Supporting corpus-based dictionary updating

Stefan Evert, Ulrich Heid, Bettina Säuberlich
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstr. 12
D-70174 STUTTGART
GERMANY
{evert,heid,tina}@ims.uni-stuttgart.de

Esther Debus-Gregor
Langenscheidt KG, MÜNCHEN, GERMANY

Werner Scholze-Stubenrecht
Duden BIFAB AG, MANNHEIM, GERMANY

Abstract
This paper describes a collection of tools that provide support for updating mono-lingual dictionaries (as well as the source language part of bilingual dictionaries). The modules in this toolbox perform various tasks including the analysis of existing dictionaries, the extraction of frequency information for words and cooccurrences from text corpora, and the display and interactive manual annotation of extraction results.

1. Introduction

The use of corpus exploration tools is quite widespread in the lexicography of European languages. Tool support ranges from concordances and statistical tools (e.g. WordSmith tools, see http://www.lexically.net/wordsmith/) over publishers' in-house systems that embed concordancing and statistics in a customised interface (e.g. Walter & Harley, 2002), to "corpus digest" software as realized in the WASPS system (Kilgarriff & Tugwell, 2001; Kilgarriff & Rundell, 2002; see also http://wasps.itri.bton.ac.uk/).

The tools developed in the project *Automatische Exzerption* of the *Transferbereich 32 (TFB)*[1] go further than this, as they are designed to support dictionary updating by combining corpus-based lexical acquisition (in the sense of a "corpus digest") with the analysis of an existing dictionary and comparison of the data obtained from both sources. This parallel approach, whose purpose is to relieve lexicographers of the routine task of verifying which facts are already present in the existing version of a dictionary that is to be updated, has first been described in (Docherty & Heid, 1998). It is motivated by the observation that many more dictionaries are updated from a previous version ("new edition", "augmented edition", etc.) than are written completely from scratch.

Furthermore, like any "corpus digest" system, our toolbox aims at generalizing and abstracting corpus evidence. Instead of displaying a large number of structurally identical example sentences, as it occurs in a KWIC concordance, the tool should provide a description of the structure common to these example sentences, a frequency-based estimate of its relevance and just a few of the actual examples. This abstraction is then

compared with the data extracted from one or more indications in the dictionary and the comparison result is presented to the lexicographer in terms of proposed inclusion or removal candidates. Through texample sentences the lexicographers have access to the corpus data, but they can just as well limit themselves (and thus limit the effort they spend) to working with the abstractions derived from the corpus.

In this paper, we summarize current work on the corpus-based dictionary updating tools developed in the TFB project (see also the report on early experiences with a precursor of the toolkit in (Heid et al., 2000)), focusing in particular on the modular architecture of the tools available for German (Section 2) and the phenomena that are covered (Section 3), as well as on *LexiView*, an interactive graphical user interface supporting the manual work of lexicographers (Section 4, see also (Heid et al., 2004)).

2. A modular architecture for dictionary updating

Figure 1 below gives a schematic overview of the TFB dictionary updating system. Its input consists of dictionary and corpus data in electronic form. Modules for dictionary analysis as well as for lexical acquisition are used to abstract descriptions of linguistic phenomena from both sources. These are represented in an XML-based internal format to allow a comparison between corpus and dictionary data. The comparison result is again represented in XML, and submitted to the lexicographer via the *LexiView* interface. (As an alternative to the XML-based internal format, a database solution is currently being investigated.) The results of this interactive selection work are exported to the publisher's dictionary writing system.

