Software Demonstration:
Computational linguistic tools for semi-automatic corpus-based updating of dictionaries

Ulrich HEID, Wolfgang WORSCH,
Stefan EVERT, Vincent DOCHERTY, Matthias WERMKE,
Stuttgart – München – Mannheim, Germany

Abstract

We will demonstrate an interface which allows the lexicographer to view the results of an automatic comparison of lexicographic descriptions from existing German dictionaries with corpus data. The second part of the paper will discuss in detail the use made of the raw material in the recent update of Langenscheidt's *Großwörterbuch Deutsch – Englisch, Der kleine Muret-Sanders*. The examples in the online-demonstration come from work on entries for headwords with the initial letter "T" in *Duden. Das große Wörterbuch der deutschen Sprache* (8 vols.; Duden GWDS) and from the German part of *Langenscheidts Handwörterbuch Deutsch-Englisch* (HWB). Both have been compared with data extracted from large newspaper corpora. The interface makes use of a standard web browser for display of lexical data.

The demonstration will be a guided tour of the data collection, from the lexicographic point of view. The first part of this paper provides the metalexicographic baseline, a short summary of the technology used to develop the data collection, a few examples of the types of data made available. The second part deals with the practical lexicographic use of the data collection in the update of *Der kleine Muret-Sanders*.

1 Introduction: the Problem

This demonstration deals with tools for dictionary updating. A large part of the activity of lexicographers is devoted to updates of existing dictionaries (often more time than to writing dictionaries from scratch): new editions typically contain more data than their predecessors, both with respect to the macrostructure and to the microstructure. Often, the objective of dictionary editors is also to provide more up-to-date material, sometimes creating space for new (uses of) words by removing less frequent ones from the existing version of the dictionary.

Next to a "sequential" update of the dictionary (carried out by working along the microstructure, in alphabetical order), sometimes there is a need for "thematic" updates, i.e. for the checking, addition or enhancement of one or more types of information only (e.g. syntactic valency, collocations, morphosyntax, specialized language uses, etc.).

How can these two kinds of dictionary updates be supported by computational linguistic tools for corpus processing? How can we integrate existing dictionary data with texts and with computational linguistic text analysis tools (and their results)? And how can such material, and the resulting synopsis facilitate the creation of new versions of dictionaries?

We will demonstrate an interface which allows the lexicographer to view the results of an automatic comparison of lexicographic descriptions from existing German dictionaries with corpus data. The examples in the demonstration come from work on entries for headwords with the initial letter ”T” in *Duden. Das große Wörterbuch der deutschen Sprache* (8 vols.; Duden GWDS) and from the German part of *Langenscheidts Handwörterbuch Deutsch-Englisch* (HWB). Both have been compared with data extracted from large newspaper corpora. The interface makes use of a standard web browser for display. The computational linguistic background has been discussed in [Docherty/Heid 1998]. We here concentrate on the tools and the workflow, as well as on an assessment of the resulting data with respect to its usefulness for the updating of a bilingual dictionary.

The remainder of this paper is organized as follows: in section 2, we discuss the kinds of data, both from the existing dictionary and from the corpus, which have so far been compared; for each item type (“Angabetyp”, in WIEGAND’s terminology) from the dictionary, we discuss whether and how it can be extracted from the dictionary, and how to extract comparable illustrative material from the corpus (section 3). We then show a few examples of the presentation of the results (section 4, more examples will be given in the demonstration, which will be a guided tour of the data collection), before concluding on a few proposals for the architecture of a more general system of this kind.

2 Metalexicographic basis

2.1 Objectives of dictionary updates

The dictionary updates we have worked on have an impact on both the macrostructure and the microstructure of the dictionary.

Typically, *macrostructural updates* concern additions to and removals from the headword list. We consider corpus frequency as just one criterion for inclusion or exclusion of items among others. Words with high lemma frequency in the analyzed corpora, but absent from the dictionary, are proposed for inclusion in the headword list. Words from the dictionary which do not appear in the corpora or very rarely, may be considered for removal[1].

