Wolfgang Teubert

SETTING UP A LEXICOGRAPHICAL DATA-BASE FOR GERMAN

Introduction

In the years to come, the main target of the Section for Computational Linguistics at the Institut für deutsche Sprache in Mannheim will be the setting up of a Lexicographical Data-Base (LEDA) in order to ensure efficiency in the collection, analysis, ordering and description of linguistic material and to support lexicographical projects at the Institut and elsewhere.

First a few words on the Institut. It is a research institution founded in 1964; its purpose is research on and description of the contemporary German language. It is not affiliated to any university but directly funded by the Federal and State governments. Its regular staff today consists of about 90 persons including an academic staff of about 45. To handle large quantities of texts for research into spoken and written German it was necessary to use a computer, to assemble machine-readable corpora and to develop appropriate programs for corpus analysis.[1] The present main-frame computer, which was installed in January 1983, is a Siemens 7.536 with a core storage of 2 megabytes, a number of tape and disk decks and 15 visual display units for interactive use.

Whereas in former years most tasks were carried out in batch mode, the terminals now make it possible for the linguist to work interactively with the computer. It was therefore a logical step to devise a more or less integrated system as a tool for the compilation of new dictionaries. This Lexicographical Data-Base is aimed at facilitating access to text samples and dictionary data and at storing preliminary versions of lexical entries for further processing.

In any lexicographical project, once the concept for the new dictionary has been established, there are three major tasks in which the computer can be employed as a tool:

(1) For each lemma, text samples have to be determined in the corpus which is the linguistic base of the dictionary. As far as this search is to be carried out by the computer, the corpus will have to be machine-readable and there have to be programs for search procedures. The corpus and the programs to be applied to it will form one component of the Lexicographical Data – Base, the so-called Text Bank.

(2) For each lemma, the lexicographer will want to compare corpus samples with the respective lexical entries of existing relevant dictionaries. In so far as these dictionaries exist in machine-readable form, the lexical entries can be looked up at the terminal. This component we call the Dictionary Bank.

(3) Once the formal structure of the lexical entries in the new dictionary has been established, the lexicographer can begin describing the lemmata according to the framework of this structure, which is usually organized hierarchically in categories and subcategories of information units. A data-base will provide the structural frame for the provisional descriptions carried out by the lexicographer, thus enabling him to achieve homogeneity and to check the thesaurus of his explicative vocabulary. This component of the Lexicographical Data-Base we call the Result Bank.

The Text Bank

Each dictionary project needs a corpus assembled according to the specific requirements of the particular lexicographical goal. Existing machine-readable corpora can form part of it, but in most cases new texts have to be processed in machine-readable form, in accordance with the special codification or pre-editing requirements demanded by the analysis programs.

At present, a number of machine-readable corpora in unified codification are available at the Institut, including the Mannheim corpora of contemporary written language, the Freiburg corpus of spoken language, and the East/West German newspaper corpus, totalling altogether about 7 million running words of text.[2] Further corpora have been taken over from other research institutions, publishing houses and other sources. These texts had been coded in various different conventions, and programs had to (and still have to) be developed to transform them according to the Mannheim coding rules.

Other texts to be included in the corpus of the Text Bank will be recorded by OCR, via terminal or by the application of an optical scanner (Kurzweil Data Entry Machine), if they are not already available on machine-readable data carriers. We hope that by the end of 1984 texts of a total length of 20 million words will be available from which any dictionary project can make its own selection.

A special query system called REFER has been developed and is still being improved (cf. Brückner 1982). The purpose of this system is to ensure rapid access to the data of the Text Bank, thus enabling the lexicographer to use the corpus interactively via the terminal.

Thus the Text Bank allows the lexicographer to gain information of the following categories:

Which word forms of a lemma are found in the corpus?
Are there spelling or inflectional variations?
In which meanings and syntactical constructions is the lemma used?
What collocations are there? What compounds does the lemma enter?
Is there evidence for idiomatic and phraseological usage?
What is the relative and absolute frequency of the lemma?
Is there a characteristic distribution over different text

types?
Which samples can best be used to demonstrate the meanings of the lemma?

To be precise, the Text Bank will not answer all these questions. Certainly no data-base system will ever be a substitute for the intellectual work of a lexicographer. But it can provide him with the data he needs for writing lexical entries more economically than in the traditional way.

The Dictionary Bank

In compiling a new dictionary, the lexicographer will be able to consult existing dictionaries and compare their descriptions with the textual samples he has found in the corpus. Since these dictionaries are, as a rule, arranged alphabetically, he is at a loss if he wants to have, say, a listing of all lemmata defined as belonging to the language of data-processing or a listing of all verbs permitting an accusative-and-infinitive construction, even if this information is provided in the lexical entries. However, if the dictionaries are available in machine-readable form it is no problem to retrieve information of this kind. As Nagao et al. (1982:52) said, "Dictionaries themselves are rich sources, as linguistic corpora. When dictionary data is stored in a data-base system, the data can be examined by making cross-references from various viewpoints. This leads to new discoveries of linguistic facts which are almost impossible to achieve in the conventional printed versions".

It is therefore desirable to collect as many dictionaries as possible in the Dictionary Bank. At the moment, there are at least five general-purpose German dictionaries on data carriers, among them the one-volume DTV-WÖRTERBUCH, the six-volume DUDEN-GROSSES WÖRTERBUCH, and the still incomplete six-volume BROCKHAUS-WAHRIG. Due to copyright problems, only one or two smaller general-purpose dictionaries will be available for the Dictionary Bank in the foreseeable future. But there are a number of other machine-readable dictionaries compiled for applications in the field of artificial intelligence and automatic translation. A research group at the Bonn University Institut für Kommunikationsforschung und Phonetik, headed by Winfried Lenders, is currently carrying out a project with the aim of connecting and, to some extent, integrating eleven of these dictionaries in a data-base called 'Kumuliertes Lexikon' (cf. Heß et al. 1983). This project is due to be completed in 1984. The complete system will then be implemented in Mannheim as the Dictionary Bank of the Lexicographical Data-Base.