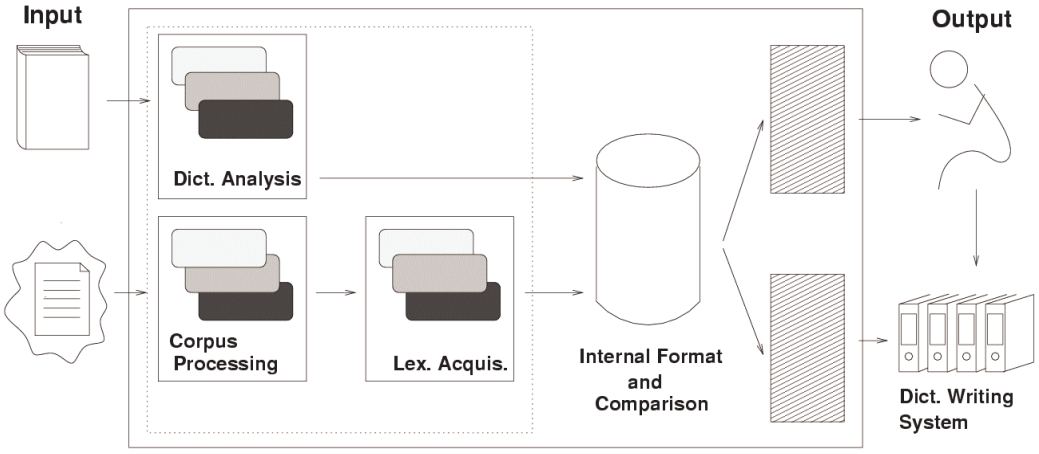

Figure 1: Schematic overview of the TFB system

2.1. Representing data from dictionaries and corpora

The existing version of a given dictionary is evidently the primary source of information for the updated version. One or more corpora should serve as the second major source of information. To be able to compare data from both sources, a flexible and sufficiently general representation is needed. To this effect, current work is based on an XML-based file format, and experiments with the use of a relational database are being carried out.[2]

The internal format is a data structure designed to represent linguistic information about lexical items and to keep track of the origin of such information. It contains the following basic data categories:

- lemma and word class (and a pointer to the information source);
- corpus frequencies of a given lemma with a given word class;
- an arbitrary number of linguistic properties of a given lemma with a given word class (and the respective source; examples are given in Section 3 below);
- collocations (and other significantly frequent word combinations) of a given lemma with a given word class (+ source), as well as the frequencies and, possibly, linguistic properties of these collocations (e.g. their preferences with respect to number or case, see (Evert, Heid & Spranger, 2004)).

In addition to pointers to the source of each element of information, example sentences taken from the corpus are also available (possibly in the form of a URLs or a similar kind of reference). Depending on the requirements of a particular dictionary updating project, any of the linguistic properties that are described (collocations, syntactic subcategorization frames, etc.) may also be illustrated with example sentences. Alternatively, only selected sentences may be given.

This data format is not intended to represent dictionary article structure. The description of certain types of linguistic facts (e.g. collocations) may appear in different types of indications of a dictionary, i.e. at different places in the microstructure: in a specific collocation item (as in the *Collins Concise German Dictionary*, 4th edition, 2003), in example sentences, in a section devoted to phrasal material of any kind (as in the *Van Dale Groot Woordenboek* series), or even hidden in the definitions themselves (as in *Cobuild)*. Reference to a given type of linguistic data in the internal format of the TFB tools is thus made by the type of phenomenon rather than indication type, even though the latter can be noted as sources.

Consequently, the TFB internal format is fundamentally different from recent attempts to propose standards for dictionary article formats (e.g., Ch. 12 of the TEI, (Mangeot-Lerebours & Andrés, 2002), etc.). On the other, and more important hand, the format is open to extensions at the level of supported data categories: when a particular category needs to be dealt with for a given dictionary, it can be added without much difficulty.

2.2. Dictionary and corpus analysis

The analysis of both sources relies on standard technologies, which are embedded as modules in the overall toolbox.

Corpus analysis is based on state-of-the-art natural-language processing technology: texts are tokenized, part-of-speech tagged, lemmatized and chunk parsed with a recursive chunker (Kermes, 2003). As an alternative to chunking, full syntactic analysis with a probabilistic parser is used for certain types of data such as noun + verb collocations (Zinsmeister & Heid, 2004). Any lexical acquisition tool could in principle be used in the TFB setup, provided that its output can easily be transformed into the internal format.

Dictionary analysis has to deal both with formalized and non-formalized indications (Heid et al., 2000: 185): formalized indications are translate directly to attribute-value structures in the internal format, whereas non-formalized ones are subjected to the same analysis as corpus material (possibly with specialised extraction rules for dictionary definitions etc.). In addition, there are tools that resolve and normalize the most frequent types of lexicographical text condensation. As each dictionary series has its own representation

format (nowadays often in XML or at least SGML encoding), the dictionary analysis tools are specific to a given dictionary or dictionary series. Perl scripts and XSLT stylesheets have proven useful for this purpose.
The components for the analysis of both sources are thus modular. For each dictionary, the types of linguistic data to be verified can be determined (from the list discussed in Section 3) and the extraction tools can be configured accordingly. As the internal format is generic and highly flexible, dictionary (updating) specifications can be defined according to the needs of the publisher. Furthermore, different corpus-based acquisition components can be plugged into the system and the results of different tools can be pooled.

