A 38 Million Words Dutch Text Corpus and its Users

J.G. Kruyt and M.W.F. Dutilh,
Instituut voor Nederlandse Lexicologie INL, Leiden, Nederland

Abstract: The use of text corpora has increased considerably in the past few years, not only in the field of lexicography but also in computational linguistics and language technology. Consequently, corpus data and expertise developed by lexicographical institutions have gained a broader scope of application. In the European context this has led to a revised view of corpus design. In line with these developments, the Institute for Dutch Lexicology (INL) has since 1994 been providing external access to steadily improving corpora via Internet. In August 1996, the *38 Million Words Corpus* was available for consultation by the international research community. The present paper reports on the characteristics of this corpus (design, text classification, linguistic annotation) and on its use, both in dictionary projects and in linguistic research. In spite of limitations with respect to corpus design, the INL corpora accessible via Internet have proved to meet external needs. By providing these facilities, the INL has acquired a much broader experience in corpus-building than before, which is essential for new, internal dictionary projects. Giving external access to corpus data which was developed primarily for internal purposes, may be profitable for all parties involved.

Keywords: LARGE ELECTRONIC DUTCH TEXT CORPUS, CORPUS DESIGN, TEXT CLASSIFICATION, TOPIC, PUBLICATION MEDIUM, LINGUISTIC ANNOTATION, ON-LINE ACCESS VIA INTERNET, CORPUS USERS

Samenvatting: Een tekstcorpus Nederlands (38 miljoen woorden) en de gebruikers ervan. Het gebruik van tekstcorpora is de laatste jaren aanzienlijk toegenomen, niet alleen op het gebied van de lexicografie maar ook in de computationele linguïstiek en de taaltechnologie. Ten gevolge daarvan kregen de corpusdata en de expertise opgebouwd door lexicografische instellingen een breder toepassingsdomein. Op Europees niveau leidde dit tot een herziene visie op corpussamenstelling. In overeenstemming met deze ontwikkelingen, geeft het Instituut voor Nederlandse Lexicologie (INL) sinds 1994 externe toegang via Internet tot steeds beter wordende corpora. In augustus 1996 was het *38 Miljoen Woorden Corpus* gereed voor consultatie door het internationale onderzoeksveld. Dit artikel beschrijft de karakteristieke kenmerken van dit corpus (corpussamenstelling, tekstclassificatie, linguïstische annotatie) en het gebruik in zowel woordenboekprojecten als in taalkundig onderzoek. Ondanks beperkingen ten aanzien van corpussamenstelling, is duidelijk gebleken dat de INL corpora die via Internet toegankelijk zijn, voorzien in een externe behoefte. Door deze faciliteiten aan te bieden, heeft het INL een veel bredere ervaring in corpusopbouw opgedaan dan voorheen. Deze is van essentieel belang voor nieuwe interne woordenboekprojecten. Het verlenen van externe toegang tot corpusdata die primair voor interne doeleinden ontwikkeld zijn, kan voor alle betrokken partijen profijt hebben.

Lexikos 7 (AFRILEX-reeks/series 7: 1997): 229-244

Trefwoorden: GROOT ELEKTRONISCH NEDERLANDS TEKSTCORPUS, CORPUSSAMENSTELLING, TEKSTCLASSIFICATIE, ONDERWERPSDOMEIN, PUBLICATIEMEDIUM, LINGUISTISCHE ANNOTATIE, ON-LINE TOEGANG VIA INTERNET, CORPUSGEBRUIKERS

1. Introduction

In the early eighties, large electronic text corpora of national languages were developed mainly for lexicographical purposes (Zampolli and Cappelli 1983). Until the early nineties, however, a major problem was the management of the huge amounts of data stored in the computer, which caused lexicographers still to work with paper copies of concordances (cf. Clear 1987). Presently, more flexible access to large corpora is feasible. Corpora published on CD-ROM are distributed by the *Linguistic Data Consortium* (LDC) in the USA and the *European Language Resource Association* (ELRA) in Europe. Several institutions with a long-standing lexicographical background provide access to corpora via the Internet, for example *CobuildDirect Service* (Krishnamurthy 1996), the Italian *DBT* (Biagini and Picchi 1996) and the Dutch corpus services of the Institute for Dutch Lexicology INL (Kruyt 1995a, b, Kruyt *et al.* 1995).