The objective of *microstructural updates* is typically the following:

- Adding (and maybe removing) example sentences;
- adding (and possibly removing) collocations;
- verifying (and possibly correcting or adding) morpho-syntactic classifications, such as indications about syntactic valency, about distributional properties, such as the preferred or exclusive use of nouns in the singular or plural, etc.

The above are just a few examples of types of data (and the pertaining item types in the dictionary text) which may need to be updated[2].

2.2 Item types analyzed

Dictionaries use different types of presentational devices to convey information about linguistic phenomena. We do not attempt to provide a metalexicographic classification of item types, here, but for our work, the rather simplistic distinction between formalized and non-formalized items has proven useful.

Certain facts are presented in *formalized items*, comparable to attribute-value-pair descriptions: a (usually fixed) set of abbreviations or other notational devices is used to indicate a certain property of an item. In this way, preferences for singular or plural use of nouns are indicated by “ohne Plural”, “Plural selten”, “meist Plural”, “nur Plural” in the grammar items of Duden GWDS; in HWB’s verb entries, complement nouns (forming collocations with the verbs) are indicated in parentheses within items giving equivalents.

Next to formalized items, the analyzed dictionaries contain *non-formalized items*, such as (made-up) example sentences, citations from the literature, from newspapers, etc.[3] These are often used by the authors of the dictionary articles to illustrate syntagmatic phenomena, such as collocations and valency, which are not coded in a formal way (as it is the case in both analyzed dictionaries); often one example sentence illustrates several phenomena. Duden GWDS follows a specific guideline for the morphosyntactic form of subject-verb collocations and verb-complement collocations. We call the resulting example items *semi-formalized*, since they are textual, but mostly ‘controlled’.

Both item types need to be analyzed, but in different ways. To compare the dictionary and the corpus, we need to produce abstractions from the descriptions available in either source, and then compare data at the level of these abstractions (see below, section 3.3).

3 Extracting and comparing data from corpus and dictionary

In the following, we will concentrate on microstructural data[4]. Table 1 summarizes the item types from Duden GWDS which have been analyzed, for distributional, valency and collocation information; in addition, table 1 contains a column indicating the kind of evidence gathered from corpora for use in the comparison.

Type of ling. information	Dictionary: Item type	Corpus: Evidence
Distribution: Preferences for Sg./Pl.	Formalized Items	Frequency Data for word forms
Valency: Verb, adj. or noun subcategorization	Formalized Items (vt...) and Example sentences	Example sentences
Collocation: (N+V, N+A,...)	Semi-formalized Items: syntagms with fixed structure	Example sentences and Frequency Data

Figure 1: Types of information, item types and corpus evidence

As certain non-formalized item types are very similar to the corpus data, the same tools for extraction of linguistic evidence (and for the automatic classification of the extracted material) can be used wherever both, dictionary and corpus, only provide example sentences.

3.1 Corpus processing infrastructure

In our experiments, all corpus-derived evidence was extracted from tokenized, part-of-speech-tagged and lemmatized German news corpora. Table 2 indicates the amounts of material available from the major sources, a total of 300 million occurrences.

Corpus	Period	Years	Word forms
Frankfurter Rundschau	1992/93	2	40 M
Stuttgarter Zeitung	1992/93	2	36 M
die tageszeitung	1987-93	7	103 M
'European News Cp.': dpa, afp...	1990-94		100 M

Figure 2: German news corpora used in the comparison

The extraction of illustrative material for the types of information given in table 1 is performed as follows:

- *Distribution*: nouns, adjectives and determiners are annotated with disjunctive sets of agreement features (for German: case, number, gender and definiteness, as it can be derived from the inflected form of certain adjectives). Within nominal groups, modelled by means of part of speech shapes (eg. "DET (ADJ) N"), clearly identifiable singular and plural uses of nouns are extracted, counted and the frequency figures for singular forms and plural forms are compared.