3. Linguistic coverage

3.1. Principles

The comparison between data abstracted from the dictionary and data extracted from corpora is carried out automatically. It is based on the frequency and significance of the targeted phenomena and results in two types of proposals for changes in the dictionary, in addition to the raw quantitative data:

- Inclusion candidates:
  Items which are prominent in the corpus, but missing in the dictionary;
- Potential removal candidates:
  Items of a certain kind that are contained in the dictionary but are not prominent in the corpus and may be removed from the next version of the dictionary.[3]

Inclusion and removal candidates concern both macrostructure ("new (head-)words") and microstructure. Since most of our lexical acquisition tools are designed for the German language, update tasks have so far targeted mainly German monolingual dictionaries as well as the German part of bilingual dictionaries (for the translation from German to a foreign language). It should be noted that the tools cannot keep track of contrastivity, which is an important criterion in bilingual lexicography. Therefore, the final selection of appropriate material for a bilingual dictionary must be made by the lexicographer, although the candidate lists may provide some guidance.
Tools for the Dutch language are being developed at the moment (corpus processing, including chunk parsing, (cf. Spranger, 2002)) and have been tested on the macrostructure of the Dutch part of a small bilingual dictionary Dutch → German (*Langenscheidt Taschenwörterbuch Niederländisch,* ca. 20,000 lemmas).
The creation of similar tools for other languages depends critically on the corpus processing infrastructure available for these languages.

3.2. Macrostructural updates

New entries proposed by the system may belong to all word classes, but most often, nouns and noun compounds are suggested, as these account for the largest part of any (German) text. Several procedures are used to avoid names in the lists of inclusion candidates: typically, only general language items are relevant for the macrostructure of a dictionary. These tools include a large database of proper names as well as general structural patterns for names (similar to those used in named entity recognition tools, e.g. *Dr first_name last_name*). Similarly, abbreviations can be filtered out or extracted specifically together with their expansion, to serve as input for abbreviation lists.

3.3. Microstructural updates

The information programme of a dictionary, i.e. its intended use and user group, essentially determine the inventory of linguistic phenomena that have to be considered in dictionary updating. Furthermore, the size of the targeted dictionary is an important parameter: for instance, certain rather specific phenomena may only be relevant for a large dictionary. The currently available tools for microstructural updates analyse the morphological, syntactic and collocational properties of words, as detailed in Table 1.

Level of description	**Phenomenon**	**Example (+ gloss)**
Morphosyntax	Number preferences of nouns	*Lebensverhältnisse* "living conditions": typically pl.
	Distribution of adjectives: predicative vs. attributive	*gestrig(e)* "of yesterday": only attrib.
	Corpus frequency of inflectional variants	*Cellos* vs. *Celli* (pl.)
Syntax	Subcategorization	
	- of verbs	*anbieten* "to offer": (subj obj indir-obj)
	- of adjectives	*unklar* "unclear": + *daß*-clause (topicalized)
	- of nouns	*Bestrebungen* "efforts": + *zu* + INF
Collocation	Noun + Adjective	*billig* + *Imitation* "cheap imitation"
	Noun + Verb	*Hund* + *ausführen* "to walk the dog"
	Verb + Adverb	*tief* + *schlafen* "to sleep deeply"
	Adjective + Adverb	*tief* + *rot* "deep red"

Table 1: Major data categories in microstructural updates

In addition to the data categories listed in Table 1, a few others have been explored in the course of separate experiments, such as significant word triples (which often result from the combination of two collocations, e.g. *scharfe Kritik üben* "criticize massively" (cf. Zinsmeister & Heid, 2003)), collocation-like combinations of verbs and auxiliaries (*jmdn nicht mehr sehen können* "to be fed up with sb", *sich sehen lassen* "to show up", etc.) or the positional and collocational preferences of adjectives taking *daß*-clauses (*ob …, bleibt unklar* "it remains unclear whether …" (see Heid & Kermes, 2002)).