The use of corpora has increased considerably in the past few years. Recent studies show the importance of corpus data for lexicography (e.g. Noël *et al.* 1995, several studies in Gellerstam *et al.* 1996 and Kiefer *et al.* 1996). Major publishers spend money on commercial corpus-based dictionaries, such as *Collins Cobuild English Language Dictionary* (1987) and *Longman Language Activator* (1993). Outside the field of lexicography, large corpora have become important for computational linguistics (Church and Mercer 1993) and language technology. From the perspective of a European infrastructure for language technology, the European Commission considered the corpus data and expertise developed by lexicographical institutions important enough to support projects in which the institutions contribute to the realization of the intended European infrastructure (cf. Kruyt 1995a, Teubert 1995, Zampolli 1996).

Corpora users have different attitudes towards corpus design. Lexicographers traditionally aim at a "representative" or "balanced" corpus, that is, the corpus should be appropriate as the basis for generalizations concerning the language as a whole. Corpus size (very large corpora) rather than corpus design is considered essential by many computational linguists using statistical methods of language analysis (cf. Church and Mercer 1993). Biber (1994) shows how complex it is to achieve "representativeness", even with the present computational methods for language analysis. Indeed, corpus practice demonstrates that lexicographical corpora for standard-language dictionaries may have very different corpus designs (Kruyt and Putter 1992, Kruyt and Van Sterkenburg 1996). The complexity of the notion "representativeness" (cf. Teubert 1995: 119), the different interests of corpus users and the costs of corpus development, have, at a European level, led to a shift of focus from building a separate, closed corpus for each project or application towards the development of

reusable, multifunctional and harmonized reference corpora for the European languages (Zampolli 1996). Flexible corpus use is ensured by the option of selecting user-defined subcorpora from a very large corpus with a composition as diversified as possible (Kruyt and Van Sterkenburg 1996).

In line with the interest in corpora and the European views on corpus development, the INL has broadened its scope in the past few years. Besides the ongoing compilation of the dictionary projects *Woordenboek der Nederlandsche Taal* (WNT) en *Vroegmiddelnederlands Woordenboek* (VMNW), the INL decided to participate in European corpus and lexicon projects. The INL also decided to make corpora accessible via Internet, so as to provide corpus facilities to the (inter)national research community. The INL opted for a phased approach, i.e. developing steadily improving corpora. Although representativeness was not aimed at, corpus design was well thought-out. In 1994, a *5 Million Words Corpus*, with a diversified composition and automatically annotated for lemma and part of speech, was made accessible via Internet (Kruyt 1995a, b). A *27 Million Words Newspaper Corpus*, with improved linguistic annotation and retrieval functionalities, followed in 1995 (Kruyt *et al.* 1995). At the end of August 1996, a *38 Million Words Corpus* with a diversified composition was made available in a similar way. This corpus is different from the former ones in various aspects: (a) size, (b) a broader coverage with respect to topic (subject domain), text types (with publication medium as parameter) and time span, (c) a more extended linguistic annotation, (d) the application of international standards for text classification and linguistic annotation, and (e) improved retrieval functionalities. The *38 Million Words Corpus* and its users will be characterized in the following sections. Where relevant, the use of the other INL corpora accessible via Internet will be discussed.

2. INL 38 Million Words Corpus 1996

2.1 Composition

The INL has been acquiring electronic texts from several publishing houses since 1992. For reasons of copyright under Dutch law, the types of use permitted by the copyright holder are specified in a written contract between the INL and the copyright holder. Most text providers so far have given permission for internal use, as well as for external consultation by Internet for noncommercial research purposes. External use was particularly relevant to the *38 Million Words Corpus*. The required permission of the copyright holders limited the availability of texts to be incorporated into the corpus. Under this restriction, a corpus as diversified as possible was aimed at, so as to offer the research community an optimal opportunity to investigate language phenomena in different text types.