- *Valency*: a cascade of extraction templates, based on the annotated types of information (sentence boundaries, part of speech, lemma), as well as on a modelling of sentence structure in terms of topological fields is used to "learn" valency patterns from the corpus text, by exploring those contexts which can be unambiguously interpreted[5]. The resulting data are used in the comparison.

- *Collocations*: Noun-adjective-, adjective-adverb-, and verb-adverb-collocations are extracted as pairs of adjacent words by means of part-of-speech sequences and subsequent ranking with respect to the "strength of their association" (by means of the log-Likelihood measure, cf. [Dunning 1993]). As noun-verb collocations do not always show up in adjacent form, a more sophisticated cascade of extraction templates is used (for details, see [Heid 1998]).

Distribution and collocation data are provided along with frequency figures from the corpus.

3.2 Extracting data from the dictionary

Formalized item types require a metalexicographic analysis and reinterpretation, and subsequent reformatting; in this way, an abstraction from the indications given in the dictionary is created[6].

Non-formalized and *semi-formalized item types* are analyzed with the same partial-parsing routines as the corpus material. A certain amount of specialization of the extraction routines is advantageous, for the analysis of semi-formalized collocation indications: Since, for example, the Duden GWDS uses a standard syntagmatic structure to illustrate certain types of noun-verb collocations, and since its illustrative syntagms are limited to exactly the relevant chunks, a special set of extraction templates can be designed to handle such cases.

For practical reasons (to make the comparison with data abstracted from corpus observation easier), the resulting data are stored in a relational database; currently, only data from formalized item types are stored this way, but it is planned to extend this to all data extracted from the dictionary and from the corpus.

3.3 Comparing data abstracted from the corpus and from the dictionary

The actual comparison is in all cases performed on abstractions, derived from the two types of sources, dictionary and corpus. These are represented in a common format, e.g. for collocations, for valency frames, etc.

In many cases, the comparison consists in checking the presence or absence of objects of a certain type in the two parts of the data collection, the repository of dictionary-derived data, and the inventory of corpus-derived data. Databases provide these comparison facilities in a straightforward way.

In the case of distributional preferences, the indications given by the dictionary (“only plural”, “mostly plural”, “rarely plural”) have to be related with frequency ranges or proportions of relative frequency of singular and plural forms observed.

4 The presentation of the results

4.1 Distributional Data

Data about the use of singular and plural forms of nouns are presented in tables derived from the database, sortable according to the alphabet, to the observed frequencies or to the marks found in the dictionary. In figure 3, a few lines from such tables are reproduced[7].

Lemma	ID	POS	GWDS-mark	% Sg.	% Pl.	total f
Tabakwaren	161246	NN	meist pl.	0	100	133
Teigware	162960	NN	meist pl.	11	89	63
Tätlichkeit	162413	NN	meist pl.	58	42	271
Tabellenstand	161265	NN	ohne pl.	91	9	81
Traumland	166882	NN	ohne pl.	94	6	69

Figure 3: Sample entries from the tables on singular/plural distribution

Note that *Tabellenstand* and *Traumland*, for example, did occur in the plural, in our corpora; the indication “ohne Plural” may be changed into “meist Sg.”.

4.2 Valency Data and Collocations

Results of the syntagmatic analysis are displayed both in a text-based (printable) table format and in a colour-coded browsable version. In both cases, a standard representation of the linguistic phenomena, an indication of the source where the pheonomenon occurred (dictionary, corpus, both), and (in the case of collocations) frequency data are indicated. In figure 4, a few noun-verb-collocations with the collocate *trüben* are given.

```
Source   Verb     Noun              Syntax  Freq.
NURCP    trüben   Freude            trans.    6
NURWB    trüben   Ruhm              trans.    1
NURCP    trüben   Stimmung          trans.    5
NURWB    trüben   Urteilsvermögen   trans.    1
NURCP    trüben   Wässerchen        trans.    9
```

Figure 4: Noun-verb-collocations with the collocate *trüben* in a table-like format

5 The Use of Language Corpora in practical Lexicography: A Project Report

In the following, we intend to demonstrate the practical use of the results of the procedures explained above.

