APPLICATIONS OF A LEXICOGRAPHICAL DATA BASE FOR GERMAN

Wolfgang Teubert

Institut für deutsche Sprache
Friedrich-Karl-Str. 12
6800 Mannheim 1, West Germany

ABSTRACT

The Institut für deutsche Sprache recently has begun setting up a LExicographical DAta Base for German (LEDA). This data base is designed to improve efficiency in the collection, analysis, ordering and description of language material by facilitating access to textual samples within corpora and to word articles within machine readable dictionaries and by providing a frame to store results of lexicographical research for further processing. LEDA thus consists of the three components *Text Bank*, *Dictionary Bank* and *Result Bank* and serves as a tool to suppport monolingual German dictionary projects at the Institute and elsewhere.

I INTRODUCTORY REMARKS

Since the foundation of the Institut für deutsche Sprache in 1964, its research has been based on empirical findings; samples of language produced in spoken or written from were the main basis. To handle efficiently large quantities of texts to be researched it was necessary to use a computer, to assemble machine readable corpora and to develop programs for corpus analysis. An outline of the computational activities of the Institute is given in LDV-Info (1981 ff); the basic corpora are described in Teubert (1982). 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 disc decks and at the moment 15 visual display units for interactive use.

Whereas in former years most jobs were carried out in batch, the terminals now make it possible for the linguist to work interactively with the computer. It was therefore a logical step to devise Lexicographical Data Base for German (LEDA) as a tool for the compilation of new dictionaries. The ideology of interactive use demands a different concept of programming where the lexicographer himself can choose from the menu of alternatives offered by the system and fix his own search parameters. Work on the Lexicographical Data Base was begun in 1981; a first version incorporating all three components is planned to be ready for use in 1986.

What is the goal of LEDA? In any lexicographical project, once the concept for the new dictionary has been established, there are three major tasks where the computer can be employed:

(i) For each lemma, textual samples have to be determined in the corpus which is the linguistic base of the dictionary. The text corpus and the programs to be applied to it will form one component of LEDA, namely the Text Bank.

(ii) For each lemma, the lexicographer will want to compare corpus samples with the respective word articles of existing relevant dictionaries. For easy access, these dictionaries should be transformed into a machine readable corpus of integrated word articles. Word corpus and the pertaining retrieval programs will form the second component, i.e. the Dictionary Bank.

(iii) Once the formal structure of the word articles in the new dictionary has been established, description of the lemmata within to the framework of this structure can be begun. A data base system will provide this frame so that homogenous and interrelated descriptions can be carried out by each member of the dictionary team at all stages of the compilation. This component of LEDA we call the Result Bank.

II TEXT BANK

Each dictionary project should make use of a text corpus assembled to the specific requirements of the particular lexicographical goal. As self-evident as this claim seems to be, it is nonetheless true for most German monolingual dictionaries on the market that they have been compiled without any corpus; this is apparently even the case for the new six volume BROCKHAUS-WAHRIG, as has been pointed out by Wiegand/Kucera (1981 and 1982). For a general dictionary of

contemporary German containing about 200 000 lemmata, the Homburger Thesen (1978) asked for a corpus of not less than 50 million words (tokens).

To be used in the text bank, corpora will have to conform to the special codification or pre-editing requirements demanded by the interactive query system. At present, a number of machine readable corpora in unified codification are available at the Institute, 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. Further corpora habe been taken over from other research institutions, publishing houses and other sources. These texts had been coded in all kinds of different conventions, and programs had to (and still have to) be develqped 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 use of an optical scanner, if they are not available on machine readable data carriers. By the end of 1985 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. For a detailed description of it, see Brückner (1982) and (1984). The purpose of this system is to ensure quick access to the data of the text bank, thus enabling the lexicographer to use the corpus interactively via the terminal. Unlike other query programs, REFER does not search a word form (or a combinantion of graphemes) in the corpus itself, but in registers containing all the word forms. One register is arranged in the usual alphabetical way, the other is organized in reverse or a tergo to allow a search for suffixes or the terminal elements of compounds. All word forms in the registers are connected with the references to their actual occurrence in the corpus, which are then looked up directly. With REFER, it normally takes no more than three to five seconds for the search procedure to be completed, and all occurrences of the word form within an arbitrarily chosen context can be viewed on the screen. Response behaviour does not depend on the size of the text bank.