Corpus texts have been selected from the INL electronic text archive according to the following criteria. The language covered is standard Dutch

and Flemish (i.e. no dialect) as used in the Netherlands and Belgium. The corpus should preferably consist of components with a more or less equal size but with different contents. Broad coverage and balanced proportions were aimed at with respect to topic, publication medium and time span. However, coverage and balance were affected by availability or copyright restrictions, and in some cases, by inappropriate text formats. The resulting corpus consists of three main components: a component with varied composition (ca. 12,7 million words), a newspaper component (ca. 12,4 million words), and a component of legal texts (ca. 12,9 million words).

The varied component covers the period 1970-1995. It includes 18 single books and one title with 24 volumes, texts from issues of seven magazines, texts from 50 daily issues of the Belgian newspaper *De Standaard* (other newspapers being incorporated in other INL (sub)corpora), texts written to be read out in TV news broadcasts for adults and for youths, 18 *Queen's Speeches*, parliamentary reports over two months, and three issues of the *Law Gazette of the Kingdom of the Netherlands*. This subcorpus covers six topics (cf. section 2.2). For most text sources, all available text material for the purpose has been included. From some magazines with a large number of annual issues, half or a quarter of the issues have been selected for reasons of balance. For a more detailed survey, see the appendix.

The newspaper component consists of issues of the *Meppeler Courant*, dating from 1992-1995. Another newspaper available at the INL, *NRC*, was not selected for this (sub)corpus, as it forms part of the contents of the INL's *27 Million Words Newspaper Corpus 1995*. The selected newspaper is published three times a week. The INL receives a selection of articles per newspaper issue. All the material available up to 1996, grouped into monthly files, has been included. Two topics ("mixed" and "sports") are covered (cf. section 2.2).

The legal text component is a compilation of Dutch legal texts operative in 1989, including 5,875 laws, orders and decrees, protocols, agreements, treaties or conventions, etc., dating from 1814 up to 1989. This subcorpus has been derived from the *NLEX* database (the version without the text added by the publisher), with exclusion of texts undated or written in French.

2.2 Text classification

The corpus texts have been classified according to two parameters, viz. publication medium (in a broad sense) and topic (subject domain). For both parameters, a set of classification categories was distinguished on the basis of external, rather than linguistic criteria (cf. Biber 1994). The value of this (traditional) type of classification for corpus linguistics is criticized, particularly with regard to topic (Sinclair and Ball 1995). However, a new, commonly accepted, linguistically founded classification scheme has not been developed yet. In classifying the corpus texts, our sole intention was to assist the researcher in defining subcorpora from the whole corpus (cf. section 2.3).

The publication medium categories distinguished are: "book", "newspaper", "magazine", "written to be spoken", "reported speech", and "miscellaneous". Reference works (*Handboek van de Nederlandse pers en publiciteit*) or specific codes (ISBN, ISSN) assigned by the publisher have been used for classifying corpus texts as "newspaper" and "magazine" or "book" respectively. "Written to be spoken" refers to a text written beforehand, which is to be read out in public. In the corpus, this category is covered by the TV broadcast texts and the *Queen's Speeches*. "Reported speech" refers to a grammatically and stylistically corrected report of spoken language, rather than to transcribed spoken language. The parliamentary reports belong to this category. "Miscellaneous" includes texts that could not be classified in one of the former categories.

For topic classification, the two-level classification scheme proposed by Norling-Christensen (1996) for topic classification in the European *PP-PAROLE* project[1] is applied, based on the topic scheme used for the corpus underlying the Danish Dictionary edited by the Society for Danish Language and Literature (DSL) in Copenhagen.

