Vincent J. DOCHERTY, Langenscheidt KG
Ulrich HEID, Institut für maschinelle Sprachverarbeitung – Computerlinguistik, Universität Stuttgart

Computational Metalexicography in Practice – Corpus-based support for the revision of a commercial dictionary

Abstract

In a cooperation between dictionary publishers and computational linguists, raw material for the revision of the German part of a bilingual German → English dictionary (*Langenscheidts Handwörterbuch Englisch*, Neubearbeitung 1991) was produced. In a case study, the entries for headwords with the initial letter "p", then, – between August 1997 and March 1998 – the full dictionary were systematically checked against a 300 million word German newspaper corpus from the late 80s and early 90s. The objective was to find evidence to support updates of the lemma inventory of the dictionary and to enhance the example and collocation coverage. The data production from the corpora is automatic, the (manual, interactive) lexicographic procedures remain unchanged. To this end, standard corpus pre-processing (tokenizing, tagging, lemmatization) and a hierarchical set of query templates for collocation extraction were used. The dictionary was transformed into a specific data format (similar to database entries), and the examples contained in the articles were prepared for automatic querying. The results are of metalexicographic interest: they show the potential of refined macrostructural selection procedures, help to improve the documentation of readings through examples, and, generally, provide an example of the use of standard computational linguistic techniques for dictionary revision. The auxiliary resources constructed from the corpora in the same process – a verb frequency lexicon for German and a collection of noun-verb collocation candidates are useful and relevant in their own right. Similarly, the tools used are mostly generic and thus reusable outside the specific context discussed here.

Keywords: Metalexicography, dictionary analysis, dictionary updates, corpus based semi-automatic lexical acquisition.

1. Context – Objectives

1.1. The application context

Langenscheidt publishers[1] are currently updating their *Handwörterbuch Englisch*, a medium size[2] German ↔ English dictionary. One of the objectives in the revision is to ensure a good coverage of up-to-date German journalistic texts.[3] Headwords and examples not particularly relevant for the work with texts of this kind may be removed, whereas items not yet covered but of high frequency in the targeted texts should be added. In addition, the collocational coverage (currently only noun-verb-collocations) is verified and enhanced.

The revision of the dictionary text itself is carried out entirely by the publisher's lexicographic team, on the basis of a WWW-browsable (in-house) data collection of all relevant project results. The task of the computational linguists is to produce raw material for the lexicographic decision-making process. Many additional parameters may influence this revision process, and the data resulting from the project are just one of these.

1.2. Objectives

Both the macrostructure and the microstructure of the *Handwörterbuch Englisch* will be substantially revised, with parts removed and new material added to the German part of the dictionary. The comparison between dictionary and corpus has led to sets of candidates for addition, modification or removal, with respect to the following types of information:

- Macrostructure
  – Corpus frequency of headwords of the dictionary entries: low frequency lemmas are candidates for removal from the dictionary.
  – Frequency of lemmas from the corpus not yet covered in the dictionary: high frequency lemmas are candidates for inclusion in the dictionary.
- Microstructure
  – Headwords without examples: for high frequency lemmas, examples (and possibly collocations) are extracted from the corpus.
  – Noun-verb collocations: For each collocation in the dictionary, we indicate under which headword it is lemmatized, such that material may be reorganized[4]; additional collocation candidates from the corpus are proposed. Similar procedures could be used for adjective-noun collocations (and other types of collocations) as well. In the project framework, the focus was on noun-verb collocations, which are hardest to extract from German corpora.

2. Methods and Tools

The tasks described above were carried out with standard tools for corpus processing. The dictionary was transformed into a homogeneous format, to facilitate the comparison with corpus data.

2.1. Dictionary analysis

This section is mainly devoted to a computational implementation of a metalexicographic analysis of the macro- and microstructure of the dictionary. Note that only as much metalexicographic work was invested as necessary to cope with the objectives of the project. A more in-depth analysis, although possible in principle, could not be made.

2.1.1. Lemma information

The dictionary is represented in a tagged format similar to SGML. The nested microstructure is flattened in the electronic version produced at the publishing house, and the usual text condensation devices are normalized. Figure 1 is the entry s.v. *bereiten*, as it appears in print. Figure 2 is the corresponding tagged text.

