Sinclair 1991, Braun et al. 2006). 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, Sinclair 1991). Church and 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.

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.

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. In section 2, we present the wider context into which the word profile system is embedded: the DWDS lexical information system. Then we give an overview of the DWDS word profile system (section 3). The syntactic relations as well as their extraction process are described in section 4 and 5. The extraction process 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 6, we apply the DWDS word profile to two different corpora and present some technicalities.

2. General context: the DWDS lexical information system

The DWDS word profile system was implemented as an additional functionality of the DWDS lexical information system: in particular it has been developed to enhance its "collocation component", i.e. the component that computes statistically salient co-occurrences on the basis of a lemmatized corpus. We will therefore present the DWDS word profile system in its wider context. The DWDS website (*www.dwds.de*) is—with approximately 5 million page impressions (PI) per month—a widely used internet platform that provides lexical word information. Currently the lexical information system contains four different types of information for a given word (Geyken 2005):

The full dictionary entry of the electronic version of the "Wörterbuch der deutschen Gegenwartssprache" (WDG, "Dictionary of Present-day German") published between 1952 and 1977 (Klappenbach et al. 1977) and compiled at the Deutsche Akademie der Wissenschaften. The print version comprises six volumes with over 4,500 pages and contains more than 60,000 headwords (more than 120,000 if compounds are counted separately).

The corpus component (currently 800 Mio tokens in total) comprises newspaper corpora, specialized corpora (e.g. spoken language, language of the former German Democratic Republic GDR), and the DWDS core corpus. The core corpus consists of 100 million tokens (comparable in size to the British National Corpus), equally distributed over time and over the following five text types: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (22%) and other non-fiction (20%), transcripts of spoken language (5%). The corpus is encoded according the guidelines of the text encoding initiative (tei-P5). It is lemmatized with the TAGH morphology (Geyken and Hanneforth 2006) and tagged with the part-of-speech tagger moot (Jurish 2004) according to the conventions of the Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999). The corpus search engine DDC (Dialing DWDS Concordancer, Sokirko 2003) supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories), filtering (author, title, text type, time intervals) and sorting options (date, sentence length). Details on the design of the corpora and on the technical background of the corpus tools are given in Geyken (2007).

An additional thesaurus component computes synonyms, hyponymy and hypernyms for lexical units on the basis of the aforementioned WDG dictionary data (Geyken and Ludwig 2003).

On the basis of the DWDS core corpus, the collocation component offers several options to compute co-occurrences for a lexical unit according to common statistical measures (mutual information, t-score and log-likelihood). It does not, however, take into account syntactic relations.

3. DWDS Word profile system

Similarly, the DWDS word profile system computes statistically salient co-occurrences on the basis of lemmatized corpora. In addition, these co-occurrences are ordered by their syntactic relations (cf. section 4). Thus it provides the user with a more fine-grained "view" on the co-occurrence properties of a word.

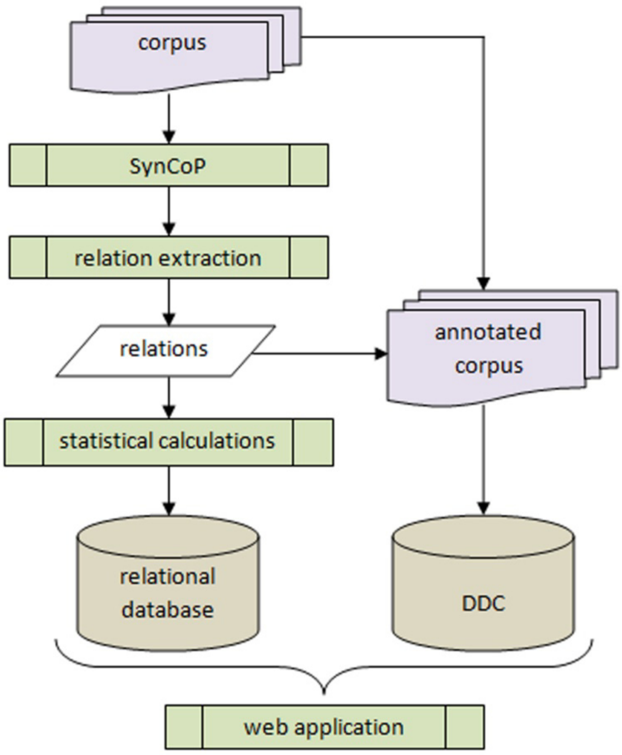

Figure 1: DWDS word profile generation system

The DWDS word profile generation process can be briefly described as follows (cf. Figure 1). Input to the DWDS word profile system is a large text corpus. The engine *SynCoP* (cf. section 3) is used to extract the syntactic relations for each lemma occurring with a sufficient frequency in the corpus. The syntactic relations (see below, section 4) 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 tuple, its frequency as well as the statistic salience is computed. We use the enhanced MI statistics suggested by Lin (1998) who defines the information I for a triple (w, r, w′) relative to its syntactic relation:

$$I(w, r, w') = \frac{\|w, r, w'\| \times \|*, r, *\|}{\|w, r, *\| \times \|*, r, w'\|}$$

Equation 1: Salience of triple (w, r, w′)

Here, *w* and *w′* are lemmas, *r* is a syntactic relation, ||*w, r; w′*|| denotes the frequency count of the triple *(w, r, w′)* in the parsed corpus. "*" denotes the wild card, and ||*w, r,* *|| is defined as the sum of the frequency counts over all lemmas *wj′* with. ||*w, r; wj′*||. Likewise ||*, *r,* *|| is defined as the sum of all triples (*w, r; wj′)* that share the relation *r*. The formula corresponds to the mutual information suggested by Hanks (1989) with the additional factor ||*, r, *||. In agreement with Kilgarriff (2004) we experienced that, in comparison to MI, (1) has the advantage not to overemphasize low frequency triples.

The collocation's tuples together with its statistic saliency are imported into a relational database (*MySql*), indexed and related to the corpus sentences by their offsets. The corpus is indexed via *DDC*, the linguistic search engine that is used for querying the corpora on the DWDS website.

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.

Word-clouds are visual presentations of a set of words, here a set of syntactic relations for a word, in which attributes of the text such as size, weight or colour can be used to represent features (e.g., salience) of the associated relations. Harvey and Keane (2007) have evaluated effectiveness of tag clouds, which are increasingly used in new web 2.0 services. The efficient visual representation of such user generated metadata is an important task. They describe the importance of font sizes and alphabetization for quickly finding relevant tags in tag clouds. The use of such distinguishing visual features is important for read effectiveness because users scan words clouds rather than read them. Kaser and Lemire (2007) present models and algorithms to improve and calculate the display of word clouds.

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

The DWDS word profile system extensively uses font size and alphabetisation to increase the readability of the used word clouds. For test and academic purposes we integrated the "older" standard presentation technique a list view which can also be used as a view.

Figure 2 gives an example of the generated word-clouds from our web front-end for the result of the verb-object and the preposition-noun-verb relation for the verb *essen* ("to eat"). For each syntactic relation the corresponding KWIC-lines in the corpus are extracted (cf. Figure 3).

4. Syntactic relations

The set of syntactic relations is predefined. Syntactic relations can be binary, such as the aforementioned adjective noun or verb object relations, 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"). Word profiles are computed for each lemma in the corpus of a certain frequency. Word profiles form an information cluster of the different syntactic relations. Syntactic relations vary with the lexical category. For example, a syntactic relation like adjective-noun is only meaningful for a lemma of the categories adjective and/or noun. Here, a difference between classical collocations and word profiles has to be noted: in linguistic literature, collocations are characterized as being unidirectional, i.e. they consist of a base and a collocate (e.g. Hausmann 1984). For example, in the collocation *confirmed bachelor*, *bachelor* is the base and *confirmed* is the collocate. The underlying motivation for this lies in the observation that the collocation is retrieved by the noun and not the adjective: hence a language learner would generate this collocation by looking for an appropriate adjective to *bachelor* and not by looking for an appropriate noun to the adjective *confirmed*. Since word profiles are generated automatically without semantic knowledge this unidirectness cannot be represented. We overcome this problem by storing syntactic relations bidirectional, i.e. the syntactic relation is stored for both the base and the collocate. Thus the completeness of the word profile for a given lemma is guaranteed.