The concrete subject of this section is *Langenscheidt’s Großwörterbuch Deutsch - Englisch*, or *Der Kleine Muret-Sanders, Band II*. The description reflects the state of compilation and editing in April 2000.

5.1 Context

Originally the *Kleiner Muret-Sanders* dates back to *Langenscheidt’s Enzyklopädisches Wörterbuch Deutsch - Englisch* (first published in 1901, the current four-volume edition published in 1962/1975). The *Großer Muret-Sanders* still is the most exhaustive English – German / German – English dictionary worldwide. It also served as the basis for the late Heinz Messinger’s *Gro"swörterbuch Deutsch - Englisch* (first edition 1972, current edition 1982), which is also known as *Der Kleine Muret-Sanders.*

Back in 1995 we began a complete revision of the English – German volume and started to plan the completely revised edition of the *KMS II* as the inhouse jargon has it. At a rather early stage, we realized that we had to do more than just a *normal* revised edition or updating. The lexicographical core of the dictionary has its roots in the *Enzyklopädische Wörterbuch,* the *Gro"se Muret-Sanders.* So the macrostructure as well as the microstructure reflected a lexicographical standard which we felt did no longer meet the requirements of the quality-conscious users of the late 1990s.

The macrostructure and phraseology reflected the German and English languages of the 1970s. This meant that the complete language development of the 80s as well as 90s had to be incorporated in the new edition. The rapid innovations in the fields of economy, science, ecology, electronics, information scienes, the media etc. over those 20 years since the last revision had brought along more new words, coinages, meanings and word combinations than never before in a comparable period of time.

Also, there had been important new developments in organizing and presenting lexicographic information – the *age of the Power*[8] *principle* had come. And in addition we had to take into consideration that the new product would be published in print form, but also and – presumably of prime importance – on electronic data carriers.

It became quite clear to us that there was little point in the cosmetics of a mere revision or even a complete revision – what was required was a completely new edition.
For this reason we decided not to use the old version of the *Kleiner Muret Sanders* as the basis for the new edition – an otherwise normal procedure in updating an existing dictionary. Instead we decided to use an at that time highly up-to-date dictionary, the *Handwörterbuch Deutsch – Englisch* with over 120 000 words and phrases. Using this dictionary we aimed at a number of headwords and phrases well over 240 000.

5.2 Why corpus-based methods

A dictionary of the size of the *Handwörterbuch*, the German equivalent of most English *Concise Dictionaries,* reflects a fairly complete picture of the languages involved. To try to double its size required a reliable method of finding out which words, meanings and phrases were actually missing. At the same time we had to find ways that would enable us to label or even eliminate words, meanings and phrases which had become outdated or obsolete over the past 20 years.

What we really wanted was the opportunity to run a corpus-based semi-automatic comparison between the *Handwörterbuch* data and a vast, up-to-date corpus of German.

At that time we had no in-house databank of a size which would be able to meet our requirements. So we decided on a cooperation with the Institut für maschinelle Sprachverarbeitung in Stuttgart[9].

5.3 Using the Results: Finding candidates for inclusion in the dictionary

We received a number of lists showing us differences between the *Handwörterbuch* and the corpora used by IMS (see above, figure 2). Obviously of top relevance to us is the list showing words which are missing in the *Handwörterbuch.* These can either be words which were simply forgotten or for various reasons not taken up into the dictionary by previous compilers, or neologisms which were coined after the publication of the *Handwörterbuch.*

Of course this list still required careful lexicographical editing, since it contained language material which we did not need. For example, we decided not add **Zumdick*[10] to our list of potential addenda (see figure 5).