> **bereiten** v/t. **1.** prepare, get *s.th.* ready; (*zubereiten*) make *some tea etc.*; (*Leder*) dress;
> **2. fig.** (*verursachen*) cause; **j-m Kopfschmerzen** *etc.* ~ *a.* give s.o. a headache *etc.*;
> → **Empfang** 2, **Ende, Freude etc.**

Figure 1: The entry s.v. *bereiten* in the printed dictionary

From this data, two types of reference resources have been created: a German lemma list and an inventory of all example sentences contained in the dictionary; for practical reasons, this inventory was created in the same format as the text corpora used in the project: thus, one and the same tool can be used on both, corpus and examples, to extract collocations.

```
<RD:Eintrag><FD:Stichwort>bereiten </FD:Stichwort><FD:Wortart>
Verb/transitiv</FD:Wortart><CR> <FD:"Erläuterung">(zubereiten)
</FD:"Erläuterung">make <FD:"Erläuterung"> <IT>some tea
etc.</FD:"Übersetzung"> </FD:"Erläuterung"> <FD:"Erläuterung">(Leder)
</FD:"Erläuterung"> dress <FD:"Erläuterung">übertragen (verursachen)
</FD:"Erläuterung"> cause <FD:Beispiel>jemandem Kopfschmerzen <BD->
<IT+><FT:Times,SR>etc. <BD><IT><FT>bereiten <IT+>a. </FD:Beispiel>
<IT>give someone a headache <FD:Verweis><FT> <QL:Suche,"[Feld
Stichwort:Empfang [Feld NUM:2]]">Empfang 2</FD:Verweis><FD:Verweis>
<QL:Suche,"[Feld Stichwort:Ende ]">Ende</FD:Verweis>,<FD:Verweis>
<QL:Suche,"[Feld Stichwort:Freude ]">Freude</FD:Verweis>
```

Figure 2: The entry s.v. *bereiten* in the SGML-like annotation

The lemma list contains the following types of information:

- the entry headword;
- a unique identifier ("ID" in figure 3: used to retrieve variants and cross-references);
- the category indication contained in the dictionary ("WC": the wordclass labels used in the dictionary were mapped on those of the corpus; using the latter for both, corpus and dictionary material, ensures comparability);
- a code for the type of entry found in the dictionary ("ET": full entry with translation; partial entry without translation; link or cross-reference);
- a code indicating the presence or absence of examples in the entry ("EX", with values "+bsp)" or "-bsp)";
- the frequency of the lemma in the corpora used in the comparison ("LF");
- the frequency of the headword word form in the corpora used in the comparison ("WF").

In figure 3, we reproduce a few lines from the data collection created from the dictionary; the entry s.v. *bereiten* is summarized in the second line of the table, marked by an arrow.

Headword	LF	WF	ID	WC	ET	EX	
bereit	24354	24116	7503	ADJA	complete	+bsp	
bereiten	6718	2964	7504	VVINF-t	complete	+bsp	←
bereitgestellt	437	1937	7505	ADJA	complete	+bsp	
bereithalten	1301	182	7506	VVINF-t	complete	-bsp	
bereitlegen	74	10	7507	VVINF-t	complete	-bsp	
bereitliegen	310	43	7508	VVINF-i	complete	-bsp	
bereitmachen	40	2	7509	VVINF-t	complete	-bsp	

Figure 3: Reformatted headword list of the dictionary

The transformation of the dictionary into the above format was carried out by means of a suite of `Perl` and `C` tools.

As can be seen easily from figure 3 the extraction results combine descriptive linguistic data and metalexicographic data: the linguistic information is needed for the comparison with corpus data (category, subcategorization), whereas the metalexicographic facts (entry with or without examples, place of lemmatization of collocations, etc.) are needed to correlate

linguistic description and metalexicographic presentation, and the frequency figures, finally, contribute the corpus dimension, at least for the macrostructural analysis.

