Technologies in Computerized Lexicography

J.G. Kruyt, *Instituut voor Nederlandse Lexicologie INL, Leiden, The Netherlands*

Abstract: Since the early eighties, computer technology has become increasingly relevant to lexicography. Computer science will probably not be the only technological discipline which may have implications for future computerized lexicography. Some developments in the fields of language technology, information technology and knowledge engineering, may support lexicographical practice and enhance the quality of the resulting dictionary. The present paper discusses how the analysis and interpretation of electronic corpus data by the lexicographer may be improved by automatic linguistic analysis, by better access to the corpus, and by a more flexible communication with the computer system. As a frame of reference, first an indication of the state of the art in computerized lexicography will be given, by a concise discussion of three projects at the Institute for Dutch Lexicology INL considered in an international context: the conversion of the *Woordenboek der Nederlandsche Taal WNT* (Dictionary of the Dutch Language Based on Historical Principles) to electronic form, the compilation of the *Vroegmiddelnederlands Woordenboek* (Dictionary of Early Middle Dutch) in a computerized lexicographer's workbench, and the *INL Taalbank* (INL Language Database). Although the topic of this paper is technology, focus is on functional rather than technical aspects of computerized lexicography.

Keywords: COMPUTERIZED LEXICOGRAPHY, ELECTRONIC DICTIONARY, ELECTRONIC TEXT CORPUS, LEXICOGRAPHER'S WORKBENCH, INTEGRATED LANGUAGE DATABASE, AUTOMATIC LINGUISTIC ANALYSIS, INFORMATION RETRIEVAL, USER INTERFACE

Samenvatting: Sinds het begin van de tachtiger jaren, is de computertechnologie in toenemende mate relevant geworden voor de lexicografie. Maar de computertechnologie zal waarschijnlijk niet de enige technische discipline zijn die implicaties heeft voor de toekomstige, computerondersteunde lexicografie. Ontwikkelingen in de taaltechnologie, informatietechnologie en kennistechnologie zijn van belang voor de ondersteuning van de lexicografische praktijk en daarmee de verhoging van de kwaliteit van het woordenboek. In dit artikel wordt besproken hoe de analyse en interpretatie van electronische corpusgegevens door de lexicograaf kan worden verbeterd door automatische linguistische analyse, door betere toegang tot het electronische tekstcorpus en door een flexibeler communicatie met het computersysteem. Als referentiekader wordt eerst een indruk gegeven van de huidige stand van zaken met betrekking tot gecomputeriseerde lexicografie, door een beknopte bespreking van drie projecten van het Instituut voor Nederlandse Lexicologie, geplaatst in een internationale context: de omzetting van het Woordenboek der Nederlandsche Taal WNT in elektronische vorm, de vervaardiging van het Vroegmiddelnederlands Woordenboek in een geautomatiseerde lexicografische werkomgeving en de INL Taalbank. Hoewel het onderwerp van dit artikel technologie betreft, valt de nadruk niet op de technische, maar op de functionele aspecten van de gecomputeriseerde lexicografie.

Trefwoorden: COMPUTERONDERSTEUNDE LEXICOGRAFIE, ELECTRONISCH WOORDENBOEK, ELECTRONISCH TEKSTCORPUS, LEXICOGRAFISCH WERKSTATION, GEINTEGREERDE TAALBANK, AUTOMATISCHE LINGUISTISCHE ANALYSE, INFORMATION RETRIEVAL (GEEN NEDERLANDS EQUIVALENT), GEBRUIKERSINTERFACE

1. Introduction

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. Harteveld 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 Zigány 1988; Boguraev and Briscoe 1989). The advances over the past years have been relevant to three projects at the Institute for Dutch Lexicology INL. The *Woordenboek der Nederlandsche Taal WNT*, the Dutch counterpart of the *Oxford English Dictionary OED*, is being converted to electronic form and will be available on CD-ROM, probably in autumn 1995. For the compilation of the *Vroegmiddelnederlands Woordenboek VMNW* (Dictionary of Early Middle Dutch), an automatized lexicographer's workbench has been developed, which ensures immediate storage of compiled entries into a dictionary database. The *INL Taalbank* (INL Language Database), originally a closed electronic text corpus intended for lexicographical purposes only, is being developed towards a dynamic multifunctional language database.

Converting printed reference works into electronic products is, of course, a passing activity, as new reference works will directly be produced in electronic form. For reasons of quality (cf. Sinclair 1987), new dictionaries will be based on the analysis of large electronic text corpora rather than on introspective methods only. This also applies to commercial dictionaries, as is evident from the corpus-based *Collins Cobuild English Language Dictionary* (1987) and *Longman Language Activator* (1993). Computerized compilation of dictionaries, which improves at least consistency in the dictionary, will become more efficient by faster and more powerful computers. But computer science will probably not be the only technological discipline which may have implications for future computerized corpus-based lexicography. Some promising developments in language technology (in a broad sense, including computational linguistics and corpus linguistics), information technology and knowledge engineering may support lexicographical practice and enhance the quality of the resulting dictionary. The present paper will discuss how the analysis and interpretation of corpus data by the lexicographer may be improved by automatic linguistic analysis, by better and more diversified access to electronic text corpora as well as to electronic reference works, and by a more flexible communication with the computer system (section 3). These developments will be related to future INL projects in section 4. The paper concludes with a discussion of more general implications for the lexicographer's knowledge and skills. As a frame of reference, first an indication of the state of the art in computerized lexicography will be given by a concise discussion of the above-mentioned INL projects in an international context (section 2). Although the topic of this paper is technology, focus will not be on technical issues but on the functional aspects of computerized lexicography relevant to the lexicographer.

2. Computerization at the Institute for Dutch Lexicology INL

2.1 Introduction

The INL has a long tradition in corpus-based lexicography. The *Woordenboek der Nederlandsche Taal WNT* is based on a large corpus of written quotations, just like its counterparts *OED, Deutsches Wörterbuch,* and other dictionaries originating in the nineteenth century. The long tradition implies that the traditional corpus-based activities are well-known and, in spite of the lexicographer's liberties, more or less standardized. Basically, this is a good condition for computerizing the lexicographical process. The compilation of the *Dictionary of Early Middle Dutch VMNW* in an automatized environment, is an example of computerized traditional lexicographic practice (2.3).