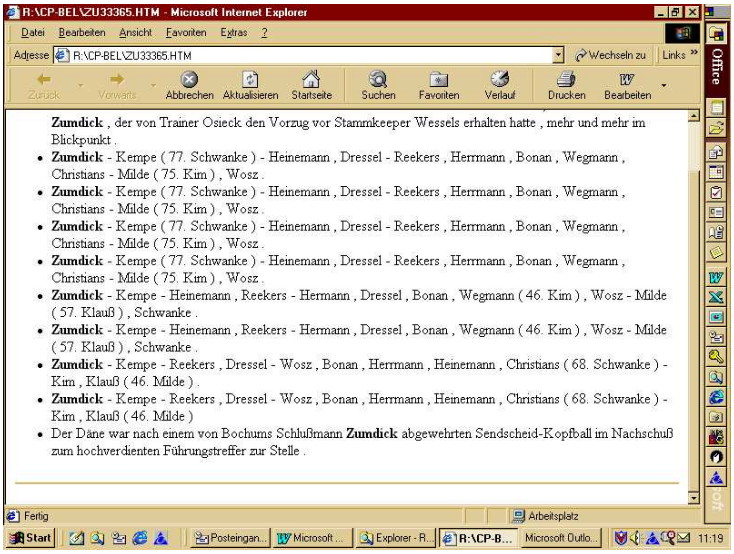

Figure 5: Evidence for *Zumdick*

We ended up with a list of possible addenda which was put at the compilers' disposal (cf. figure 6 for a screenshot from a typical work environment). It was and is up to them and their lexicographical expertise to decide whether to take up a word or leave it out.