2.1.2. Extracting collocations from the dictionary

In the *Handwörterbuch*, noun-verb collocations appear in different types of items. There is no specific item type (WIEGAND's terminology) for collocations (other than, e.g. in the Van Dale bilingual dictionaries), but collocations are distributed over the following three item types, the first two covering, in a 2:1 relation, together almost all of the noun-verb collocations contained in the dictionary:

- Example sentences: the items tagged as `FD: Beispiel` (see above, figure 2) contain syntagms which are or contain collocations. Note that `FD: Beispiel` is polyfunctional: not all examples are collocationally relevant (though most are, indeed); this leads to some noise in our data.
- Meaning discrimination glosses: the items tagged as `FD: Erläuterung` (see above, figure 2) are also polyfunctional: they contain synonyms (as in the above entry: "**bereiten** – (*zubereiten*) – make *some tea etc.*"), but also collocates ("**bereiten** – (*Leder*) – dress"), domain indications and others; collocates can be extracted from nominal glosses in verb entries.
- Cross references: some cross references lead to collocations: s.v. *Mund*, there is a reference to *stopfen*$_3$, where we find the example "*fig.* **j-m den Mund** ~ silence s.o." Since collocational cross references are covered "somewhere else" in the microstructure, we have not fully exploited cross references yet.

From the glosses, we can currently only extract the noun and verb lemmas; from the examples, the syntactic construction of the verbal collocate can also be extracted (e.g. *den Weg bereiten*, with transitive *bereiten*, vs. *sich auf den Weg machen* with the preposition *auf*).

To verify the dictionary's collocational coverage, all examples were extracted, preprocessed as described in section 2.2 and made accessible to corpus query, preserving information about the headword under which the example is listed and the exact text of the example. Figure 4 contains all collocations with the verbal collocate *bereiten* found in the dictionary[5], along with the entry where the collocation is lemmatized (last column, right).

Verb lemma	Noun lemma	Synt.	Lemmatized s.v.
bereiten	Aufnahme	nop	(Aufnahme)
bereiten	Boden	nop	(Boden)
bereiten	Empfang	nop	(Empfang)
bereiten	Kopfschmerzen	nop	(Kopfschmerzen, bereiten)
bereiten	Kopfzerbrechen	nop	(Kopfzerbrechen)
bereiten	Lauferei	nop	(Lauferei)
bereiten	Ovation	nop	(Ovation)
bereiten	Qual	nop	(Qual)
bereiten	Verdruß	nop	(Verdruß)
bereiten	Weg	nop	(Weg)
bereiten	Willkommen	nop	(Willkommen)
bereiten	Überraschung	nop	(Überraschung)

Figure 4: Collocations with the collocate *bereiten* in example sentences of the dictionary

2.2. Corpus pre-processing

The analysis of the German corpora relies on standard tools and methods for low-level processing. The corpora are tokenized (word and sentence boundaries) and part-of-speech tagged with the STTS tagset[6] using SCHMID's decision tree tagger.[7] The tagging process includes lemmatization, based on morphological and part-of-speech information.

2.3. A lemma frequency list from German corpora

The corpus exploration work relies on the CQP/XKWIC corpus query tools (see (Christ 1994)), a query package supporting regular expressions over word forms and annotations of any type, as well as set operations on the extraction results. Extraction templates (i.e. complex queries with variables) make use of information about sentence boundaries, sequencing and adjacency of word forms, lists of lemmas (e.g. for function words), and boolean expressions over word forms, lemmas and/or part-of-speech shapes.

The creation of a lemma frequency list from corpora is mostly trivial, except for German verbs with separable prefixes. We have extracted about 10,000 simplex verb lemmas from a 200+ million word corpus, plus an additional 20,000 lemmas of verbs with separable prefixes (example: *er trennt das Stück ab*). Since separable verbs can appear in separated (as above) and in non-separated forms (e.g. *weil er ... abtrennt/abgetrennt/abzutrennen hat*), it is important to carry out independent frequency counts for each type of occurrence. The identification of separated forms relies on corpus queries (usually the prefix appears at the right sentence boundary, "rechte Satzklammer"), and on a subsequent comparison of lemma candidates with those derived from "non-separated" contexts.[8]

2.4. Corpus queries for collocations

The German corpora were used to extract a reference list of noun-verb collocations, with the emphasis on light verb constructions ("Funktionsverbgefüge"). This extraction work was exclusively based on corpus query, i.e. on symbolic, not on statistical techniques.