Our corpus texts appeared to cover only part of the (sub)categories of the *PAROLE* scheme. The resulting topic categories are "HEALTH" with subcategories "health" and "psychology", "HUMANITIES" with subcategories "philosophy" and "language", "LEISURE" with subcategories "leisure" and "sports", "SCIENCE" with subcategories "astronomy" and "environment", "SOCIETY" with subcategories "social studies", "politics" and "law". A final category "MIXED" refers to texts covering a broad variety of topics, e.g. newspapers. For magazines, topic classification is based on branch-codes listed in the reference work *Handboek van de Nederlandse pers en publiciteit*, which have been translated into the *PAROLE* topic categories. For books published since 1980, so-called *CIP* (*Cataloguing in Publication*) data are available in the source. *CIP* data include up to three codes, which have their origin in different Dutch classification schemes (e.g. *UDC*, *NUGI*, *SISO*), as well as keyword terms. This data is reinterpreted in terms of the *PAROLE* topic scheme. Books without *CIP* data are classified on the basis of the title of the book or information in the front or back matter (cf. Dutilh and Kruyt 1992). Newspapers and TV news texts, covering many topics, have been classified as "mixed". However, the sports pages of the *Meppeler Courant* could be classified as "leisure / sports", based on the title of these pages as encoded in the electronic files. The classification of the remaining texts was based on general knowledge about the text.

2.3 Access to the corpus data

A retrieval (corpus query) system has been developed which enables the researcher to search for single words or for word patterns in the corpus, including some rather primitive, predefined word classes (e.g. past participle) and syntactic patterns (e.g. noun phrase NP, prepositional phrase PP) which can be customized and extended by the user. The result of a query is, in the end, a

series of concordances (keywords in context) meeting the query specifications. A major problem in information retrieval is the effectiveness of a search (recall and precision) (cf. Kruyt 1995b), in our case, the extent to which the query system retrieves the exact linguistic data the researcher needs from the corpus (no more and no less). Two functionalities of the corpus retrieval system reduce the overflow of data in the output of a query. One is linguistic annotation in terms of lemma (headword) and part of speech (POS). The other is the option to select a user-defined subcorpus from the whole corpus.

The researcher may address a query to a subcorpus selected from the perspective of his research purposes rather than to the whole corpus. Corpus composition (section 2.1), text classification (section 2.2), and text date enable the user to select subcorpora easily according to the parameters corpus component, topic, publication medium, and period. Selection of one, more or all of these options results in the display of text source surveys on the screen. The researcher has the opportunity to select individual text sources from these surveys. In this way, each researcher can define his own subcorpora, based on selection at the level of individual texts. This reduces an overflow of output caused by data meeting the query but coming from texts without relevance to the research purposes. For each defined subcorpus, its size can be displayed on the screen.

The other functionality reducing an overflow of irrelevant data is linguistic annotation, the explicit encoding of linguistic features in the electronic text (cf. Grefenstette 1996). The main function of linguistic annotation is that searches may be specified in terms of various linguistic features. The word-forms (tokens) in the corpus texts have automatically been annotated with lemma (headword) and two types of part of speech (POS)[2]. One POS scheme includes thirteen basic POS categories (Van der Voort van der Kleij *et al.* 1994). The other POS scheme is fine-grained, each POS being subcategorized in terms of type and/or characteristic features, conformant with the European *MECOLB* standard[3] (Raaijmakers and Dutilh 1995). For example, the *MECOLB* POS tag for the word-form "loopt" ("walks") is "VRB (intrans, indic, pres, sg, 2/3)", specifying it as an intransitive verb with its values for the features mood, tense, number and person. As a result of the linguistic annotation, a query may include references to specific word-forms, to specific basic parts of speech, to *MECOLB* parts of speech, *MECOLB* POS subtypes and features, and to headwords (lemmas), either separately or combined in one query definition. For example, the following searches may be expressed in the formal query language:

(1) "Search the occurrences of the word-form 'werk' (work)". The first output is a list containing "werk" specified as noun and "werk" specified as verb. The user makes a selection and the relevant concordances appear on the screen.

(2) "Search the occurrences of the word-form 'werk' under the condition that 'werk' is a noun". This query, with the double specification ("werk" plus

"as noun"), immediately results in the relevant concordances (i.e. without occurrences of the verb form).

(3) "Search the occurrences of the lemma 'president' (president)". The output is a series of concordances with occurrences of singular "president" and plural "presidenten".

(4) "Search the occurrences of the lemma 'president' followed by a prepositional phrase PP, within a distance of (say) 7 arbitrary word-forms". The output concordances show occurrences of word-forms of the lemma "president" (cf. example (3)) only if they are followed by a PP within the specified distance (and not all the other ones without the PP). For example: "president van Amerika" (president of America), "presidenten uit diverse landen" (presidents from various countries), instead of "president", "presidenten" not followed by the PP specification.