1908-03-05	Zeitung:Morgen-Ausgabe	Er steht , wie die Schwester mit Behagen immer weiter	ißt	.
1912-12-31	Belletristik	Doch siegte am Ende mein Hunger über die Gewissensbisse und ich	aß	mit großem Behagen .
1950-12-31	Belletristik:Roman	Ja , so sagte er , verrecken , und er war voller Bitterkeit , aber dabei	aß	er sein Käsebrot mit allem Behagen .
1923-07-26	Zeitung	Ab und zu kam eine Dämonin , totschwarz , entblößte einen Raubzahn , snobte auf Morphinistin ,	aß	dann aber doch mit Behagen zweimal Schlackwurst mit Kraut « (also noch über der Orska) . Von einem Ungarn wird gesagt : » Er aß schon mit den Händen , stahl aber offenbar noch mit allen Vieren .
1926-12-31	Gebrauchsliteratur:Belletristik	Die anderen	essen	den Fisch mit Behagen , u. nur wir , die verrückten Ingleses , glauben zu sterben .

Figure 3: concordance lines for the syntactic relation *mit Behagen essen* (‘to eat with relish’)

Currently, 17 binary syntactic relations and one ternary syntactic relation are extracted from the corpus for the DWDS word profile system. The syntactic categories are closely related to the ENGCG tag set (see Halteren 1999) which are assigned by the SynCoP engine (see section 5). The following syntactic categories are currently used for the word profile system; its part-of-speech (pos) categories correspond to the widely used STTS tagset (Schiller et al. 1999):

1. Eight binary relations with respect to the head functions. For the relations the reverse relations are also explicitly represented:

relation (of/has)	example	translation
active-clause subject	der *Mann*[2] *tötet*[1]	the *man*[2] *kills*[1]
passive-clause subject	der *Mann*[2] wird *getötet*[1]	the *man*[2] is *killed*[1]
active-clause object	die Besatzung *sagt*[1] die *Wahrheit*[2]	the crew *tells*[1] the *truth*[2]
passive-clause object	Die Besatzung bekam die *Wahrheit*[2] *gesagt*[1]	the crew was *told*[1] the *truth*[2]
indirect object	der Mann *gibt*[1] der *Frau*[2] das Buch	the man *gives*[1] the book to the *women*[2]
auxiliary	Der Mann *wird*[2] *schlafen*[1]	the man *is going to*[2] *sleep*[1]
modal auxiliary	Der Mann *muss*[2] *schlafen*[1]	the man *has to*[2] *sleep*[1]
verb particle	Ich *stelle*[1] das Buch *zurück*[2]	I *put*[1] the book *back*[2]

2. Seven binary relations with respect to the modifier functions. For the relations the reverse relations are also explicitly represented (of/has):

relation (of/has)	example	translation
genitive attribute	das *Auto*[1] des *Mannes*[2]	the man’s[2] car[1]
determiner	*das*[2] *Auto*[1]	the[2] car[1]
preposition	*im*[1] *Auto*[2]	*In*[1] the *car*[2]
modifying noun	eine *Flasche*[2] *Wein*[1]	one *bottle*[2] of *wine*[1]
modifying adjective	der *intelligente*[2] *Mann*[1]	the *intelligent*[2] *man*[1]
modifying ad-adjective	der *sehr*[2] *intelligente*[1] Mann	the *very*[2] *intelligent*[1] man
modifying quantifier	*zwei*[2] *Autos*[1]	two[2] cars[1]

3. Two binary relations with respect to the coordination functions. Here, the coordination is considered symmetrical and gives no rise to a separate inverse relation:

relation	example	translation
noun coordination	der *Mann*^{1} und die *Frau*^{2}	the *man*^{1} and the *woman*^{2}
adjective coordination	der *große*^{1} und *geheimnisvolle*^{2} Mann	the *tall*^{1} and *mysterious*^{2} man

4. One ternary relation which concerns prepositional phrases functioning as facultative/mandatory adverb as well prepositional phrases in light-verb constructions. For the relation the reverse relations is also explicitly represented:

relation	example	translation
adverbial PP/light-verb	der Mann *lebt*^{1} *in*^{2} der *stadt*^{3}/*in*^{2} *Kraft*^{3} *treten*^{1}	the man *lives*^{1} *in*^{2} *the town*^{3}/ *to become effective (to enter*^{1} *in*^{2} *power*^{3}*)*

5. Extraction of syntactic relations with SynCoP

The extraction of the syntactic relations is based on the *Syntactic Constraint Parser* (SynCoP) an engine which performs the syntactic dependency annotation of the corpora fully automatically (Didakowski 2007). SynCoP is based on *finite state techniques* which have been used successfully in automatic corpus annotation tasks (cf. Koskenniemi 1990 or Abney 1996). Especially *weighted finite state transducers* (WFST)—a special kind of finite state machines—are used (see Mohri 2004).