An important aspect of collocational information are the morphosyntactic preferences of collocations. For example, *Rolle* "role" in *eine … Rolle spielen* "play a … role", often combines with an adjective (indicated by the ellipsis) into a word triple. In addition, the combination has a massive preference for the singular. Thus, in the dictionary, we need to have *eine besondere, wichtige, zentrale, wesentliche Rolle spielen; eine geringe, untergeordnete Rolle spielen*. Some of the morphosyntactic preferences in collocations are on the verge of idiomatization – or at least many lexicographers would see two different readings of *Schritt* in (i) *ein gewaltiger, historischer, bedeutender Schritt* ("step", all with

a preference for singular), as opposed to (ii) *gerichtliche, juristische, rechtliche Schritte* ("measures", all with a marked preference for plural). See (Evert, Heid & Spranger, 2004) and (Evert, 2004) for more details.

As can be seen from the list of phenomena above, the TFB tools do not support the acquisition of lexical semantic data. This has to do with the fact that an automatic mapping of corpus data onto the readings in a dictionary is impossible in the general case, at least with the currently available tools and resources. Statistical methods could provide information about tendencies, but in the course of the project it turned out that most lexicographers prefer yes/no-statements over probabilistic ones. A drawback of this choice is that all microstructural data for a given lemma (of a given category) are returned as a single collection without internal structure, and not separated according to readings.

4. Interactive work with the system: LexiView

LexiView is a graphical user interface for interactive work with the results of the comparison between dictionary and corpus. It is implemented in Java, so that it works on a wide range of platforms.

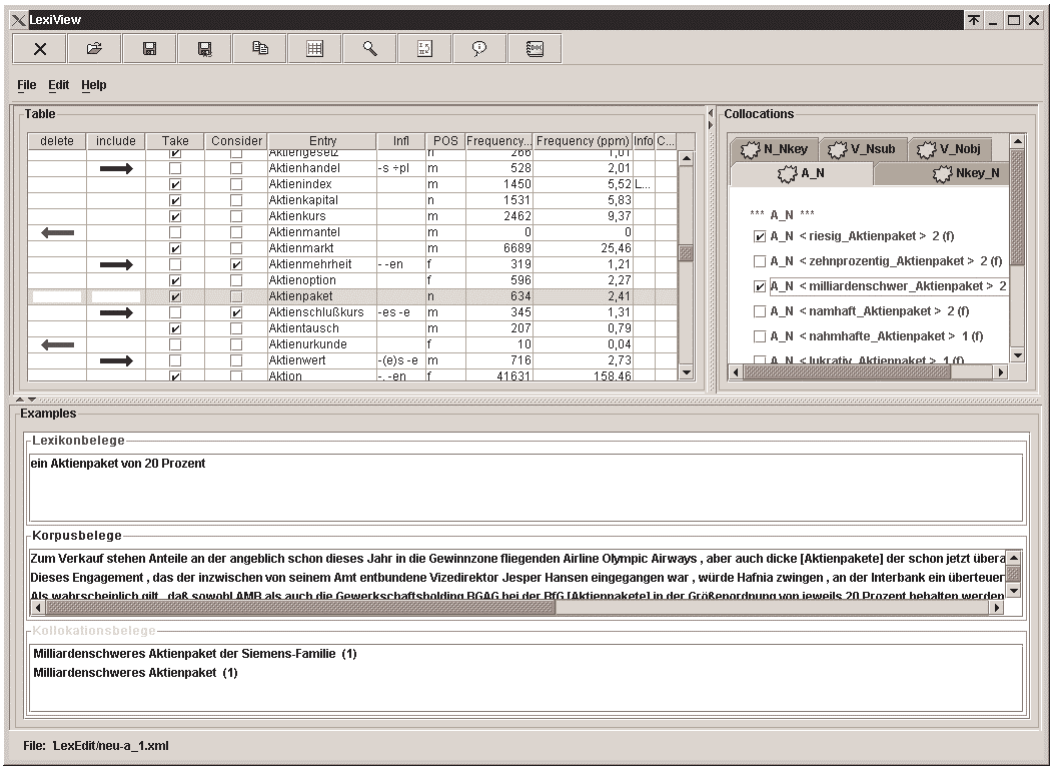

Figure 2: A screen dump from *LexiView*, showing information on the noun *Aktienpaket* "equity stake" (morphological information, corpus frequencies, collocations with adjectives, and examples) as well as candidates for inclusion (→) and removal (←)