In line with the institute's policy of corpus-based lexicography, the INL started building a large electronic text corpus of present-day Dutch in the early eighties, in view of a dictionary of 20th and 21st century Dutch, planned after completion of the *WNT.* As a consequence of the growing global interest in large electronic text corpora in the past few years, this corpus will be a component of a multifunctional collection of electronic texts, rather than used for lexicographical purposes only (2.4).

The reason for converting the *WNT* into an *Electronic WNT* (2.2) is not only to have flexible and fast access to the wealth of information included in this dictionary. It will also be an easy accessible, valuable reference work during the compilation of the envisaged dictionary of 20th and 21st century Dutch. The *Electronic WNT,* covering the Dutch language from the 16th-20th century, will additionally be an important component of the future *INL Integrated Language Database of 12th-21st Century Dutch* (4).

In these projects, the computer is used in essentially three types of processes. In the *Dictionary of Early Middle Dutch* project, focus is on system development, carried out by computer scientists. The linguistic encoding of text corpora belongs to the field of language technology, more specifically natural language processing (3.1), and is carried out by computer linguists, whereas computer scientists are responsible for the implementation of the encoded corpora into storage and retrieval systems. The development of the *Electronic WNT* is first of all a matter of text technology. By text technology, we mean processing of text (rather than language) by the computer mainly based on characteristics of the (typo)graphical form and textual structure of a text. In the *Electronic WNT* project, text technology more specifically concerns the automatic assessment of the information categories (contents) of text fragments, on the basis of the graphical form and structure of the dictionary text as well as the lexicographical structure of the entries (form) (2.2). This job requires a special combination of linguistic and programming expertise. Most software for the INL projects was developed in-house.

2.2 *Electronic Woordenboek der Nederlandsche Taal WNT*

The *Electronic WNT* project started in 1984, a little later than its model project, the *New OED* project (Simpson 1986). Mainly due to limiting financial conditions, the project has not yet resulted in an *Electronic WNT*. Cooperation with the Dutch electronic publishing firm *AND* in the past two years, will result in the publication of a CD-ROM, probably in autumn 1995. The basis for it will be a *WNT*-text file encoded for information categories. That is, the running *WNT*-text will be interrupted by tags specifying the type of information conveyed by a text fragment. The encoding enables the user to have multi-path access to the information in the electronic dictionary. Not only the headword, traditionally the entry to dictionary information, but each encoded information category can be used in queries, either separately or in combination (cf. Kruyt 1989).

The conversion of the printed dictionary text into an encoded text file requires essentially three steps. First, the printed text is to be converted into its equivalent in machine-readable form. Ideally, the corrected machine-readable text file should be the input for the automatic encoding for information categories. In practice, this is not feasible, due to structural ambiguity in the dictionary entries and to 'lexicographical economy', i.e. all means applied by the lexicographer for reasons of economy of space, such as abbreviations, dashes replacing words, incomplete references to repeated sources, etc. A prior text revision supporting the automatic encoding, is required. This may be done during the conversion to machine-readable form (the first step), or as an intermediate second step (cf. Kruyt and Van der Voort van der Kleij 1992-93). Except for some minor points, this approach is similar to the one followed in the *New OED* project (cf. Berg et al. 1988).

Since 1982, text processing facilities have been utilized in the production of the printed *WNT*. This directly yields corrected machine-readable text files.

Ca. 69.000 columns of dictionary text were available in printed form only, and were to be converted to machine-readable form. Initially this was done by retyping the dictionary text, followed by manual correction, later on by optical character recognition (OCR) with computer-aided correction (Kruyt and Van der Voort van der Kleij 1992, 1993), and most recently by retyping the dictionary text twice, the correction being supported by the automatically detected differences between the two versions. The conversion and textual revision steps were carried out by different partners in the course of the project. As a consequence, the input texts for the automatic encoding differ with respect to the supporting text revisions. But this is not the only factor complicating the preparation of a uniformly encoded *WNT*-text file. The quality of the first *Electronic WNT on CD-ROM* will mainly be determined by characteristics of the dictionary itself (inconsistency caused by lexicographers' liberties, revised lexicographical methods, etc.; cf. Moerdijk 1994) and the feasibility to compensate for that.

The *Electronic WNT* project differs from the *NEW OED* project in some respects. Due to limiting financial conditions, our project has become a longstanding project, which subsequently caused the project to be less attractive for external funding by commercial firms. This, of course, also applies to similarly comprehensive, rather ambitious and financially risky projects. As a consequence, the project structure and the choices to be made could not always be the most desirable ones. This is the major reason for the different methods applied by the various institutions involved, referred to above. On the other hand, limiting conditions stimulated us to achieve the largest possible efficiency. The conversion to machine-readable form by OCR at the time, though dismissed in the *New OED* project (Berg et al. 1988) and still a controversial tool for computerizing large scale resources (Doorn et al. 1993), resulted in a considerable reduction of costs per column, to a considerable saving of time for the subsequent text revision process, and to a higher quality of the output, due to a high degree of computerization of the whole process at our institute (Kruyt and Van der Voort van der Kleij 1992-93).

The *WNT* is not yet completed. In view of the complete *Electronic WNT*, the question arises how to compile the remaining volumes. The traditional way implies a similar conversion and encoding process as described above. Going over to compilation in a database structure, as has been done in the *Woordeboek van die Afrikaanse Taal WAT* project (Harteveld 1991, 1994), implies a fundamental revision of the lexicographer's working method. The question was hardly relevant to our institute. The financial basis for progress and continuity of the *Electronic WNT* project was insufficiently firm, and the date of completion of the *WNT* by the traditional method was close enough for not to take unknown risks.