Statistical approaches, such as the well-known Mutual Information and t-score measures usually rely on adjacent items (which would be useful for adjective-noun collocations) or on windows of a fixed size. They produce information on the fact that two lexemes cooccur, but almost nothing else. We are interested, however, in a broad syntactic classification of the extracted collocations, such as "verb-object" vs. "subject-verb" (*Frage aufwerfen* vs. *eine Frage stellt sich*), the use of determiners and prepositions (*ein Ende finden, sein Veto einlegen, zur Rede stehen, Rechnung tragen*, etc.), singular vs. plural nouns (*Grenzen setzen*). In order to extract such information, the extraction tools need to consider syntactic phenomena such as the three different word order models of German, separated verb prefixes or different syntactic verb types (reflexive/non-reflexive, different subcategorization types). These phenomena are hard if not impossible to express in terms of "windows" (cf. (Breidt 1993), who notes similar problems), but are captured with our query patterns: we obtain the intended broad syntactic classification of the extracted collocations automatically. Note that no robust parsing of the corpora was available for the project work yet; with chunked or parsed data available, statistical measures could have been more successfully applied (see (Krenn 1998) for similar experiments).

The corpus queries for collocations operate only on corpus sentences where the verb complex is at the right sentence boundary[9] and directly follows a nominal or prepositional group. The head of this noun group and the main verb of the verb complex are identified and their cooccurrence is counted.[10]

The queries have been organized hierarchically. Each query template identifies a subset of the corpus sentences with specific syntactic properties:

- reflexive vs. non-reflexive verbs[11];
- "prepositional" vs. "accusative/dative" constructions (*im Vordergrund stehen* vs. *eine Frage stellen*);
- details of the noun group, at the levels of determiners (no determiner vs. definite vs. indefinite), of adjectival modification of the noun (*ein jähes Ende finden*), and of the presence or absence of genitives and/or prepositional phrases to the right of the noun *(im Zusammenhang mit x stehen)*.

More details on the extraction work can be found in (Heid 1998).

2.5. A corpus of example sentences from the dictionary

The procedures described in sections 2.2 and 2.4 are applied to all example sentences from the dictionary. As mentioned in section 2.1.2, the glosses are also explored for collocation candidates. The data from both sources are merged. An example, for the noun *Empfang*, is given in figure 5. Each line is a data set from the dictionary: preposition (or "nop", for: "no prep."), noun, verb, entry where the collocation is lemmatized:

```
nop   Empfang bereiten       (Empfang)
nop   Empfang bescheinigen   (bescheinigen)
---   Empfang bestätigen     (bestätigen)
nop   Empfang geben          (Empfang)
---   Empfang nehmen         (nehmen)
in    Empfang nehmen         (Empfang)
```

Figure 5: Collocations with the base noun *Empfang* in the dictionary

3. Results

This section summarizes our results and gives examples. A full-scale assessment of the lexicographic use of the data is under way; we thus can only show the types of data produced and indicate the number of instances of each type produced by the automatic procedures, but we cannot yet quantify in detail the economies realized, in comparison with the usual methods of dictionary revision.

3.1. Tools – Methods – Resources

The comparison of the dictionary and the corpus is based on generic computational lexicographic tools and resources:

1. A simplified data collection created from the machine-readable dictionary: an example is given in figure 3.

2. Frequency data for word forms and lemmata, as well as collocational data, extracted from the corpora.
3. Tools for the manipulation of data represented in ASCII data collections of the kind illustrated in figure 3.

The data in item (1), above, can be extracted with more or less the same kind of `Perl` and `C` tools from any dictionary. The preprocessing of textual corpora with a view to producing data of type (2) is a set of standard procedures used in other work at IMS as well. The extraction tools have been designed to operate on any German corpus available.[12] The same tools and data have been used for research on collocations (cf. (Heid 1998)), and similar ones have been developed for the extraction of evidence for subcategorization properties of verbs, nouns and adjectives (cf. (Eckle-Kohler/Heid 1996b), (Eckle-Kohler 1998)).