SynCoP consists of a *grammar compiler*, a *grammar-driven parser,* and a *preprocessing module* which comprises tokenizing and the recognition of multi-word units. The Engine admits specification of the parser along with the preprocessing module by means of a grammar which is written in XML. Thus the engine can be easily adapted to individual conceptions of analysis. The writing of a grammar proceeds in three steps:

1. the pure technical design,
2. the interface design for the syntactic rule writing,
3. the writing of syntactic rules.

Here, the grammar compiler converts a given grammar into a specification which is used for the analysis process.

The morphological analysis is performed with the TAGH morphology, which is also used for the morphological annotation of the DWDS core corpus in the lexical information system of the DWDS: The TAGH morphology is a complete morphology which analyses productive German derivation and composition. Like SynCoP, TAGH is implemented with weighted finite state transducers.

The input of the system is a corpus of raw text and the system returns the syntactically annotated corpus as output. Figure 4 illustrates the basic mechanism of the SynCoP engine.

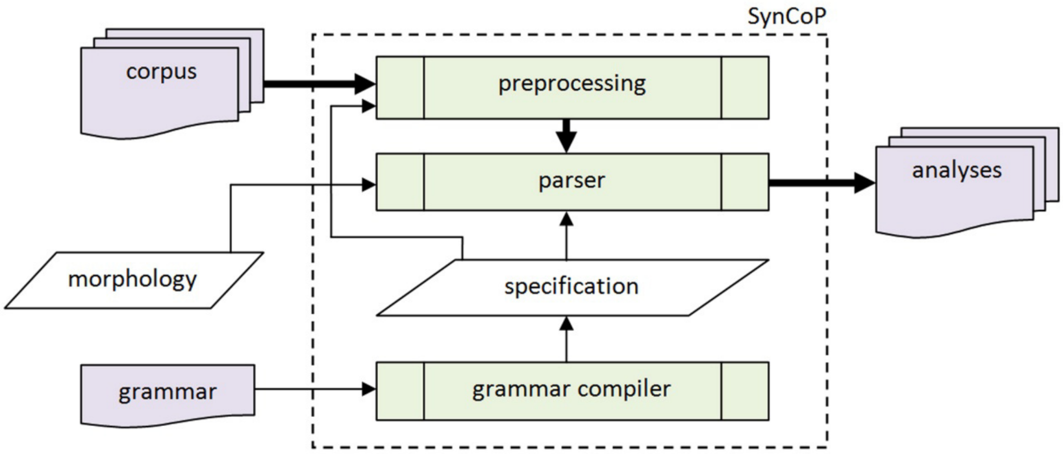

Figure 4: general overview of the system SynCoP

To build up our word profile system, information about the linking of words is needed. To provide such information, we implement a dependency parser within the SynCoP engine by means of a hand-written grammar (note that also constituency parsers can be implemented). In our implementation, we combine *syntactic tagging* (Karlsson 1995) with *chunking* (Abney 1991): The parsing is done by the marking of non-recursive phrases (chunks), main-clauses and sub-clauses, as well as by the syntactic tagging of modifier/coordination functions (determiner, genitive attribute, noun coordination, etc.) and head functions (subject, object, main verb etc.) within main-clauses or sub-clauses. In this approach the chunks can be seen as local dependency structures that are integrated into a global dependency structure by syntactic functions (Didakowski 2005).

The rules for chunking and for the syntactic tagging of head and modifier/coordination functions are implemented independently by our grammar. The grammar consists of five modules that are applied sequentially during the parsing process:

1. the morphology interface that maps the tag sets used by the TAGH morphology and by the grammar,
2. the chunking of non-recursive phrases and the tagging of contained modifier/coordination functions,
3. the syntactic tagging of modifier/coordination functions that are related to the chunks,
4. the marking of sub-clauses and the tagging of contained head functions,
5. the marking of main-clauses and the tagging of contained head functions.

A general problem with annotation tools working with finite state techniques is the cut-off of relevant syntactic readings in early processing steps. Such a cut-off occurs if a decision is made although not enough context is considered. This happens for example by greedy disambiguation strategies which are applied on chunk level (Abney 1995). In our approach, all local ambiguities are maintained during the five analysis steps to avoid such a fatal cut-off of syntactic readings.