In its final version, the Dictionary Bank will provide a fully integrated version of the eleven dictionaries to the level of the lexical entry. A complete integration within the microstructure of the lexical entry, however, is neither possible nor desirable. A complex categorization of the parts of speech and the inflectional classes had to be developed to make up for the differences in the source dictionaries, since no information was to be lost. Unification was not possible on the level of semantic and pragmatic description. Here, the dictionary source for each information item has to be retrievable to assist the lexicographer in the evaluation.

The Result Bank

Once the formal micro-structure of the lexical entry in a projected dictionary is designed, the macro-structure of the dictionary (consisting of files, where each file represents a lexical entry) and the micro-structure (consisting of records, where each record represents an information category) can be constructed within the framework of any standard data-base management system. The format of the lexical entry is then at the disposal of all members of the dictionary staff. Each lexicographer can now store preliminary versions of whole lexical entries or selected information categories; each stored record can be altered, expanded or corrected at any time, or can be used for consultation and comparison in order to achieve homogeneity of description (cf. Guckler 1983).

Descriptive uniformity in the morphosyntactical categories seems easy enough. But the exact analysis of the first volumes of the BROCKHAUS-WAHRIG dictionary (cf. Wiegand and Kučera 1982) has shown an amazing number of discrepancies, inaccuracies and even gross mistakes, even though this dictionary claims to have used the full potential of modern data-processing. More difficult is homogeneity in the semantic description of the vocabulary, representing a partly hierarchical, partly associative net of conceptual relations. The words used in semantic explications must be used only in the same sense or senses in which they are defined under their respective headwords.

Using a Result Bank also helps the lexicographer avoid the discrepancies commonly found in the cross-references of synonyms, antonyms, hyponyms and so on. Furthermore, it helps him collect and compare the related elements of groups such as

all verbs with the same sentence pattern;
all adjectives used predicatively only;
all nouns with a particular inflectional pattern;
the vocabulary of automobile engineering;
all words rated as obsolete, etc.

When the final version is stored in the Result Bank, it can be used as copy, using standard editing programs to produce the printed dictionary directly from the result bank. This final version will then be integrated in the Dictionary Bank, provided the files and records of the Dictionary Bank and the Result Bank are properly matched. Thus the Lexicographical Data-Base will reflect the lexicographical knowledge of all the dictionary projects which employed it as a tool. The Result Bank can further be used as a master dictionary (as defined by Wolfart 1979) from which derived printed versions for different purposes can be produced.

Work on the Result Bank has not yet begun. At the moment, several data-base management systems are being tested as to their applicability, and information is being gathered on dictionary and terminology projects that employ or plan to employ similar techniques.

'Manual of Hard Words'

The Lexicographical Data-Base will first be applied to the current main dictionary project of the Institut für deutsche Sprache, the 'Manual of Hard Words', which at present is still in its planning stage. By 'hard words' we mean those lexemes in languages for special purposes which also occur in texts read by the lay public, such as instructional leaflets, income tax forms, legal literature, newspaper articles dealing with politics, economics, and the sciences. Special corpora for each of the subjects to be covered in this dictionary will be assembled from selected texts, and their findings will be compared with corresponding samples from general present-day language corpora (cf. Mentrup 1983). Even in its initial version, however, the Lexicographical Data-Base will be accessible and applicable for other linguistic projects as well.

Notes

[1] A description of the Mannheim corpora and the available programs for text handling is found in: LDV-INFO 1. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung, Mannheim 1981. Cf. also Teubert (1982).

[2] For a list of all machine-readable corpora in modern German, cf. Dokumentation Textkorpora des neueren Deutsch. Institut für deutsche Sprache, Mannheim 1982.

References

Brückner, T. (1982) "Programm-Dokumentation REFER, Version 1" in LDV-INFO 2. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung. Mannheim: Institut für deutsche Sprache

Guckler, G. (1983) "A computer-based monolingual dictionary: a case study" in Lexicography: Principles and Practice ed. by R.R.K. Hartmann. London-New York: Academic Press

Heß, K. et al. (1983) Maschinenlesbare deutsche Wörterbücher. Tübingen: Niemeyer

Mentrup, W. (1983) "Lexikographische Konzepte zur Beschreibung 'schwerer Wörter'. Probleme und Vorschläge" in Wortschatz und Verständigungsprobleme ed. by W. Mentrup. Düsseldorf: Schwann

Nagao, M. et al. (1982) "An attempt to computerize dictionary data bases" in Lexicography in the Electronic Age ed. by J. Goetschalckx and L. Rolling. Amsterdam: North-Holland

Teubert, W. (1982) "Corpus and lexicography" Paper presented at the Second Scientific Meeting on Computer Processing of Linguistic Data at Bled, Yugoslavia

Wiegand, H.E. and Kučera, A. (1982) "Brockhaus-Wahrig. Deutsches Wörterbuch auf dem Prüfstand der praktischen Lexikologie. II. Teil" in Studien zur neuhochdeutschen Lexikographie (Band I) ed. by H.E. Wiegand. Hildesheim-New York: Olms

Wolfart, H.C. (1979) "Diversified access in lexicography" in Dictionaries and Their Users ed. by R.R.K. Hartmann. Exeter Linguistic Studies (Vol.4)