3.2. Macrostructural updates

Macrostuctural updates concern proposed candidates for inclusion and for removal. Both types are made accessible to the lexicographers in alphabetical order and by frequency. In addition, inclusion candidates are given in derivational families.

For each inclusion candidate, about 10 example sentences (extracted from the *Frankfurter Rundschau*) are given. Similarly, if the dictionary contains example sentences (or syntagms) for items which qualify as removal candidates, given their low corpus frequency, the examples can be displayed: frequency is but one of many parameters according to which lexicographers decide that a word should be in the dictionary. Other criteria are, for example, contrastive relevance, disponibility, text-type specific relevance, etc.

Here are a few inclusion candidates (with frequency figures in the 200 M word corpus used): the word *Deutschland* (94,304) does not appear in the macrostructure, because the dictionary has an appendix with geographical names. *Investor* (7,284), *Parlamentswahl* (3,400), *Pressemitteilung* (2,023), *Kindergartenplatz* (1,536) and many others are not contained in the macrostructure so far. Many inclusion candidates belong to the fields of politics and economy. Compounds with "*Schadstoff-*" as a first element include *Schadstoffeintrag*, *-konzentration*, *-messung*, *-reduzierung* and *Schadstoff-Mobil*, in addition to *Schadstoff-ausstoß*, *-emission*, *-belastung*, *-norm(en)*, *-richtlinien*, which are already in the dictionary.

The corpus material provides over 2,000 lemma candidates with a frequency higher than 500 (in 200 M) which are not yet in the dictionary. Going down to a frequency of 100 in 200 M, almost 10,000 candidates are found, and with a threshold frequency of 50, we find over 18,000 candidates. Similarly, we find 25,000 lemmas from the dictionary not more than 20 times in the 200 M words of our corpora; 10,000 of these 25,000 are nouns with a corpus frequency between 5 and 0, another 9,000 nouns with frequencies between 6 and 20.

3.3. Microstructural updates

3.3.1. Examples

The analysis shows that only about 25-30% of the headwords contained in the dictionary are documented with examples.

Currently, we only provide a list of items without examples, along with corpus sentences from the *Frankfurter Rundschau*, where these are available. Examples seem to be particularly necessary with polysemous verbs; for example, *klären* (10,359 in 200M: purify, clear (up), clarify, be settled, be solved), *tauchen* (6,467, 6 equivalents) are described only by means of glosses, but without examples.

3.3.2. Collocations

The analysis of noun-verb collocations in the dictionary and the comparison with our corpus-derived collection of collocations lead to two types of results: on the one hand data on the lemmatization practice in the dictionary (are the collocations found in noun or in verb entries, in both, or elsewhere), on the other hand material enabling us to complete the collocational description given in the dictionary.

The lemmatization practice in the dictionary had not been computationally controlled before our project. A few rules have been observed with few exceptions: light verb constructions with semantically (almost) empty verbs (*bringen, führen, kommen, gelangen*, etc.) are found in the entries of the nominal bases; on the other hand, collocations with salient, rather specific verbal collocates are found under the verb rather than under the noun (e.g. *Vorschlag* + *unterbreiten* under the verb only), which is the perspective of the "passive" dictionary (see also (Heid to appear)).

There are comparatively few cases where collocations are lemmatized twice; however, about one third of all collocations contained in the dictionary are only found in the collocate entries, not in those of the bases.

All in all, the dictionary is quite rich in collocations: it contains about 20,000 noun-verb collocations, in a total of 75,000 headwords. It offers particularly good coverage of the base nouns of somewhat higher frequency, with much less attention paid to the less frequent cases.

The dictionary has a fairly broad coverage of semi-specialized items (close to terminology) from a wide range of domains, whereas the corpus (evidently) provides more material for the language of sports, politics and economy. An example is the series of combinations with the base *Partie* in figure 6 :

Partie	dict.	corpus
eine ... ~ machen	+	–
eine ~ gewinnen	–	+
eine ~ verlieren	–	+
in einer ~ stehen	–	+
eine ~ endet (...)	–	+
eine ~ entscheiden	–	+

Perspektive	dict.	corpus
~ stimmt	+	–
aus einer ~ sehen	+	+
aus einer ~ betrachten	+	–
eine ~ eröffnen	–	+
eine ~ geben	–	+