An important precondition for the full automatic annotation of large text corpora is robustness. It is nearly impossible to write a grammar which covers all German sentences. A fundamental reason for this is the variability of the syntactic constructions in German (as in many other languages). Some syntactic constructions are very frequent, some are quite rare. Some are easy and some others are hard to explain. Another reason is the phenomenon of gradual grammaticality: some syntactic constructions are fully grammatical, whereas some others are “less grammatical”. Thus in our approach, we do not try to validate every sentence. This would be too time-consuming and in the end would be a hopeless task. Instead we try to extract the syntactic information as much as we can. Here, only the syntactic information supported by a sentence is relevant. In our approach, robustness is achieved by both local structures and the possibility of underspecified syntactic functions.

To allow highly efficient processing of the text corpora, a non-recursive model of the German language is assumed. This means the embeddings of phrases or clauses are bounded. Additionally tail recursion is treated as iteration. This is a common approach in full automatic corpus annotation, and seems to be "absolutely sufficient" (see Koskenniemi 1990).

Furthermore, SynCoP is required for a variety of different phenomena:

- the resolution of case/number/gender agreement phenomena, which are important to determine subject-verb relations,
- the recognition of verb particles, which are used for the correct lemmatization of complex verbs,
- the preference of readings in sentences which contain global ambiguity,
- the possibility of violating syntactic rules to cover gradual grammaticality.

The problem of free word order in German does not arise in this formalism because the possible variants of functions are defined *a priori*. Thus the engine 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 analyzed partially.

The analyses returned by the parsing process contain information about chunks, main-clauses, sub-clauses and syntactic functions. A simple example for this is given by the analysis of the following sentence—the title of a movie directed by Rainer Werner Fassbinder's (1974):

Angst essen Seele auf. (lit. fear eats soul up, engl. fear eats the soul) (example 1)

Labelled bracketing and syntactic tags are used here to represent the syntactic structure:

[[Angst@HEAD]$_{np}$@SUBJ essen@FMAINV [Seele@HEAD]$_{np}$@OBJ *auf@FPARTV* .]$_{cl}$

In this analysis the noun chunks "Angst" and "Seele" are marked by brackets ([...]$_{np}$) and the syntactic tag *@HEAD* within the chunks indicates the syntactic head of the chunks (which is necessary to infer a local dependency structure). The sentence as a whole is marked by brackets ([...]$_{cl}$), too. Within this clause, the syntactic tags for the head functions subject (*@SUBJ*), object (*@OBJ*), main verb (*@FMAINV*), and verb particle (*@FPARTV*) are assigned (the tags are strongly related to the ENGCG tag set). SynCoP returns such structures in an XML format. In this representation, the dependency relations and consequentially the different word profile relations are not directly accessible. To overcome this problem, word profile relations are inferred from such structures by interpreting the syntactic tags. Here, the word profile relations are inferred for each main-clause and sub-clause separately. The extracted dependency tree for the example sentence above is shown in Figure 5. With this dependency tree, a list of bidirectional word profile relations can be extracted. The word lemmas are used in the construction of the relation list. Here, the verb lemma is composed of the verb particle and the stem of the main verb:

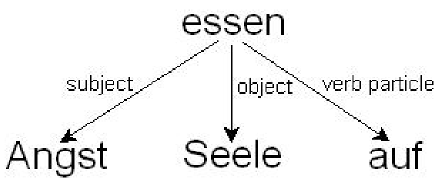

Figure 5: dependency tree for the active sentence: "Angst essen Seele auf."

- Angst – *active-clause_subject_of* – aufessen (engl: fear – eat up)
- Seele – *active-clause_object_of* – aufessen (engl: soul – eat up)
- auf – *verb_particle_of* – aufessen (engl: up – eat up)
- aufessen – *has_active-clause_subject* – Angst (engl: eat up – fear)

- aufessen – *has_active-clause_object* – Seele (engl: eat up – soul)
- aufessen – *has_verb_particle* – auf (engl: eat up – up)

This approach can be re-illustrated by a more complex example sentence:

> Jeder <u>Aspekt</u> des Vertrags von Rom sowie der im Anschluß an seine Unterzeichnung getroffenen Entscheidung und alle Folgen und Auswirkungen, die ein britischer Beitritt nach sich ziehen dürfte, sind von allen Seiten <u>beleuchtet</u> worden.
>
> Each <u>aspect</u> of the Treaty of Rome as well as the decision agreed upon following its signature and all consequences that Britain's accession to the EU could involve, have been <u>highlighted</u> by all sides (example 2)