2.3 Lexicographer's workbench for the *Dictionary of Early Middle Dutch VMNW*

The compilation of the *Dictionary of Early Middle Dutch VMNW*, covering the Dutch language of the 13th century, started in 1988. The dictionary is based on an electronic corpus of Early Middle Dutch texts (mainly the *Corpus Gysseling*), containing ca. 1,6 million word forms. The corpus has interactively been encoded for part of speech, inflection, and present-day Dutch head word. Characteristics of the dictionary and the corpus are described in Pijnenburg (1991).

From 1989 up to 1993, the Electronic Data Processing (EDP) department at the INL developed a lexicographer's workbench for the *VMNW* project. This is an information system in which the electronic corpus, the lexicographer's working environment and the lexicographical database are mutually linked subsystems integrated into a relational database system. Compilation is computerized to a large extent. The four lexicographers have on-line access to the database system through workstations in a local area network. The system allows the lexicographers to select text materials from the corpus (basically but not exclusively the headword instances) and to copy it to their working environment. The lexicographer's analysis and interpretation of the text materials is supported by different views on the data, mainly tables and concordances (selected word forms in their local context), as well as by various sorting and rearranging options according to the parameters identified in the database system. For the word forms, these parameters are the above-mentioned linguistic categories and position in the document, and for the documents, date and place of origin, and text genre. The linguistic encoding allows for queries addressing both word form and part of speech level, separately and in combination. The system enables the analyzing and interpreting lexicographer to classify concordances according to lexicographical criteria (e.g. the headword's meaning) by marking them with a code. Subsequently, these codes become parameters for selection, sorting and rearrangement actions, in addition to the just-mentioned parameters already identified in the database system. The result of the lexicographer's investigations is recorded into skeleton template entries displayed on the screen, with separate fields for the various information categories in the dictionary. Consistency and efficiency are enhanced by some built-in system facilities. For example, specific fields, such as part of speech, are immediately checked by the system for their formal contents. Selected quotations are directly copied from the corpus subsystem to the lexicographical database. The printed version of the dictionary is derived from the lexicographical database. For a more detailed description from the lexicographer's point of view, we refer to Schoonheim (in press).

When compilation started in 1988, it was not obvious to what extent and how the computer could be utilized in the project. At the time, the *Cobuild* project demonstrated the constraints of computer-aided corpus-based lexicography (Clear 1987). Due to technical limitations, the *Cobuild* lexicographers worked off-line for reasons of convenience and economy. Concordances were analyzed on the basis of paper copies. The template entries were completed in written form, then converted to machine-readable form by keyboarding, and finally loaded into a relational database. At the beginning of 1987, Clear stated: "it is still the case that the technology of microcomputers and network communications is unable to offer an economically competitive system which will allow a large team of lexicographers to compile dictionary entries without using pen and paper" (Clear 1987: 47). The feasibility of a comprehensive lexicographer's workbench was also topic of a lively debate among lexicographers and technicians, during the European Science Foundation Summer School on "Computational Lexicology and Lexicography", in Pisa, in 1988. Obviously, the technicians had an optimistic view, whereas the lexicographers were very sceptical about it. Against this background, the decision to develop a comprehensive lexicographer's workbench for the compilation of the *VMNW* in 1988, may be called rather progressive.

A similar integrated system for corpus-based lexicography was built in the *Hector* project, a feasibility study on high-tech corpus lexicography performed by Oxford University Press and Digital Equipment Corporation from 1990/91 up to mid 1993 (Glassman et al. 1992; Atkins 1992-93). The main difference is corpus size: 17,3 million words in the *Hector* project versus 1,6 million words in the *VMNW* project. Minor differences concern the use of three versus one screen by the lexicographer, the distribution of tasks over the system components, and access to other reference works in the *Hector* system. Both projects show the present technical limits. Handling large amounts of data, complex sorting and rearrangement actions and other technically complex processes result in unbearable performance, frozen screens, locking problems, etc. But the overall impression is that the lexicographer's work was faster, more thorough, more consistent, and "infinitely more fun" (Atkins 1992-93:6).

2.4 INL Language Database

In 1980, the INL started building a large, electronic text corpus of present-day Dutch for lexicographical purposes. A corpus-based dictionary of present-day Dutch is envisaged after completion of the *WNT* (Van Sterkenburg 1983). In accordance with the lexicographic views at the time, the corpus should be a representative sample of the Dutch (written) language (Martin et al. 1985). As machine-readable text was hardly available, most texts were converted from printed to machine-readable form by OCR (Van der Voort van der Kleij 1986). The so-called '*50 Million Words Corpus*' now consists of ca. 1600 full texts, with a total amount of ca. 50 million word forms (tokens), corresponding with ca. 700.000 different word forms (types). With a few exceptions, the texts date from 1970-1990. The corpus covers several genres within the category fiction (ca.

30%) and a broad variety of topics, representing the main domains in society and science, within the category non-fiction (ca. 70%). An on-line retrieval program enables the user to define subcorpora on the basis of the parameters author, title or character string in title, and text number, both prior to and after the formulation of a query. As a consequence, queries may concern the whole corpus or a user-defined subcorpus. Queries are still at the level of word form. However, the corpus is being linguistically encoded and retrieval on headword and part of speech is currently developed for part of it. Output data include tables with type/token frequencies and distribution of word forms over the sources, concordances with the keyword in a user-defined context, and the keyword in an electronic version of the traditional quotation slip. The analysis of the output concordances is supported by several sorting options.

Recent developments at a European level have resulted in a revised view on the function of the *50 Million Words Corpus* and other electronic text collections at the INL. The *50 Million Words Corpus* is one of the large electronic corpora of a national language, started in the early eighties for lexicographical purposes (cf. Zampolli and Cappelli 1983). The recent international interest in very large electronic text corpora (cf. 3), made the national language corpora attractive for broader application than for lexicographical purposes only. The European Commission, aiming at a European infrastructure for language technology, supported a preparatory study into the feasibility of a network of harmonized text corpora of the national languages, which could meet the needs of diverse (including commercial) user groups (*NERC*-project). This study has a follow-up in the *PAROLE*-project, in which thirteen academic and industrial participating partners, representing eleven language areas, aim at the specification and implementation of the envisaged corpora network. Participation of the INL in the European projects has intensified the awareness of the need for an approach that is more oriented towards the external user, rather than towards the institutional lexicographers only. This is relevant to the further development of the INL corpora (Kruyt 1995; Van Sterkenburg and Kruyt in press). In addition to the closed and static *50 Million Words Corpus*, an open-ended and dynamic collection of corpora is aimed at, which can be used for a wide range of research and applications. Focus will be on external access to specific corpora selected by the user. For a flexible selection of user-defined subcorpora, the texts at the INL need to be classified according to as many as possible meaningful parameters. The retrieval of linguistic data requires the texts to be linguistically encoded. Access to the corpora and the linguistic data should be facilitated by user interfaces that are as user-friendly as possible, even for non-experienced users.