In addition, REFER features the following options:

- The lexicographer can search for a word form, for word forms beginning or ending with a specified string of graphemes or for word forms containing a specified string of graphemes at any place.
- The lexicographer can search for any combination of word forms and/or graphemic strings to occur within a single sentence of the corpus.
- REFER is connected with a morphological generator supplying all inflected forms for the basic form, e.g. the infinitive (cf. fahren (inf.) --- fahre, fährst, fahrt, fährt, fuhr, fuhren, fuhrst, führe, führen, führst, gefahren). This will make it much easier for the lexicographer to state his query.
- For all word forms, REFER will provide information on the relative and absolute frequency and the distribution over the texts of the corpus.
- The lexicographer hat a choice of options for the output. He can view the search item in the context of a full sentence, in the context of any number of sentences or in the form of a KWIC-Index, both on the screen and in print.
- For each search procedure, the linguist can define his own subcorpus from the complete corpus.
- Lemmatized registers are in preparation. They will be produced automatically using a complete dictionary of word forms with their morphological descriptions. These lemmatized registers not only reduce the search time, but also give the accurate frequency of a lemma, not just a word form, in the corpus.
- Register of word classes and morphological descriptions (e.g. listing references of all past participles) will be produced automatically by inverting the lemmatized registers. Thus the linguist can search for relevant grammatical constructions, like all verb complexes in the passive voice.
- Another feature will permit searching for an element at a predetermined sentence position, like all finite verbs as the first words of a sentence or all nouns preceded by two adjectives.

Thus the text bank is a tool for the lexicographer to gain information of the following kind:

- 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 employed?
- What collocations are there? What compounds is the lemma part of?
- Is there evidence for idiomatic and phraseological usuage?
- 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?

Preliminary versions of the text bank are in use since 1982. Not only lexicographers but also grammarians employ this interactive system to gain the textual samples they need. A steadily growing number of service demands both from members of the Institute and from linguists at other institutions are being fulfilled by the text bank.

III DICTIONARY BANK

If access to the textual samples of a corpus is an indisputable prerequisite for successful dictionary compilation, consultation of other relevant dictionaries can facilitate the drawing up of lexical entries. It is virtually impossible to assemble a corpus so extensive and encompassing that it will suffice to describe the whole vocabulary of a language, even within the limits of the particular conception of any dictionary (unless it were a pure corpus dictionary). A dictionary of contemporary language should not let down its user if he is reading a text written in the early 19th century though it will contain words and meanings of words not found in a corpus of post World War II texts. This holds even more for languages for special purposes; they cannot be described without recurrence to technical dictionaries, collections of terminology and thesauri, because the more or less standardized meanings cannot be retrieved from their occurrences in texts.

According to Nagao et al. (1982), "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 of various viewpoints. This leads to new discoveries of linguistic facts which are almost impossible to achieve in the conventional printed versions". A dictionary bank will therefore form one of the components of the Lexicographical Data Base.

Since 1979 a team at the Bonn Institut für Kommunikationsforschung und Phonetik is compiling a 'cumulative word data base for German', using 11 existing machine readable dictionaries of various kinds, including dictionaries assembled for Artificial Intelligence projects, machine translation systems and, for copyright reasons, only two generals purpose dictionaries. Programs have been developed to make up for the differences in the description of lemmata and to permit automatic cumulation. For further information regarding this project, see Hess/Brustkern/Lenders (1983) and Brustkern/Schulze (1983, 1983a). The cumulative word data base, which is due to be completed in 1984, will then be implemented in Mannheim and form the core of the dictionary bank of LEDA.

In its final version, the dictionary bank will provide a fully integrated cumulation of the source dictionaries, down to the level of lexical entries, including statement of word class and morphosyntactical information. A complete integration within the microstructure of the lexical entry, however, seems neither possible nor even desirable. Automatic unification cannot be achieved on the level of semantic and pragmatic description. Here, the source for each information item has to be retrievable to assist the lexicographer in the evulation.

The dictionary bank will be a valuable tool not only for the lexicographer but also for the grammarian. Retrieval programs will make it possible to come up with a listing of all verbs with a dative and accusative complement, or of all nouns belonging to a particular inflectional class. Since the construction of the dictionary bank and the result bank will be related to each other, every time a new dictionary has been compiled in the result bank, it can be copied into the dictionary bank, making it a growing source of lexical knowledge. The dictionary bank can then be used as a master dictionary as defined by Wolfart (1979), from which derived printed versions for different purposes can be produced.

IV RESULT BANK

Whereas text bank and dictionary bank supply the lexicographer with linguistic information, the result bank will be empty at the beginning of a project; it consists of a set of forms which are the frames for the word articles. Into these forms the lexicographer enters the (often preliminary) results of his work, which will be altered, amended or shortened and interrelated with other word articles (e.g. via synonymy or antonymy) in the course of compilation; he copies into those forms relevant textual samples from the text bank and useful information units from the dictionary bank.