In example 2 we focus on the word *Aspekt* (aspect), which is marked by brackets. Further, we focus on the passive construction, as well as on the long dependency and on the noun coordination. These aspects of the sentence above are annotated by SynCoP as follows:

> *[[*Jeder@*DN>* Aspekt@*HEAD]*$_{np}$@*SUBJ* *[*des@*DN>* Vertrags@*HEAD]*$_{np}$@*<GN* ... und@*CC* *[*alle@*DN>* Folgen@*HEAD* und@*CC* Auswirkungen@*HEAD]*$_{np}$@*SUBJ* sind@*FAUXV* ... beleuchtet@*FMAINV* worden@*FAUXV* *.]*$_{cl\_passive}$

The syntactic tag *@DN>* stands for a noun-determiner relation and the tag *@<GN* stands for a noun-genitive relation. Here, the arrow "<" or ">" gives the direction of the head of the relation. The syntactic tag *@CC* stands for a coordination relation, and the syntactic tag *@FAUXV* stands for a verb-auxiliary relation. The sentence is marked as a passive clause by the bracketing ([...]$_{cl\_passive}$). The meaning of the other tags can be taken from the first example. A dependency tree can be extracted from the information provided by the annotated sentence fragment. Such a dependency tree is shown in figure 6.

Now the bidirectional word profile relations can be extracted from the dependency tree with respect to the word "Aspekt". For this purpose, we focus only on the edges of the tree which are related to this word:

- Aspekt – *passive-clause_subject_of* – beleuchten (engl: aspect – highlight)
- jeder – *determinier_of* – Aspekt (engl: each – aspect)
- Vertrag – *genitive_attribute_of* – Aspekt (engl: treaty – aspect)
- beleuchten – *has_passive-clause_subject* – Aspekt (engl: highlight – aspect)
- Aspekt – *has_determinier* – jeder (engl: aspect – each)
- Aspekt – *has_genitive_attribute* – Vertrag (engl: aspect – treaty)
- Aspekt – *noun_coordination* – Folge (engl: aspect – consequence)
- Aspekt – *noun_coordination* – Auswirkung (engl: aspect – implication)
- Folge – *noun_coordination* – Aspekt (engl: consequence - aspect)
- Auswirkung – *noun_coordination* – Aspekt (engl: implication - aspect)

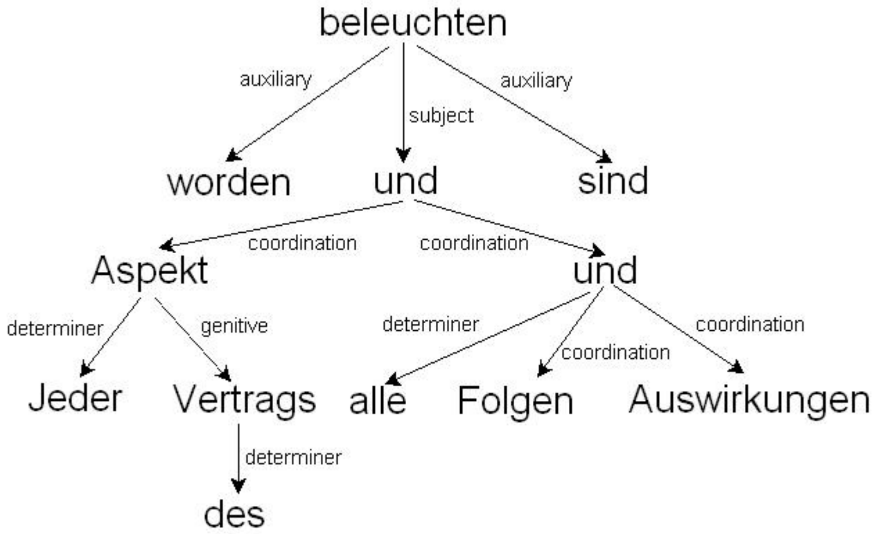

Figure 6: dependency tree for the passive sentence: “Jeder Aspekt des Vertrags ... und alle Folgen und Auswirkungen ... sind ... beleuchtet worden.”

6. Word Profiles for two large German corpora

The DWDS word profile tool was applied to three different corpora: the DWDS core corpus, a large balanced corpus of German texts of the XXth century (cf. section 2), the weekly newspaper *Die Zeit* (electronic archive from 1997-2006) and 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.