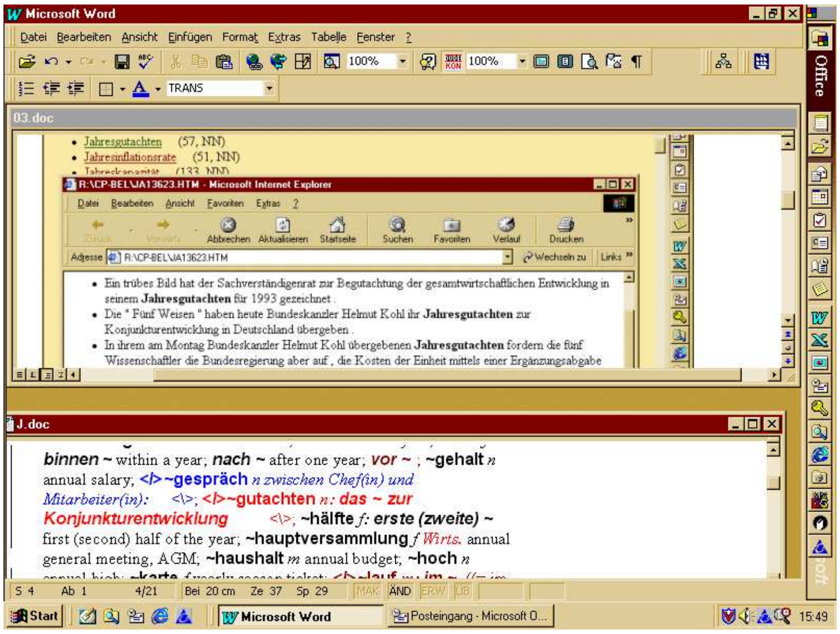

Figure 6: User Interface: Dictionary entry and corpus evidence

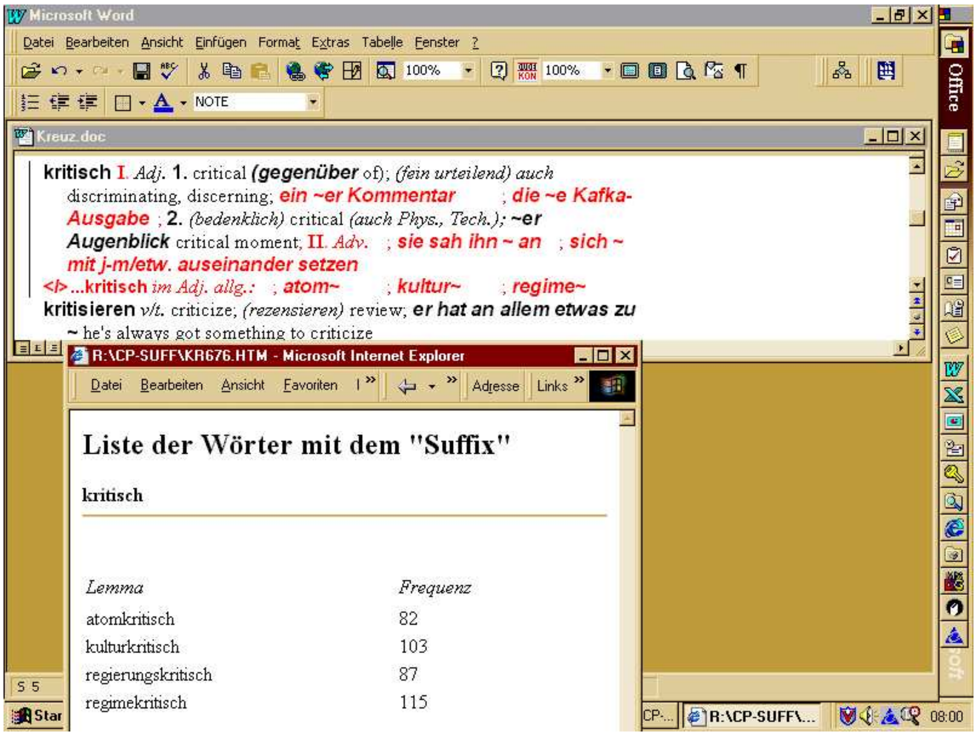

Figure 7: Inclusion candidates by morphological properties

It goes without saying that in this process we also performed a close a check of morphological elements such as prefixes and suffixes (inclusion candidates can be displayed by elements of compounds or by affixes, cf. figure 7).

This method of electronically checking a dictionary text against a vast language corpus has the additional advantage that you find real gaps, i.e. words which have long been forgotten or ignored by lexicographers. One word which does not feature as an entry in any dictionary – at least in the dictionaries we know – is the German word **Babypause**.

5.4 Using the Results: Collocations

Using sophisticated corpus tools also enables us to look at words in their typical context, i.e. to make sure that we include collocations which are the natural patterns of any language (as Broder Carstensen once put it).

Syntax	*Noun + Verb*	trans.	`Daten löschen`
trans.	`Daten aufnehmen`	trans.	`Daten nennen`
trans.	`Daten austauschen`	trans.	`Daten sammeln`
trans.	`Daten auswerten`	trans.	`Daten senden`
trans.	`Daten enthalten`	trans.	`Daten speichern`
trans.	`Daten erfassen`	trans.	`Daten übertragen`
trans.	`Daten erheben`	trans.	`Daten vernichten`
trans.	`Daten geben`	intr.	`Daten vorliegen`
trans.	`Daten liefern`	trans.	`Daten zeigen`

Figure 8: Collocations with the noun *Daten*

An example is the material available for the noun *Daten* displayed in figure 8; the graphical environment is reproduced in figure 9.