Access via terminal is not only possible to any file representing a word article but also to any record representing a category of explication. The result bank, which can be constructed within the framework of any standard data base management system, thus permits consultation and comparison on any level of lexical description. Descriptive uniformity in the morphosyntactical categories seems easy enough. But as has been shown in a number of studies, e.g. by Mugdan (1984), most existing dictionaries

abound in discrepancies and inaccuracies which easily can be avoided by cross-checking within the result bank. 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 heard words. These tasks can be carried out easier within a data base system. Furthermore, the result bank will support collecting and comparing the related elements of groups such us:

- all verbs with the same sentence patterns
- all adjectives used predicatively only
- all nouns denoting tools
- all words rated as obsolete
- the vocabulary of automobile engineering.

Files will differ from word class to word class, as particles or adverbs cannot be describend within the same cluster of categories as nouns or verbs. Similarily, macrostructure and microstructure will not be the same for any two dictionaries. Still categories should be defined in such a way that the final version of the dictionary can be copied into the dictionary bank without additional manual work.

After the dictionary has been compiled, it can be used as copy, using standard editing programs to produce the printed version directly from the result bank. At that level, strict formatting is no longer necessary and should be abandoned, whereever possible, in favour to economy of space.

Work on the result bank will begin in autumn 1984. The pilot version of it will be applied to the current main dictionary project of the Institute, i. e. the "Manual of Hard Words", which at present is still in its planning stage. Even in its initial version, however, LEDA will be accessible and applicable for other lexicographical projects as well.

REFERENCES

Tobias Brückner. Programm Dokumentation Refer Version 1. LDV-Info 2. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung. Mannheim: Institut für deutsche Sprache, 1982, pp. 1-26.

Tobias Brückner. Der interaktive Zugriff auf die Textdatei der Lexikographischen Datenbank (LEDA) Sprache und Datenverarbeitung 1-2/1982, 1984, pp. 28-33.

Jan Brustkern/Wolfgang Schulze. Towards a Cumulated Word Data Base for the German Language. IKP-Arbeitsbberichte Abteilung LDV. Bonn: Institut für Kommunikationsforschung und Phonetik der Universität Bonn, 1983, pp. 1-9.

Jan Brustkern/Wolfgang Schulze. The Structure of the Word Data Base for the German Language. IKP-Arbeitsberichte Abteilung LDV, Nr. 1. Bonn: Institut für Kommunikationsforschung und Phonetik der Universität Bonn, 1983, pp 1-9.

Klaus Heß/Jan Brustkern/Winfried Lenders. Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration. Tübingen, 1983.

LDV-Info. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung, Mannheim: Institut für deutsche Sprache, 1981 ff.

Joachim Mugdan. Grammatik im Wörterbuch: Wortbildung. Germanistische Linguistik 1-3/83, 1984, pp. 237-309.

M. Nagao, J. Tsujii, Y. Ueda, M. Takiyama. An Attempt to Computerize Dictionary Data Bases. J. Gotschalckx, L. Rolling (eds.). Lexicography in the Electronic Age. Amsterdam, 1982, pp. 51-73.

Wolfgang Teubert. Corpus and Lexicography. Proceedings of the Second Scientific Meeting "Computer Processing of Linguistic Data". Bled, Yugoslavia, 1982, pp. 275-301.

Herbert Ernst Wiegand / Antonin Kucera. Brockhaus-Wahrig. Deutsches Wörterbuch auf dem Prüfstand der praktischen Lexikologie. I. Teil: 1. Band (A-BT); 2. Band (BU-FZ). Kopenhagener Beiträge zur Germanistischen Linguistik, 18, 1981, pp.. 94-217.

Herbert Ernst Wiegand / Antonin Kucera. Brockhaus-Wahrig. Deutsches Wörterbuch auf dem Prüfstand der praktischen Lexikologie. II. Teil: 1. Band (A-BT); 2. Band (BU-FZ); 3. Band (G-JZ). Germanistische Linguistik 3-6/80, 1982, pp. 285-373.

H. C. Wolfart. Diversified Access in Lexicography. R.R.K.Hartmann (ed.). Dictionaries and Their Users. Papers from the 1978 B.A.A.L. Seminar on Lexicography. (=Exeter Linguistic Studies, Vol.4). Exeter, 1979, pp. 143-153.