Figure 6: Collocations in the dictionary and in the corpus: examples of comparison results

Newspaper style has more abstract uses than are covered by the dictionary: the examples s.v. *Perspektive* in figure 6 illustrate a reading of this noun (cf. *eine Perspektive geben, eröffnen*: '`prospect(s)`'), so far not illustrated in the dictionary. A further example: for the base noun *Fehler* (`error, mistake`), the dictionary has 17 collocations; the entry s.v. *Fehler* itself contains *Fehler machen, Fehler begehen, in (einen ...) Fehler verfallen.* Additional verbs are found in the respective verb entries for *einsehen, nachsehen, anstreichen, aufzeigen, ausbessern, ausmerzen, ausschalten, beheben, beschönigen, entdecken, feststellen, stehenlassen* and *in (den ...) Fehler zurückfallen.* Finally, the dictionary has *etwas steckt voller Fehler.* The corpus provides the following 6 additional verbs which may be worth entering in the dictionary:

- *Fehler ausbügeln, korrigieren, berichtigen*;
- *(ein) Fehler schleicht sich ein, unterläuft jemandem*;
- *jemand leistet sich (...) Fehler.*

In addition, the corpus has less typical, yet somehow relevant verbs, such as *Fehler eingestehen, einräumen, zugeben, nachweisen, vermeiden, finden, enthalten, entschuldigen.* As noise, it contains the trivial combinations *zu (...) Fehler führen, der Fehler liegt (wo), (...) Fehler bemerken, wiederholen, erkennen.* These examples show that the corpus-derived data enable us to further improve the collocational description contained in the dictionary.

3.4. An interface for the inspection of the results

All data extracted from the dictionary and the corpus are interactively accessible to the lexicographer through an interface based on a web browser. This choice has the advantage of being a well-known technology: both lexicographers and developers know how to handle it. In figure 7, a sample screen is reproduced: the top part contains pointers to the available functions, and a clickable alphabet. The left frame contains candidates, the right frame additional information; our example shows inclusion candidates, listing lemmas and frequency according to the heads of compounds (here *-Hose*).

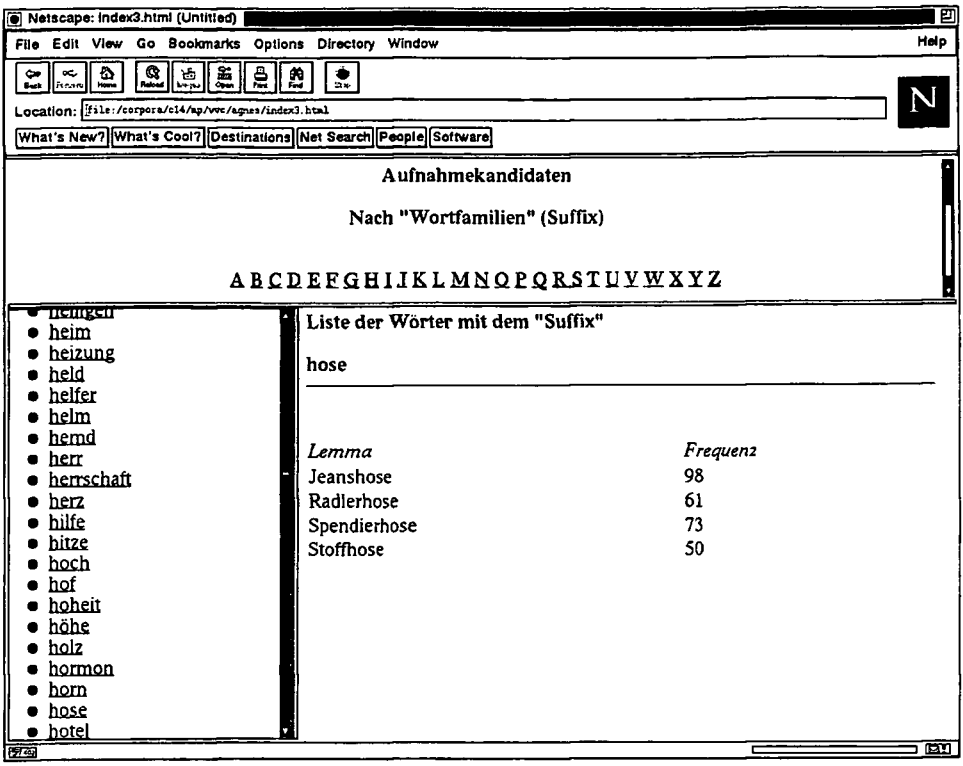

Figure 7: A GUI for the inspection of data from corpus and dictionary: compound inclusion candidates with the head *-Hose*