In the first place, *LexiView* is used to display lists of inclusion and exclusion candidates at lemma level. Typically, the existing nomenclature of the dictionary as well as the inclusion and removal candidates are presented in one common, alphabetically sorted list. Candidates for inclusion or removal are indicated by means of colour coding or e.g. an arrow pointing to the right ("in") or to the left ("out"). The tabular listing of lemmas (the

top left pane, marked “Table”, in the screenshot in Figure 2) may include (formalized) indications pertaining to morphosyntactic and distributional properties of the lemma.
Secondly, *LexiView* provides access to syntagmatic data (e.g. collocations, subcategorization frames), which may be of different types: for example, collocations are classified according to the word classes of their elements, as illustrated in Table 1 above. For these data, corpus frequencies (or cooccurrence probabilities) are indicated and used as a sorting (and inclusion) criterion (see the top right pane, marked “Collocations”, in Figure 2).
Thirdly, example sentences extracted from the corpus both for the lemma and for each syntagmatic fact observed can be displayed, as well as examples from the dictionary. These can be copied to other applications, e.g. a text processor used to write new dictionary articles.
LexiView is generic and user-configurable. Except for a few fields, most of its layout and contents can be customized, according to which types of information were produced by the previous steps and according to the needs of the lexicographer. The fields in the “Table” pane can be edited manually, and comments can be added to each lemma. In addition, the rows can be sorted alphabetically or by frequency, columns can be resized and reordered, and new editable columns can be added.
Each lemma has one or more checkboxes: the lexicographer can select or unselect items from the list for the nomenclature of the new dictionary. Default selection of words (e.g. those above a given frequency threshold) is possible, but has to be performed with a separate tool before the data are loaded into *LexiView*. The same selection mechanism is available in the syntagmatic section (the “Collocations” pane). When the lexicographer has selected the items and the syntagmatic data needed for the new version of the dictionary, the data can be exported to a file in XML format, in a text-based format, or (by use of an XSLT stylesheet) in the format required by the publisher’s dictionary writing system. By means of this exportation facility, *LexiView* – and through it the whole dictionary updating chain – can be brought closer to the lexicographers’ production tools, without needing to interface directly with these.

5. Conclusion

The TFB tools for dictionary updating are a modular tool suite combining different kinds of computational lexicographic procedures: dictionary analysis, corpus-based lexical acquisition, a comparison of corpus-derived data with an existing version of the dictionary to be updated, and a comfortable interactive user interface for manual selection of inclusion and removal candidates. The modularity of the toolbox lies in the fact that the types of linguistic phenomena to be covered can be selected according to publishers’ needs, as can the lexical acquisition tools to be applied for the extraction of data of a certain kind from a corpus. Work on German has been based on ca. 350 million words of German newspaper text so far, but other material can also be used when it becomes available. Moreover, the architecture of the TFB toolbox allows for a relatively easy extension to other languages (again, only for monolingual description).
With the help of a database-driven implementation of the comparison between corpus and dictionary data, we expect that we will also be able to provide lexical profiles with respect to different corpus sources. For this application, geographically or sociolinguistically different corpora, corpora for specific domains of knowledge, etc. need to be available.

First experiments are being carried out with German newspaper text from Switzerland and Austria, as well as with Dutch from Belgium and the Netherlands.
Lexical acquisition still needs to be improved with respect to its precision and recall, but also with respect to the types of lexicographically relevant data that can be extracted from corpora. For subcategorization, Spranger (2004) is working along these lines.
In the medium term, the tools described here, which are oriented exclusively towards lexicographers, may also give rise to new lexicographic products for the end user, including a new way of presenting information about linguistic properties of lexical items.

Endnotes

[1] The project was a cooperation between the publishing houses Langenscheidt KG, München, and Duden BIFAB AG, Mannheim, on the one hand and the Institut für maschinelle Sprachverarbeitung (IMS) at the Universität Stuttgart on the other. We gratefully acknowledge the financial support granted to IMS by the Deutsche Forschungsgemeinschaft, DFG, from 10/2001 to 12/2003, under its *Transferbereiche* programme.
[2] A database solution offers the possibility to use an arbitrary number of corpora, without the need for additional representational devices.
[3] We are aware that frequency is only one criterion among many to decide upon inclusion or removal, which depend on the profile of the intended user. However, size restrictions often force lexicographers to restrict the nomenclature of a dictionary to, say, 40,000 items. Frequency data from several hundred million words of text provide a useful (but not exclusive) selection criterion in this case. Work reported in (Zinsmeister & Heid, 2004) may lead to an automatic classification of German noun compounds into lexicalized ones vs. productively formed words. The former tend to be semantically opaque, whereas the latter are semantically predictable. Under certain conditions, especially when there are tight space constraints, the former may be included in the dictionary, whereas the latter may be left out. In general, however, we do not believe that the selection procedure can be automatized.