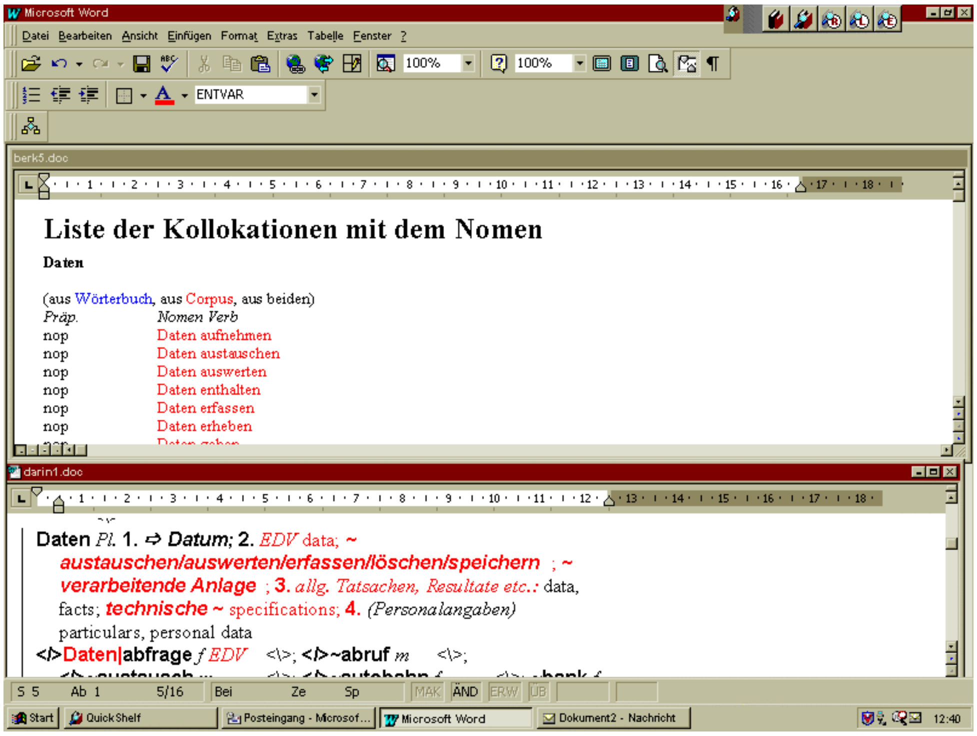

Figure 9: Working with the collocation lists for *Daten*

In editing existing as well as new dictionary entries it proved useful to have the opportunity to look at the use of a word in real language; below, in figure 10, a few corpus examples for *Quereinsteiger* are reproduced.

5.5 Using the Results: Finding dictionary corpses

The use of corpus tools not only showed us words to be taken up into the new *Großwörterbuch*, it also showed us dictionary corpses, old or rare words which miraculously had survived gen-

- Einer dieser Neulinge und **Quereinsteiger** ist der 30jährige Bogdan Mazuru aus Rumänien .
- Nicht angeben könne sie jedoch , wie hoch der Anteil der Kinder sei, die die Schulen nun als **Quereinsteiger** besuchen, aber mit ihren Eltern noch keine eigene Wohnung gefunden haben, sondern in Hotels und Wohnheimen leben.
- Der FDP-Fraktionschef im Bundestag, Hermann Otto Solms, sagte im Mitteldeutschen Rundfunk , der neue Wirtschaftsminister dürfe kein **Quereinsteiger** sein .
- Die Hochschule auf der Suche nach immer besseren Materialien Nach zwanzig Jahren wieder ein neuer Fachbereich an der Darmstädter TH 34 Erstsemester und 15 **Quereinsteiger**, die nach dem Grundstudium der Physik, Chemie oder einer Ingenieurwissenschaft wie Elektrotechnik und Maschinenbau nun mit dem Vordiplom in der Tasche im fünften Fachsemester weiterstudieren, haben sich für die neue Forschungsrichtung entschieden.

Figure 10: Evidence for *Quereinsteiger*

erations of lexicographers and revisions: *Immobilienmagnat* or *immobilisieren* are such words which used to be entries in the *Handwörterbuch* but had zero-evidence in the corpus.

5.6 Assessment

We see from the examples mentioned that the corpus data is of invaluable use to lexicographers. Using the tools enables us not to just rely on our language competence and language sensitivity when it comes to decide whether and how to include a word in a dictionary. With the available corpus evidence we have a clear picture of how a language works, and the corpus evidence at the same time enables us to present a clear picture to the users.

Looking at all this one might get the idea that the "Lexicographer's Creativity"[11] is no longer required, and might even be a hindrance. Is the dictionary publishers' dream to produce up-to-date dictionaries at the touch of a button without having to worry about paying salaries and additional wage costs about to come true?

We do not think so. To make this point let us give one more example, which shows that corpus evidence does not necessarily make the lexicographer redundant.

One spin-off of the Cobuild language database is the CD-ROM called *Cobuild English Collocations.* Looking up collocates for the noun **deadline**, you get in the frequency ranking in third place with 278 hits the date ***15th***. This unexpected information is due to the fact that shortly before the expiry of the UN ultimatum (or rather: deadline) against Saddam Hussein on 15 January 1991 plentyof newspaper text must have been scanned into the Cobuild database.