4. Conclusion

We reported on tools for comparing macrostructure and microstructure of a commercial dictionary with German text corpora, and we exemplified our results with data from *Langenscheidts Handwörterbuch Englisch.*

In 1986, SCHAEDER said that the use of computational techniques "*kann* [...] dazu beitragen, die Lexikographie ökonomisch effektiver und empirisch kontrollierter bzw. kontrollierbarer zu gestalten" (Schaeder 1986:266). In 1992, HEYN presented a metalexicographic analysis of a printed dictionary which covered the entire macro- and microstructure (Heyn 1992). What has been done here, is thus not very new in itself.

However, we have by now enough elements at hand, it seems, to computationally underpin metalexicographic dictionary analysis to a point where it can be practically used, and to combine it with standard low-level techniques for corpus exploration. The result is a series of comparative data which allows the lexicographer to more easily control the empirical side, as SCHAEDER said, of lexicographic work, making it economically more viable and leading to an improved product.

Many more possibilities for the combination of information from corpus and dictionary are open and yet unexplored: which examples are redundant? Which participles are lexicalized (and need to have separate entries), which ones can be derived from the verbs? How should complex collocations combining noun + verb and noun + adjective (e.g. *ehrendes Angedenken bewahren*) be treated?

If we assume that many dictionaries will in the future be mutiply queriable databases, the presentational side of these questions may become less important. Yet the tools provided here may also serve to restructure existing lexicographic material for electronic products.

5. Notes

[1] The WVC project (Wörterbuch validierung mit Corpora) described in this paper, was carried out in a cooperation between the institutions of the authors. The pilot study in 1996 also involved Judith ECKLE-KOHLER and Oliver CHRIST, of IMS-CL. The WVC project itself was carried out by Stefan EVERT (infrastructure, see section 2.1, section 2.5), Judith ECKLE-KOHLER (collocation extraction from corpora, see section 2.4), as well as Arne FITSCHEN and Agnès REY (data collections and WWW-browsable presentation of the data, see section 3.4). Wolfgang WALTHER, of Langenscheidt KG supervised the technical aspects related with the dictionary. They all deserve our gratitude for their contributions to the project.

[2] The dictionary has about 75,000 headwords in the German → English part.

[3] The corpora used in the project to test and enhance the coverage are mainly based on material available for research purposes: besides the LDC's European Languages News Corpus (ca. 100 M words for German) and the 2 years of *Frankfurter Rundschau* contained in the CD-ROM MC1 of the European Corpus Initiative, 2 years of *Stuttgarter Zeitung* and 7 years of *die tageszeitung* are used. We are aware that a more "representative" corpus would have been a much better and more trustworthy source of information. But, unlike for English, where the BNC would provide an excellent basis for the kind of work undertaken in the project, there is no sizeable corpus available for German which would give even cover of the written text production of the 1980s and 90s. We are aware that this biased corpus relativizes some of our results; but the methodological outcome of the project remains valid.

[4] See, for details on this topic, (Heid to appear)

[5] The table indicates the verb lemma, the noun form, a preposition if relevant (otherwise "nop" for "no preposition", since the dative and accusative case are not given separately), and the headword where the collocation was found.

[6] STTS stands for Stuttgart-Tübingen TagSet. STTS is compatible with and trivially mappable onto the EAGLES morphosyntax specifications ELM-DE (cf. (Teufel/Stöckert 1996)). It contains 54 tags with categorical, distributional and lexical distinctions (see (Schiller/Teufel/Thielen 1995)).

[7] See http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html. Tagging accuracy with STTS is around 97%. The examples from the dictionary did not cause particular problems.