A major result of these developments is the facility of on-line access by Internet to the on-line retrieval program developed for a new, linguistically encoded INL corpus, the *5 Million Words Corpus 1994*[1]. A total of seventeen text sources, most of them dating from 1989-1994, have been classified according to publication medium (book, newspaper, magazine, written-to-be-spoken) and to topic (politics, journalism, leisure, linguistics, environment, business and employment). These classifications, as well as bibliographic references to the texts, have been implemented in the retrieval program as parameters for the definition and selection of subcorpora. The texts have been automatically encoded for headword and part of speech by linguistic software developed at the INL, and have subsequently been loaded into the on-line retrieval program developed for this corpus (Van der Voort van der Kleij et al. 1994). The user can search for single words or word patterns, including a set of predefined syntactic patterns that can be customized by the user. Queries concern the levels of word form, headword and part of speech, separately or in combination by use of Boolean operators and proximity searches. Output data include intermediate tables with the possibility of selecting specific items, and ultimately concordances of the searched items in a user-defined context size. Under limitations due to copyright, the output data can be transferred to the user's computer by e-mail. Other facilities include a variety of sorting options. For 1995, two more corpora accessible in a similar way are planned: a newspaper corpus (27 million words) and a diversified corpus, the latter with extended linguistic encoding.

The European user-oriented multifunctional approach determines corpus development at the INL in the short term. The user group will, of course, include the INL lexicographers. But unlike the corpus in the *VMNW* project, the function and development of the present-day INL corpora is no longer exclusively determined by an internal dictionary project. In the longer term, the INL aims at the integration of its linguistic resources into an *INL Integrated Language Database of 12th-21st Century Dutch*. Which characteristics this language database may have and how its relationship with the envisaged dictionary project may be, is topic of section 4. First, we will outline some recent interdisciplinary developments that may have significance for lexicography, including the INL projects.

3. Recent interdisciplinary developments

3.1 Introduction

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 (Zernik 1991), or for building a lexical knowledge base, which not only contains lexical information but also has a conceptually based organisation 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.

A discussion of the various approaches in the different fields is outside the scope of this paper (for a historical review of NLP, see Sparck Jones 1995). Here, it is relevant that the automatic analysis of language has become an interdisciplinary topic of interest, and that some developments may have relevance to corpus analysis and computerized corpus-based lexicography. We particularly refer to the need for sophisticated means for access to and analysis of the huge amounts of corpus data. We will give some examples of promising developments.

3.2 Linguistic analysis

At the level of word form, the lexicographer's analysis of corpus data may be supported by statistical tools. Church et al. (1991) show how mutual information statistics (the probability of observing two words together compared with the probability of observing them independently) and the t-test can be used as measures of similarity and dissimilarity, respectively, of the words 'strong' and 'powerful'. They argue that the use of statistics of this kind may support the lexicographer to sharpen the focus of definitions, highlighting salient facts and omitting remote possibilities, and to formulate explicit rules for choosing among near synonyms. These tools have been implemented in the *Hector* project (Atkins 1992-93). Other examples of tools based on statistics are 'Collocate' and 'Typical', being developed by Sinclair (1994, 1995). 'Collocate' evaluates the significance of the individual collocate in concordances. The output is a list of word forms with significance scores for their co-occurrence with a particular keyword. The tool demonstrated that *eye* is mostly associated with the metaphorical uses, with collocates such as *caught, naked, evil,* whereas *eyes* was used more in the physical sense with collocates such as *brown, narrow, blue.* 'Typical' is intended to find 'typical' citations for a certain word form and sorts concordance lines in order of the combined significance value of the collocates in a line. The output shows a grouping of concordance lines which contain the same collocate, provided with the significance value, and ordered from high to low significance. 'Typical' is developed to provide reliable and useful examples of words in use, and it also appeared to be helpful for disambiguating different senses of words. For a different approach for automatic selection of the most representative concordances, we refer to Collier (1994).

The analysis at other language levels than word form requires a corpus encoded for linguistic features. Up to now, the encoding has often been done manually or interactively (cf. 2.3, *VMNW*). With the present-day multi-million words corpora, this is no longer feasible. In the framework of improving NLP (3.1), much effort is spent in developing software for automatic morphological analysis, part of speech tagging, lemmatizing and syntactic parsing, in particular for the English language (see issues of e.g. *Computational Linguistics, Literary and Linguistic Computing, Computer and the Humanities*). For several languages, automatic part-of-speech taggers and morphological analyzers have been developed. Lemmatizers are not yet available at a large scale. Automatic syntactic parsing of unrestricted text is feasible at the level of phrasal groupings; the quality at the sentence level is still rather poor, even for English. Automatic semantic tagging and knowledge-based approaches are less developed (cf. Pustejovsky et al. 1993). From the point of view of the lexicographer, these efforts are relevant to the automatic linguistic encoding of corpora. A corpus encoded at whatever linguistic level, allows retrieval on the encoded linguistic parameters (cf. 2.3, 2.4). Statistic devices can be applied on encoded linguistic features as well. For example, Church et al. (1991) compute lexical preferences among subjects, verbs and objects, and they suggest that tables of SVO associations could be used for partitioning concordance lines into senses.