With a human filter between data input and output this slight "slip of data" could have been avoided.
We mention this incident of "data error" to make the point that apart from an increasingly sophisticated technology in gathering, processing and evaluating language data it is still up to the competence of the skilled lexicographer to produce correct dictionary entries.

In this production process advanced corpus tools are a great help, at some stage they will be (or already are) similarlay indispensable as the old and large foliant volumes like OED, Grimm or Larousse used to be – but they remain nothing more than tools.

Notes

[1]There are many other criteria for deciding about removal and inclusion of headwords. The work with Duden GWDS showed very clearly that a news corpus alone is insufficient for deciding about removal candidates, since many words, for example denoting objects of daily life, do not appear in the texts sufficiently often; corpus frequency then does not say much about the relevance of the respective items for the dictionary user. Similarly, in a bilingual dictionary, rare items may be kept because of their contrastive relevance.

[2]Not all item types can profit equally from automatic corpus analysis, in an update: evidently, current corpus linguistic tools have little to contribute (in an automatic way) to reading distinctions, diasystematic marks (“marques d’usage”), etc.

[3]We have not dealt, in the experiments described here, with definitions, illustrations of meanings and reading differenciation. Work on bilingual dictionaries has been restricted to the analysis of the German part of the dictionary, from a monolingual point of view.

[4]Results of the macrostructural comparison will also briefly be discussed in the demonstration; for a description with respect to the extraction problem, see [Docherty/Heid 1998].

[5]See [Eckle-Kohler 1999] for details of the extraction routines and the underlying linguistic assumptions. Since the presence of a certain valency pattern with a given predicate can be concluded from a minimal number of sentences that unambiguously illustrate the pattern, the emphasis in the corpus-based extraction, is on precision: we maximize the amount of correct valency hypotheses within the total of the hypotheses derived from the corpus. Consequently, no frequency data can be provided.

[6]We do not elaborate, here, on the manipulation of the dictionaries available in an SGML-annotated format. If a dictionary publisher holds descriptive data in a (pre-lexicographic) database, the SGML-based conversion step is not even necessary. Examples of the conversion steps can be found in [Docherty/Heid 1998].

[7]The lines are not taken in this order from the tables, but compiled from different tables, for the purpose of illustrating a few interesting cases.

The abreviations in the table should be decoded as follows: ‘ID’ = number of item; ‘POS’ = word class; ‘GWDS-mark’ = note on singular/plural use from GWDS; ‘% Sg.’ and ‘% Pl.’ = percentages observed in the corpus; ‘total f’ = total number of occurrences in 200 M words.

[8]Cf. Langenscheidt’s Power Dictionary Englisch and Langenscheidt’s Power Wörterbuch Franz"osisch.

[9]The computer data of the typeset text was transferred to Dr. Heid’s institute and there it was compared to the institute’s corpus of at that time roughly 300 millions lexical items. We saw the first preliminary results on the fringe of the Euralex conference at Gothenburg in 1996.

[10]ZUMDICK, as the football experts will know, was a reasonable goalkeeper who played for various clubs in the German Bundesliga, now acting as manager of VfL Bochum.

[11]This was the title of a plenary speech held by Ladislav Zgusta at the Euralex conference in Gothenburg in 1996.

References

[Docherty/Heid 1998] Vincent J. Docherty, Ulrich Heid: “Computational Metalexicography in Practice – Corpus-based support for the revision of a commercial dictionary”; to appear in: *Proceedings of the* EURALEX *International Congress 1998*, (Liège), 1998.

[Dunning 1993] Ted Dunning: “Accurate Methods for the Statistics of Surprise and Coincidence”. Computational Linguistics, 19/1, 61-74, 1993.

[Eckle-Kohler 1999] Judith Eckle-Kohler: *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*, (Berlin: Logos) 1999, ix + 402 SS.

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