References

Atkins, B. T. S. 1992. 'Tools for Computer-Aided Corpus Lexicography: the Hector Project', *Acta Linguistica Hungarica* **41**(1-4), 1992-93, 5 – 71.

Docherty, V. J. and Heid, U. 1998. 'Computational Metalexicography in Practice – Corpus-Based Support for the Revision of a Commercial Dictionary' in *Proceedings of the 8th Euralex International Congress*, Liège, 333 – 346.

Dunning, T. 1993. 'Accurate Methods for the Statistics of Surprise and Coincidence', *Computational Linguistics* **19**(1), 61 – 74.

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.

Evert S.; Heid U.; Spranger, K. 2004. 'Identifying Morphosyntactic Preferences in Collocations' in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Heid, U.; Säuberlich, B.; Debus-Gregor, E.; Scholze-Stubenrecht, W. 2004. 'Tools for Upgrading Printed Dictionaries by Means of Corpus-Based Lexical Acquisition' in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Heid, U.; Evert, S.; Docherty, V.; Worsch, W.; Wermke, M. 2000. 'A Data Collection for Semi-Automatic Corpus-Based Updating of Dictionaries' in Heid, U.; Evert, S.;

Lehmann, E.; Rohrer, C. (eds.), *Proceedings of the 9th Euralex International Congress*, Stuttgart, Germany, 183 – 195.

Heid, U. and Kermes, H. 2002. 'Providing Lexicographers with Corpus Evidence for Fine-Grained Syntactic Description: Adjectives Taking Subject and Complement Clauses' in *Proceedings of the 10th Euralex International Congress*, Copenhagen, Denmark.

Kermes, H. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. PhD Thesis, IMS, University of Stuttgart (AIMS, volume 9, number 3).

Kilgarriff, A. and Rundell, M. 2002. 'Lexical Profiling Software and its Lexicographic Applications – A Case Study' in *Proceedings of the 10th Euralex International Congress*, Copenhagen, Denmark, 807 – 818.

Kilgarriff, A. and Tugwell, D. 2001. 'Word Sketch: Extraction and Display of Significant Collocations for Lexicography' in *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, 32 – 38.

Mangeot-Lerebours, M. and Andrés, F. 2002. 'An XML Markup Language Framework for Lexical Database Environments: the Dictionary Markup Language' in *Proceedings of the LREC Workshop on International Standards of Terminology and Language Resources Management*, Las Palmas, Spain, 37 – 45.

Schulte im Walde, S.; Schmid, H.; Rooth, M.; Riezler, S.; Prescher, D. 2001. 'Statistical Grammar Models and Lexicon Acquisition' in Rohrer, C.; Roßdeutscher, A.; Kamp, H. (eds.), *Linguistic Form and its Computation*, Palo Alto: CSLI Publications, 387 – 440.

Spranger, K. 2002. *A Lexically-Informed Chunking Analysis as a Starting Point for the Extraction of Linguistic and Terminological Information from Dutch Text.* Unpublished Master's Thesis, IMS, University of Stuttgart.

Spranger, K. 2004. 'Beyond Subcategorization Acquisition – Multi-Parameter Extraction from German Text Corpora' in *Proceedings of the 11th Euralex International Congress*, Lorient, France.

Walter, E. and Harley, A. 2002. 'The Role of Corpus and Collocational Tools in Practical Lexicography' in *Proceedings of the 10th Euralex International Congress*, Copenhagen, Denmark, 851 – 857.

Zinsmeister, H. and Heid, U. 2003. 'Significant Triples: Adjective+Noun+Verb Combinations' in *Proceedings of Complex 2003*, Budapest, Hungary.

Zinsmeister, H. and Heid, U. 2004. 'Collocations of Complex Nouns' in *Proceedings of the 11th Euralex International Congress*, Lorient, France.