The kind of tools discussed here support the analysis of corpus data by the lexicographer, firstly by allowing queries at various linguistic levels, secondly by computing lexical patterns that are not easily observable by human analysis, and finally by the facility of concordance sorting according to relevance. If implemented in a lexicographer's workbench, the tools provide the lexicographer with the facility to have different views on large amounts of corpus data with a speed and flexibility that is inconceivable within the traditional method of manually arranging quotation slips. Another application of the tools may be the classification of corpus texts on the basis of internal linguistic characteristics (cf. Biber 1993), rather than on the more common external parameters (topic, bibliographic data etc.; cf. 3.3).

3.3 Access to data: information retrieval

With the increasing availability of electronic reference works and other textual information that may have relevance to lexicography, the lexicographer (as many other people) is dependent on research in information retrieval (IR). The aim of IR is to provide the user with exactly the information he needs from the huge amounts of electronic data nowadays created. The effectiveness of a document IR system is determined by its recall (the fraction from all relevant documents available that has actually been found) and its precision (the fraction from all found documents that actually is relevant). Most IR systems require the documents in a textual database to be 'indexed', i.e. provided with an abstract representation that reflects the contents of the document as good as possible. Different techniques have been developed (for an evaluation, see Wiesman 1995). In most current systems, the document representation consists of a number of words from the text which are considered to be representative for that text. The majority of IR systems only extract single word forms. This gives some serious problems affecting the effectiveness: (1) the user is forced to formulate a query using words that are literally present in the texts, (2) the user is not able to formulate a query starting off with a vague notion of what he is looking for, and (3) the system does not take into account that words may have several meanings and that many terms may have more or less the same meaning, being a characteristic of the natural language used both in the text in the documents and by the user (Karssen et al. 1994; cf. Wiesman 1995). The relationship with the problems in NLP is evident. In addition to statistical techniques, resources like machine-readable dictionaries and phrase lists, and relatively simple NLP-techniques like tagging and partial parsing, have proven to be useful contributors to improving IR effectiveness, while more sophisticated techniques are still in an experimental stage (Smeaton 1994). Future IR is expected to be concept-based.

The development of new, multilingual concept-based IR techniques is the aim of a European research project described by Karssen et al. (1994). The text in the documents and the query stated in natural language by the user is translated into concepts denoting the meaning of the natural-language utterances (at the level of phrases in the sentence). Indexing consists of determining the concepts that denote the meaning of the texts, while retrieval comes down to translating the query into representative concepts and matching these concepts with the ones representing the documents. This should be realized by the following method. After preprocessing, the texts are annotated with tags denoting the lemma and morphological category of each word, by use of a lexicon. Then, a syntactic analyzer eliminates possible morphological ambiguities of words by deducing the syntactic role they play. The parse-structure of each utterance, denoting the morphological and syntactic categories of all its words, is input for a module that extracts meaningful chunks of text, phrases representing important notions of the text. After these fairly standard techniques have been applied, the chunks are input for a semantic analyzer which generates their representing concepts by calculating the right word sense (i.e. disambiguating the meaning of each of the individual words). The semantic analyzer makes use of a conceptual structure already available as a product, consisting of 25.000 concepts organizing some 100.000 word meanings. It should be noted that this approach is not yet implemented into an operational system. The project, funded by the European Commission and carried out by universities and commercial companies, gives an impression of what is going on in the field of IR.

Progress in IR is relevant to lexicography not only for reasons of easy and accurate access to electronic reference works available at libraries or, for example, the World Wide Web. A major concern in IR is to determine as good as possible what a text is about (the indexing stage). This is exactly what publishing houses and corpus builders do when they classify texts according to topic. The better the IR methods, the higher the quality of text classification according to topic, which is important for the user-defined selection of subcorpora (2.4).

3.4 Access to data: user interface

Another aspect of IR relevant to lexicographers, is the user interface, roughly speaking the way in which a computer system communicates with the user. One aspect of the lexicographer's scepticism with respect to a computerized lexicographer's workbench concerns the rather poor possibilities to keep a good overview of the data. Not only lexicographers, but also many other users these days are non-experts in the field of computers and are growing accustomed to user-friendly systems. This has led to research into methods for supporting the user in his queries. Wiesman (1995) evaluates some systems that have an intelligent search-intermediary between the user and the proper retrieval system, helping the user formulate and reformulate his queries. We mention here two techniques applied in search-intermediaries: a natural language interface, which allows the user to formulate a query in his own language rather than in a formal language, and a thesaurus or knowledge base with domain knowledge, by which the query can be expanded or restricted by replacing the concept by a more general or more specialistic concept, respectively. A concept-based system additionally allows the user to fuzzily describe what he is looking for and then comes up with suggestions corresponding to the user's notion (Karssen et al. 1994). NLP and knowledge banks are apparently relevant to this specialism as well.

De Smedt et al. (1994), in a tentative project proposal in the field of medical information science, focus on a user-oriented presentation of information. Their aim is twofold. The contents of the information to be presented by the system should automatically be customized for the individual user. Secondly, depending on the type of information to be presented, the system should present the information as text or as picture, text and pictures being coherently combined in the output of the system. The approach is a knowledge-based one. User characteristics and various types of information are joined in a knowledge graph, which is the internal representation of the message to be communicated. The output messages, being different depending on user-characteristics, are derived from the same abstract knowledge representation in the system. Essentially four processes are involved in the information system envisaged. The 'determine mode' process determines the modality (text vs. picture) of each fragment of the information to be presented. The 'generate expression' process generates Dutch sentences in a visual format suitable for communicative purposes. The 'generate graphics' process produces the pictures. In the 'format multimodal text' process, the sentences and pictures are integrated and structured in a way required by the input message.

A final aspect discussed here is access to several databases rather than to a single one. The *VMNW*-system contains three subsystems integrated in one database system. When various linguistic databases are available at an institute or even at other places, integration of different databases into one database may be no longer feasible or efficient. This implies the need for an architecture which allows the user to retrieve information from several databases by use of a uniform interface. Merz and King (1994) describe a query facility for heterogeneous, non-integrated databases, from a technical point of view. The databases differ in their models (e.g. relational vs. hierarchical) and other technical aspects. These differences are maintained. A multi-database query language provides a uniform interface for retrieving data from different databases. The multi-database query is decomposed into subqueries with operators supported by the individual database management systems. The global query execution relies on a relational database manager. From the linguistic point of view, Calzolari (1991) discusses the idea of accessing different types of linguistic data and tools available at the institute, in a lexicographic workstation which is conceived as a central module able to link a number of different components. Christ (in press), started the implementation of a modular architecture of a corpus query system, which accesses data originating from different sources, also remote ones.