We decided to combine the electronic ZEIT archive and the DWDS core corpus (henceforth referred to as DWDS/ZEIT corpus), first, because both corpora taken together cover the entire 20th century as well as up-to-date texts, and second, because DWDS core corpus and ZEIT compare in that they both use a similar proportion of standard German. We opted for building a separate word profile on the basis of the BILD archive (henceforth referred to as BILD corpus) in order to be able to investigate the impact of corpus differences on word profiles. Both corpora differ not only in their text composition but also with respect to their size: the DWDS/ZEIT corpus contains 140,000 documents with approximately 160 million tokens whereas the BILD corpus consists of 555,000 documents and comprises 90 million tokens.

For both corpora, the above mentioned (section 4) syntactic relations were extracted. For the ZEIT/DWDS corpus it took 2 days on a 8-processor computer to extract 68 million syntactic relations corresponding to 1.26 million lemma-pos[1] pairs. 171,000 (42,929; 8,500) lemma-pos pairs occur 10 (100; 1,000) times or more in the corpus. For BILD, it took 1,5 days on a 8-processor computer to extract 37 million syntactic relations corresponding to 791,165 lemma-pos pairs. 105,204 (26,594; 5,108) lemma-pos pairs occur 10 (100; 1,000) times or more in the corpus.

The calculation of the statistical values (MI, salience) took approximately 2 days for both corpora. The storing and indexing in the relational database model and the DWDS linguistic search engine required another 3-4 days. The long database creation process is due to the high indexing effort to gain high performance querying of the syntactic relations and corresponding KWIC-lines in the corpus. In total, the word profile generation for the DWDS/ZEIT corpus (resp. BILD corpus) required 7 (5) days.

For both corpora, a prototype containing all lemma-pos pairs with a frequency greater than 10 is accessible on the Internet under *http://odo.dwds.de/wortprofi*l. The user can type in any word (in

[1] Pos stands for part-of-speech

lemma form). The lemma is then expanded to one or more lemma-pos pairs[2]. Their corresponding word profiles are displayed as word-clouds. There are as many word-clouds as relations for the word. As a default, only those relations are displayed where the triples (w, r, w′) occur at least five times in the corpus. For each relation, the 20 most salient triples are displayed. It is possible via the interface to modify those settings: for high-frequent lemma-pos pairs it is useful to increase the number of displayed triples whereas for low-frequent lemma-pos pairs it is sometimes necessary to lower the occurrence threshold to less than 5.

7. Conclusion and discussion

We have presented the DWDS word profile system, a software-tool that extracts statistically salient co-occurrences from corpora and clusters them according to their syntactic categories. Due to the difficulties of German, in particular its free word order and long distance dependencies, shallow approaches like phrase chunking are not sufficient for a satisfactory extraction of syntactic relation. Our system uses a syntax parser based entirely on weighted finite state transducers which combines satisfactory extraction of syntactic relations with good performance. Currently, we have built a prototype for two corpora of 160 m tokens (resp. 90 m tokens) that are accessible via the Internet. We will integrate the word profile as an additional information source for the DWDS web-platform.

The feedback by users of our Internet prototype confirms the assumption in section 2 that using word-clouds instead of tables or lists facilitates the work with word profiles. The main focus of our future work will be in the following areas: evaluation of the quality of word profiles, the influence on different corpora on the word profiles and the enhancement of our system for the requirements of language teaching.

In the near future we plan to evaluate more systematically the quality of the extracted word profiles in terms of correctness and completeness of the extracted triples. In agreement with Kilgarriff (2004) we are less worried with correctness since we suppose that these errors will be filtered out statistically. As one possible baseline for completeness we could compare the extracted relations with a large monolingual print dictionary. The following example with the noun *Angst* (anxiety, fear) shows that the automatically extracted syntactic relations compare fairly well to the constructions listed in the electronic version of the WDG [cf. section 2. The WDG lists here 9 verbs. 6 (8) of them are statistically salient with a frequency greater than 5 (3) in the word profile. Only one entry of the WDG was not extracted by the word profile (and not present in the corpus) whereas 4 (7) salient word triples of the word profile with a frequency greater than 5 (3) are not listed in the WDG. We plan to do this comparison on a larger scale in the near future].