[8] Details on the procedures can be found in (Eckle-Kohler/Heid 1996).

[9] This includes all verb-last cases (*weil...eine Rede gehalten hat*), but also verb-second cases with auxiliaries as finite verb forms (*...kann...eine Rede halten*). Since the main verb at the right sentence boundary may be accompanied by auxiliaries, a part-of-speech-modelling of all combinations of main and auxiliary verbs in such verb complexes is used in the queries.

[10] Light verb constructions tend to be of high frequency, thus come out top ranked if, for a given noun, we list all verbs which the noun can be the object of; for collocation listings, we can set a frequency threshold (e.g. 10 occurrences in the corpus). We plan to experiment with statistical measures for better separating typical collocations from trivial ad hoc combinations.

[11] *Haben* and *sein* may also appear in collocations (*Angst haben*); these cases are captured by a separate template set.

[12] Note, however, that collocational processing of a corpus of 100 M words is relatively time-consuming.

6. References

Elisabeth Breidt: "Extraction of v–n–Collocations from Text–Corpora: A Feasibility Study for German". In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. 2.6.1993, Ohio State University, Columbo.* Association for Computational Linguistics, 1993.

Oliver Christ: "The XKwic User Manual", internal report, Stuttgart: IMS, 1994.

Judith Eckle-Kohler: "Methods for quality assurance in semi-automatic lexicon acquisition from corpora"; in: *Proceeding of the* EURALEX *1998 International Congress (Liège)*, 1998.

Judith Eckle, Ulrich Heid: "Creating verb frequency lists for German", ms., technical report, Stuttgart: IMS, 1996.

Ulrich Heid, Judith Eckle: "Extracting raw material for a German subcategorization lexicon from newspaper text"; in: *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX '96*, (Budapest), 1996.

Ulrich Heid: "Finding Hidden Collocations: a computational analysis of a commercial dictionary", to appear in: Arne Zettersten (Ed.): *Proceedings of the ninth international Symposium on lexicography at Copenhagen university*, (Tübingen: Niemeyer).

Ulrich Heid: "Towards a corpus-based dictionary of German noun-verb collocations", in: *Proceedings of the 1998* EURALEX *International Congress (Liège)*, 1998.

Ulrich Heid, Vincent J. Docherty, Judith Eckle-Kohler: "Computational linguistic support for the revision of a commercial dictionary: corpus-based updates of entries and of collocation information" in: *Proceedings of the DGfS-CL Fachtagung: "Anwendungen in der Computerlinguistik"* (Heidelberg: IBM 1997).

Helbig, Gerhard: "Probleme der Beschreibung von Funktionsverbgefügen im Deutschen", in: Helbig, Gerhard: *Studien zur deutschen Syntax*, Bd.2, (Leipzig) 1984.

Matthias Heyn: *Zur Wiederverwendung maschinenlesbarer Wörterbücher*; Eine computergestützte metalexikographische Studie am Beispiel der elektronischen Edition des "Oxford Advanced Learner's Dictionary of Current English"; Lexicographica Series Maior 45; (Tübingen: Niemeyer); 1992.

Matthias Kammerer: *Bildschirmorientiertes Abfassen von Wörterbuchartikeln*; Dargestellt am Beispiel des Frühneuhochdeutschen Wörterbuches; Lexicographica Series Maior 68; (Tübingen: Niemeyer); 1995.

Brigitte Krenn: "Acquisition of Phraseological Units from Linguistically Interpreted Corpora – A Case Study on German PP-Verb Collocations". in: *Proceedings of the 3rd International Symposium on Phraseology*. Universität Stuttgart, IMS-CL, 1998.

Burkhard Schaeder: "Die Rolle des Rechners in der Lexikographie"; in: *Studien zur neuhochdeutschen Lexikographie VI.1*. Ed. v. Herbert Ernst Wiegand. Germanistische Linguistik 84-86, 1986. Hildesheim, Zürich, New York: Georg Olms 1986.

Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen: "Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS", Stuttgart/Tübingen, 1995.

Bruno Maximilian Schulze: *MP user manual*, Stuttgart: IMS, 1996.

Simone Teufel, Christine Stöckert: "EAGLES specifications for German morphosyntax", Stuttgart, 1996.