Most of the methods reported here are still in an experimental or prototype stage and much fundamental research is still needed for their implementation into real systems (cf. De Smedt et al. 1994). The studies however demonstrate that the developers of computer systems start getting more attention for user-friendly systems for use by the non-technical, unexperienced user.

3.5 Concluding remarks

In this section, some interdisciplinary developments have been outlined that may lead to more sophisticated means for access to and analysis of the huge amounts of corpus data in a lexicographical workbench or in a corpus system interface. In the mid-eighties, there was a discussion at our institute about how to reduce the amount of concordances retrievable from the *50 Million Words Corpus* for words with a high frequency, in order to keep the collection of concordances manageable for lexicographical analysis. At the time, statistic-based random reduction seemed to be most feasible. In this section, we have shown that the lexicographer can in principle investigate the whole amount of corpus data available, and manage the data by restricting queries to subcorpora defined by the lexicographer, by grouping concordances along linguistic criteria (head word, part of speech, collocates etc.), and by sorting them according to relevance (cf. 3.1). This, of course, is a much better method than the random reduction.

The developments outlined in this section may increasingly enhance the linguistic analysis of corpus data. Prior classification of texts for retrieval purposes may become unnecessary if this activity could be replaced by high-quality on-line information retrieval. 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.

4. Towards an INL Integrated Language Database of 12th-21st Century Dutch

The developments outlined in the previous section are important to future INL objectives. Within a few years, after completion of the *WNT* and the *VMNW* dictionaries, the INL will have a rich collection of electronic dictionaries, text corpora, and lexical databases, covering many centuries of the Dutch language: the linguistically encoded *Corpus Gysseling* and the electronic *Dictionary of Early Middle Dutch VMNW* covering the 12th century (cf. 2.3), the electronic *Dictionary of the Dutch Language based on Historical Principles WNT* covering the 16th up to the 20th century (2.2), linguistically encoded present-day corpora (2.4), as well as electronic lexica (word lists with lexical information) such as the CELEX data for Dutch, and the Dutch spelling guide *Herziene Woordenlijst Nederlandse Taal* (1990). The period of Middle Dutch is covered by the *Dictionary of Middle Dutch* (Verwijs en Verdam 1885-1929), which will be available in electronic form when the concrete plans to digitize the dictionary will be realized.

Given these collections and given the international recognition of a multifunctional use of electronic corpora and lexica (2.4), it is not surprising that the INL aims at making the data collections reusable for a wider range of users than the INL lexicographers only, by integrating the data into an *INL Integrated Language Database of 12th-21st Century Dutch.* Although detailed linguistic and technical concepts for the integrated language database are not yet available, some basic outlines are clear. The database will be two-dimensional. One is the time dimension; data cover the 12th-21st century. The other is the linguistic dimension; for each century (or whatever period), various types of linguistic data are available: texts (including the quotations in the *WNT* dictionary), dictionary data (including etymology, subcategorization, selection restrictions, chronological, regional and subdomain information, etc.) and various types of linguistic data (e.g. morphological analyses, elaborate morpho-syntactic information, etc.). All these data will be linked in a linguistically well-founded way along the two dimensions. The database will additionally include the types of features and relationships established in thesauruses and knowledge bases. The envisaged integrated language database will be a kind of knowledge base from which the user can easily retrieve information at different linguistic levels and in different representations, about the Dutch language (and culture) over many centuries. To illustrate this, we give two conceivable examples of potential queries. The user enters a meaning description, and the system returns all words that have, or had that meaning in modern or older Dutch, with information about their etymology, domain of use, their use in collocations, etc. Or, a user formulates a query for a modern Dutch word, selects a particular meaning, and, on request, the system provides him with lists of modern or older Dutch synonyms, antonyms, hyponyms, etc., or words with a similar feature (e.g. animate, abstract, tool), whether or not presented in their natural textual contexts derived from texts dating from various periods of time specified by the user. The timing for a feasible implementation of the concept will be dependent on, among other things, the results obtained in the specialisms outlined in section 3. The language database will be useful for diachronic and synchronic, literary and linguistic research, as well as for historians, lawyers, and, last but not least, lexicographers.

The lexicographers of the planned dictionary of present-day Dutch will, of course, have access to the Integrated Language Database. In view of its broader purpose, it is however unlikely that the language database will become a subcomponent of the lexicographer's workbench in a similar way as in the *VMNW* project. More probably, a separate lexicographer's workbench will be developed which meets the very specific needs imposed by the character of the lexicographer's work (cf. 2.3). When the dictionary will have been completed, it will become an additional component of the Integrated Language Database.

5. Conclusion and discussion

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. Compared to the present state of the art (section 2), 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 (section 3). 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). More than ever before, lexicography is influenced by other disciplines.

These developments have implications for the basic knowledge, skills and interests of the lexicographer, traditionally being a linguist. Experience (also at our institute) has shown that no good system can be built by system developers (technical specialists) alone. Users have an important contribution in specifying the functional requirements the system will have to meet, and in evaluating prototypes of the system. Applied to a lexicographer's workbench for computerized corpus-based lexicography, this implies that lexicographers need at least be interested in developments such as those sketched in section 3. In addition to the development of a lexicographic concept for the dictionary, they should preferably contribute to the concept for its implementation into the workbench, including the dictionary database. This tends to require more than superficial interest.

Even when the interest and knowledge are present, it will be difficult to get to grips with the new developments in the different fields and their relevance to lexicography. How to select relevant information from the huge flows of information and keep an overview? How to assess relevance to lexicography and feasibility? How to apply new developments in a project running over many years (cf. 2.2) ? How to provide colleagues with actual relevant knowledge, etc. This requires a thorough understanding of relevant developments in the other disciplines, and ... a lot of time. Should all this be the task of a specialist in theoretical lexicography (cf. Swanepoel 1994: 13, 14)? Or will the present interdisciplinary efforts applied to lexicography lead to a new specialism within language technology: lexicographic technology?