We also plan to investigate the differences of word profiles between the DWDS/ZEIT corpus and the BILD corpus. The following example the with the verb *übertragen* (“to transmit”) shows which type of differences might be expected: Here the DWDS/ZEIT-corpus has a much larger variety of collocating direct objects. Many of them correspond to support verb constructions and hence a formal language: *Ermächtigung*, *Befugnis* (both authorization), *Aufgabe* (task), *Daten* (data), *Verantwortung* (responsibility), *Zuständigkeit* (competency), *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

[2] This is due to the STTS tagset which is used for our extraction task where, for example, two different adjective tags and four different verb tags are used.

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).

A third aspect of our future work is to make the use of word profiles easier for language learning purposes. In particular, we will use a simplified tag set and a more systematic description of the word profile differences between corpora. Additionally, we intend to store the extracted relations in 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.

References

Abney, S. P. (1991). “Parsing by chunks”. In Berwick, R. C.; Abney, S.; Tenny, C. (eds.). *Principle-Based Parsing*. Boston: Kluwer Academic Publishers. 257-278.

Abney, S. (1995). “Chunks and dependencies: Bringing processing evidence to bear on syntax”. In Cole J.; Green, G.; Morgan, J. (eds.). *Computational Linguistics and the Foundations of Linguistic Theory.* Stanford: CSLI. 145-164.

Abney, S. (1996). “Partial Parsing via Finite-State Cascades”. In *Proceedings of the ESSLLI ‘96 Robust Parsing Workshop. 8-15.*

Bolla, E.; Drumbl, J. (2008). *Theoretische und praktische Aspekte der Wortschatzarbeit mit Korpusinstrumenten: ein Werkstattbericht.* In press.

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

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*. 76-83.

Didakowski, J. (2008). “SynCoP - Combining syntactic tagging with chunking using WFSTs”. *In Proceedings of FSMNLP 2007*.

Didakowski, J. (2005). *Robustes Parsing und Disambiguierung mit gewichteten Transduktoren*. *Linguistics in Potsdam 23*. Potsdam: Universitätsverlag Potsdam.

Evert, S. et al. (2004). “Supporting corpus-based dictionary updating”. In *Proceedings of the 11th Euralex International Congress*. Lorient, France. 255-264.

Geyken, A.; Ludwig, R. (2003). *Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache* [on-line]. TaCoS. *http://kollokationen.bbaw.de/doc/ExtrHyp.pdf* [Access date: 27 March 2008].

Geyken, A. (2005). “Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS)”. *BBAW Circular* 32. Berlin: BBAW.

Geyken, A.; Hanneforth, T. (2006). “TAGH - A Complete Morphology for German based on Weighted Finite State Automata”. In *Proceedings of FSMNLP 2005*. 55-66.

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.

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

Harvey, M.; Keane, M. (2007). “An Assessment of Tag Presentation Techniques”. In *Proceedings of the 16th international conference on World Wide Web*. Alberta: ACM. 1313-1314

Hausmann, F. J. (1984). “Wortschatzlernen ist Kollokationslernen”. In *Praxis des neusprachlichen Unterrichts*. Vol. 31. 395-406.

Halteren, H. (1999). *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers.

Kaser, O.; Lemire, D. (2007). “Tag-Cloud Drawing: Algorithms for Cloud Visualization”. *The Computing Research Repository (CoRR)*. abs/cs/0703109.

Klappenbach, R.; Steinitz, W. (eds.) (1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG)*. Berlin: Akademie-Verlag.

Koskenniemi, K. (1990). "Finite-state parsing and disambiguation". In *Proceedings of the the 13th International Conference on Computational Linguistics (COLING 90)* 2. 229-232.

Karlsson, F. et al. (1995). *Language independent system for parsing unrestricted text.* Berlin, New York: Mouton de Gruyer.

Kilgarriff, A. et al. (2004). "The Sketch Engine". In *Proceedings of Euralex 2004*. Lorient, France. 105-116.

Mohri, M. (2004). "Weighted Finite-State Transducer Algorithms: An Overview". In Martin-Vide, C.; Paun, G.; Mitrana, V. (eds.). *Formal Languages and Applications (Studies in Fuzziness and Soft Computing)*. Berlin, Heidelberg, New York: Springer Verlag. 551-563.

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

Schiller, A.; Teufel S.; Stöckert, C. (1999). "Guidelines für das Tagging deutscher Textcorpora mit STTS". Technical report. Universität Stuttgart, Tübingen.

Sokirko, A. (2003). "DDC". *Computational linguistics and intellectual technologies.* Protvino, Russia.