

THE USE OF COMPUTER TECHNOLOGIES IN THE LEXICOGRAPHY

Since the early eighties, computer technology has become of increasing importance for lexicography. The compilation of dictionaries is being more and more computerized (cf. Clear 1987 vs. Glassman et al. 1992). Electronic dictionaries have obvious advantages over printed dictionaries with respect to access to the dictionary information and reusability of the product (e.g. Hartevelde 1991). For this reason, comprehensive reference works are converted from printed to electronic form (e.g. the Oxford English Dictionary OED; Simpson 1986). A variety of topics concerning machine-readable dictionaries is covered by a new specialism: computational lexicography (cf. Magay and Zigany 1988; Boguraev and Briscoe 1989).

In the past decade, machine-readable dictionaries and electronic text corpora have become relevant to specialisms in the fields of computational linguistics, information technology, and knowledge engineering. These specialisms have a common key problem: how to provide computer systems with linguistic knowledge and with world or specific-domain knowledge, in order to improve them. This knowledge is needed by computer systems that process (i.e. 'understand' or 'produce') natural (human) language for some purpose, such as machine-translation, automatic text summarizing, man-machine communication in natural language (dialogue systems), as well as selective retrieval of relevant documents from a large text database. Machine-readable dictionaries and electronic text corpora are resources from which, to some extent, knowledge can be extracted for building a computational lexicon, which is considered a major bottleneck for natural language processing NLP (Zemik 1991), or for building a lexical knowledge base, which not only contains lexical information but also has a conceptually based organization and an inference mechanism (Boguraev and Levin 1990). Very large electronic text corpora are additionally used for empirical and statistical methods of automatic language analysis (Church and Mercer 1993). They contain sentence and word usage information that was difficult to collect until recently and consequently was largely ignored by linguists.

The quality of future computerized corpus-based lexicography will rely on progress not only in the more or less traditionally related fields of linguistics and computer technology, but also in the fields of language technology, information technology and knowledge engineering. The efficiency of dictionary compilation and the quality of future dictionaries may be improved by advanced means supporting the analysis and interpretation of corpus data, as well as by flexible access to a variety of electronic resources of information. Here, we left aside the potentially favourable effects of the attempts to bridge the gap between dictionary compilers and theoretical lexicographers on the one hand, and between the makers of dictionaries for human use and those making computational lexica on the other (cf. Swanepoel 1994). If computers would ever succeed in interpreting text at a semantic level in some way, then the corpus system could even provide the lexicographer with preliminary interpretations of concordances. However, research in automatic machine-translation, for example, has demonstrated that we have no reason for being optimistic about the term in which this might be feasible. Even with respect to the present results and the tools available, the question always is what is actually ready for implementation into rather complex corpus query systems and lexicographer's workbenches.