Notes

1. In order to get free on-line access to the INL 5 Million Words Corpus 1994 for non-commercial purposes, a personal user agreement has to be signed. An electronic user agreement form can be obtained from our mailserver "Mailserv@Rulxho.LeidenUniv.NL". Type in the body of your e-mail message: "SEND [5mln94]agreemnt.use" (without the quotes). Please make a hard copy of the agreement form, sign it, keep a copy yourself, and return a signed copy to Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden, The Netherlands. Fax: 31 71 272115. After receipt of the signed user agreement, you will be informed about your user name and password. If you need additional information, please send an e-mail to "Helpdesk@Rulxho.LeidenUniv.NL".

References

Atkins, Beryl T.S. 1992-93. Tools for Computer-aided Corpus Lexicography: The Hector Project. *Acta Linguistica Hungarica* 41(1-4): 5-71. Budapest: Akadémiai Kiadó.

Berg, Donna Lee, Gaston H. Gonnet and Frank W. Tompa. 1988. *The New Oxford English Dictionary Project at the University of Waterloo.* Waterloo, Ontario: UW Centre for the New Oxford English Dictionary OED-88-01.

Biber, Douglas. 1993. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19(2): 219-241.

Boguraev, B. and T. Briscoe (Eds.). 1989. *Computational Lexicography for Natural Language Processing.* London: Longman.

Boguraev, Branimir and Beth Levin. 1990. Models for Lexical Knowledge Bases. *Electronic Text Research. Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*: 65-78. Waterloo: UW Centre for the New OED and Text Research.

Calzolari, Nicoletta. 1991. Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base. Zernik, Uri (Ed.). 1991. *Lexical Acquisition: Exploiting On-Line Resources to build a Lexicon*: 191-208. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Christ, Oliver. In press. A Modular and Flexible Architecture for an Integrated Corpus Query System. *Proceedings of COMPEX'94, 3rd Conference on Computational Lexicography and Text Research*: 23-32. Budapest.

Church, Kenneth, William Gale, Patrick Hanks and Donald Hindle. 1991. Using Statistics in Lexical Analysis. Zernik, Uri (Ed.). 1991. *Lexical Acquisition: Exploiting On-Line Resources to build a Lexicon*: 115-164. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19(1): 1-24.

Clear, Jeremy. 1987. Computing. Sinclair, J.M. (Ed.). 1987. *Looking Up. An Account of the Cobuild Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*: 41-61. London and Glasgow: Collins ELT.

Collier, Alex. 1994. A System for Automatic Concordance Line Selection. Jones, Daniel (Ed.). 1994. *Proceedings of the International Conference on New Methods in Language Processing*: 95-100. University of Manchester, United Kingdom.

Collins Cobuild English Language Dictionary. 1987. London: Harper Collins Publishers.

Doorn, Peter, Eric Helsper, Rene van Horik, Ellen Leenarts and Carlo Vreugde (Eds.). 1993. *Optical Character Recognition in the Historical Discipline. Proceedings of an International Workgroup organized by: Netherlands Historical Data Archive and Nijmegen Institute for Cognition and Information.* St Katharinen: Max-Planck-Institut für Geschichte, Scripta Mercaturae Verlag.

Glassman, Lucille, Dennis Grinberg, Cynthia Hibbard, James Meehan, Loretta Guarino Reid and Mary-Claire van Leunen. 1992. Hector: Connecting words with Definitions. *Screening Words: User Interfaces for Text. Proceedings of the Eighth Annual Conference of the UW Centre for the New OED and Text Research*: 37-74. Waterloo: University of Waterloo.

Gysseling, Maurits. 1977-87. *Corpus van Middelnederlandse teksten (tot en met het jaar 1300).* s' Gravenhage: Martinus Nijhoff.

Harteveld, P. 1991. Die rekenarisering van die leksikografiese prosesse in die Buro van die WAT. *Lexikos* (AFRILEX-reeks 1): 128-157.

Harteveld, Pieter. 1994. The Computerization of the Lexicographical Processes at the Bureau of the Woordeboek van die Afrikaanse Taal (WAT). Martin, Willy, Willem Meys, Margreet Moerland, Elsemiek ten Pas, Piet van Sterkenburg and Piek Vossen (Eds.). 1994. *Euralex 1994 Proceedings*: 449-458. Amsterdam.

Herziene Woordenlijst Nederlandse Taal. 1990. s'-Gravenhage: SDU Uitgeverij.

Karssen, Zeger, Gemme Schwartzenberg and Joost de Jonge. 1994. Understanding Conceptual Information Retrieval. Noordman, L.G.M. and W.A.M. de Vroomen (Eds.). 1994. *Informatiewetenschap 1994, Wetenschappelijke Bijdragen aan de Derde StinfoN-conferentie*: 27-38. Tilburg.

Kruyt, J.G. 1989. Gecomputeriseerde woordenboeken voor mens en computer. *Jaarboek van de Stichting Instituut voor Nederlandse Lexicologie 1988*: 53-72. Leiden: INL.

Kruyt, J.G. 1995. Nationale tekstcorpora in internationaal perspectief. *Forum der Letteren* 36(1): 47-58.

Kruyt, J.G. and J.J. van der Voort van der Kleij. 1992. Towards a Computerized Historical Dictionary of Dutch: from Printed Dictionary to Correct Text File. Kiefer, Ferenc, Gabor Kiss and Julia Pajzs (Eds.). 1992. *Papers in Computational Lexicography COMPLEX '92*: 203-210. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Kruyt, Johanna G. and John J. van der Voort van der Kleij. 1992-93. Towards a Computerized Historical Dictionary of Dutch. *Acta Linguistica Hungarica* 41(1-4): 159-174. Budapest: Akadémiai Kiadó.

Kruyt, J.G. and J. van der Voort van der Kleij. 1993. Converting the Historical Dictionary of Dutch to Electronic Form. Doorn, Peter, Eric Helsper, Rene van Horik, Ellen Leenarts and Carlo Vreugde (Eds.). 1993. *Optical Character Recognition in the Historical Discipline. Proceedings of an International Workgroup organized by: Netherlands Historical Data Archive and Nijmegen Institute for Cognition and Information*: 131-138. St Katharinen: Max-Planck-Institut für Geschichte, Scripta Mercaturae Verlag.

Longman Language Activator. 1993. Essex: Longman Group UK Limited.

Magay, T. and J. Zigány (Eds.). 1988. *BudaLEX '88 Proceedings. Papers from the 3rd International EURALEX Congress.* Budapest: Akadémiai Kiadó.

Martin, W., F. Platteau and R. Heymans. 1985. *Naar een Corpus voor een Woordenboek Hedendaags Nederlands. Mogelijkheden en Beperkingen van het Gebruik van Corpora in Lexicografisch Onderzoek.* Ongepubliceerd rapport. Universitaire Instelling Antwerpen.

Merz, Ulla and Roger King. 1994. Direct: A Query Gacility for Multiple Databases. *ACM Transactions on Information Systems* 12(4): 339-359.

Moerdijk, A. 1994. *Handleiding bij het Woordenboek der Nederlandsche Taal (WNT).* 's-Gravenhage: SDU Uitgeverij.

Pijnenburg, Willi J.J. 1991. Das >Vroegmiddelnederlands Woordenboek (1200-1300)<: Seine Bedeutung für die computergestützte Lexicographie in Belgien und in den Niederlanden. Gärtner Kurt, Paul Sappler and Michael Trauth (Eds.). 1991. *Maschinelle Verarbeitung altdeutscher Texte IV, Beiträge zum Vierten Internationalen Symposion Trier 28. Februar bis 2. März 1988*: 60-67. Tübingen: Max Niemeyer Verlag.

Schoonheim, Tanneke. In press. The Vroegmiddelnederlands Woordenboek. *International Medieval Research — A New Methodological Annual. Proceedings of the First International Medieval Congress.* Leeds 1994.

Simpson, John. 1986. Opening Address: The New OED Project. *Information in Data. Proceedings of the First Conference of the UW Centre for the New Oxford English Dictionary*: 1-6. Waterloo.

Sinclair, J.M. (Ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing.* London, Glasgow: Collins ELT.

Sinclair, John M. 1994. LIS/MAS Software. *MLAP93-21 MECOLB Quarterly Report II August, September, October 1994.* Mannheim: Institut für Deutsche Sprache.

Sinclair, John M. 1995. LIS & MAS. *MLAP93-21 MECOLB Progress Report II August 94 - January 1995.* Mannheim: Institut für Deutsche Sprache.

Smeaton, Alan F. 1994. Using Syntactic Level Language Analysis for Document Retrieval: Where Does it Fit and Does it Help? Noordman, L.G.M. and W.A.M. de Vroomen (Eds.). 1994. *Informatiewetenschap 1994, Wetenschappelijke Bijdragen aan de Derde StinfoN-conferentie*: 15-25. Tilburg.

Smedt, Koenraad de, Wim Claassen and Carla Huls. 1994. GeKnIPT: op kennis gebaseerde informatiepresentatie (een projectvoorstel). Noordman, L.G.M. and W.A.M. de Vroomen (Eds.). 1994. *Informatiewetenschap 1994, Wetenschappelijke Bijdragen aan de Derde StinfoN-conferentie*: 157-165. Tilburg.

Sparck Jones, Karin. 1995. Natural Language Processing: A Historical Review. Zampolli, Antonio, Nicoletta Calzolari and Martha Palmer (Eds.). 1995. *Current Issues in Computational Linguistics: In Honour of Don Walker*: 3-16. Pisa: Giardini editori e stampatori in Pisa and Dordrecht: Kluwer Academic Publishers.

Sterkenburg, P.G.J. van. 1983. Dutch Lexicology and the Computer. Zampolli, A. and A. Cappelli (Eds.). 1983. *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries. Proceedings of the European Science Foundation Workshop, Pisa,* 1981: 201-206. Pisa: Giardini Editori e Stampatori in Pisa.

Sterkenburg, P.G.J. van and J.G. Kruyt. In press. *Dutch Electronic Corpora: Their History, Applications and Future.* Computer and the Humanities.

Swanepoel, Piet. 1994. Problems, Theories and Methodologies in Current Lexicographic Semantic Research. Martin, Willy, Willem Meys, Margreet Moerland, Elsemiek ten Pas, Piet van Sterkenburg and Piek Vossen (Eds.). 1994. *Euralex 1994. Proceedings*: 11-26. Amsterdam.

Verwijs, E. and J. Verdam. 1885-1929. *Middelnederlandsch Woordenboek.* 's-Gravenhage: Martinus Nijhoff.

Voort van der Kleij, J.J. van der. 1986. The Use of Optical Character Readers in Machine Lexicography. Zimmerman, H.H. (Ed.). 1986. *Proceedings IV Forum Information Science and Practice, Standardization in Computerized Lexicography*: 285-294. Saarbrücken: European Science Foundation and Institut der Gesellschaft zur Förderung der angewandten Informationsforschung.

Voort van der Kleij, John van der, Stephan Raaijmakers, Mathijs Panhuijsen, Martijn Meijering and Rogier van Sterkenburg. 1994. Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem. Noordman, L.G.M. and W.A.M. de Vroomen (Eds.). 1994. *Informatiewetenschap 1994, Wetenschappelijke Bijdragen aan de Derde StinfoN-conferentie*: 181-194. Tilburg.

Wiesman, Floris. 1995. Information Retrieval: een overzicht. *Informatie* 37(3): 182-192.

Woordenboek der Nederlandsche Taal·WNT. 1864-. Leiden.

Zampolli, A. and A. Cappelli (Eds.). 1983. *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries. Proceedings of the European Science Foundation Workshop, Pisa, 1981.* Pisa: Giardini Editori e Stampatori in Pisa.

Zernik, Uri. 1991. Introduction. Zernik, Uri (Ed.). 1991. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*: 1-26. Hillsdale, New Jersey: Lawrence Erlbaum Associates.