(5) "Search the occurrences of the lemma 'werken' (to work) with *MECOLB* feature 'present tense'". The output is a series of concordances with occurrences of the verb "werken" as far as they have been annotated by the feature "present tense".

From the perspective of search effectiveness, the annotated corpus has an added value with respect to a "raw" (not annotated) corpus or "raw" texts available on CD-ROM, Internet, etc. "Raw" text can essentially be addressed at the level of word-form (token) only, whereas annotated text can be addressed at all linguistic features expressed by the encoding (including combinations). Due to the headword and POS annotation, the researcher does not need to specify the whole paradigm of a word. This is not only a matter of user-friendliness. For ambiguous word-forms such as "school" (ambiguous for noun "school" and verb form "sheltered") or "sleep" (ambiguous for noun "train", and verb forms "polished" and "drags"), the headword and POS encoding enables the search engine (the computer program) to discriminate between the different headwords. As a consequence, the researcher will only retrieve the occurrences of the headword he is interested in, and not all the others as well. The facility of combining search specifications or conditions in one query (examples (2), (4) and (5)) allows an additional curtailment of the output. As opposed to these positive effects on effectiveness, it should be noticed that incorrect encodings and encodings that could not be disambiguated, result in some overflow of data (found but not intended by the researcher) and/or some deficiency of data (intended but not found). For a more detailed description of retrieval facilities, see Kruyt *et al.* (1995).

2.4 Use of the 38 Million Words Corpus

The *38 Million Words Corpus* was ready for consultation via Internet at the end of August 1996[4]. It is consulted by lexicographers in dictionary and lexicon

projects as well as by individual users for various purposes. In the short term, it will be used in university courses in corpus linguistics. The INL keeps record of particular user data, not only so as to trace potential misuse, but also and more positively, to obtain insight into the needs of the corpus users. Where relevant, other INL corpora will be referred to, particularly the earlier *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995* which also are accessible via Internet. Published work refers to these corpora rather than to the most recent one.

2.4.1 Use by lexicographers

The *38 Million Words Corpus* is used in the preparatory phase of a new dictionary project at the INL, which will start after completion of the *Woordenboek der Nederlandsche Taal* (*WNT*) in 1998. Depending on the concept for this dictionary, texts will be selected from the INL text archive and corpora for incorporation into the closed corpus for the dictionary. The assumption is that additional texts are to be selected and acquired.

In addition to this strictly internal use, the *38 Million Words Corpus 1996* is being consulted within the framework of several international corpus-based lexicon projects. The Dutch-Flemish project *Referentie Bestand Nederlands* (*RBN*) (*Reference Database of the Dutch Language*), a project under the authority of *Commissie Lexicografische Vertaal Voorzieningen* (*CLVV*) (*Committee for Lexicographical Translation Facilities*) and supervised by Prof. dr. W. Martin, aims at the development of a lexical database for the purpose of noncommercial dictionaries with Dutch as either source or target language. The INL corpora were used for the composition of the entry list. Prof. Martin selected a subcorpus of ca. 10 million words from the *38 Million Words Corpus* to be consulted by the lexicographers for determining the contents of several fields of the microstructure (e.g. lexical and grammatical collocations, idioms). Lexicographers from several cities in the Netherlands and Belgium work on the INL computer system daily (cf. diagram on p. 238).

The EC-funded project *LE-PAROLE*[5] aims at the development of comparable corpora (each 20 million words) and lexica (each 20,000 entries) for 12 Western European languages, according to European standards with respect to linguistic background, contents, linguistic annotation schemes, text representation and access. The INL is responsible for the Dutch corpus and lexicon. The entry list for the Dutch lexicon has been determined on the basis of linguistically annotated INL corpora containing a total of ca. 54 million words: the *27 Million Words Newspaper Corpus 1995*, a *15 Million Words Corpus* with diversified composition and the varied component (ca. 12 million words) of the *38 Million Words Corpus*. The varied component and the newspaper component of the *38 Million Words Corpus* have been selected as subcorpora for determining syntactic complementation patterns for various types of POS.

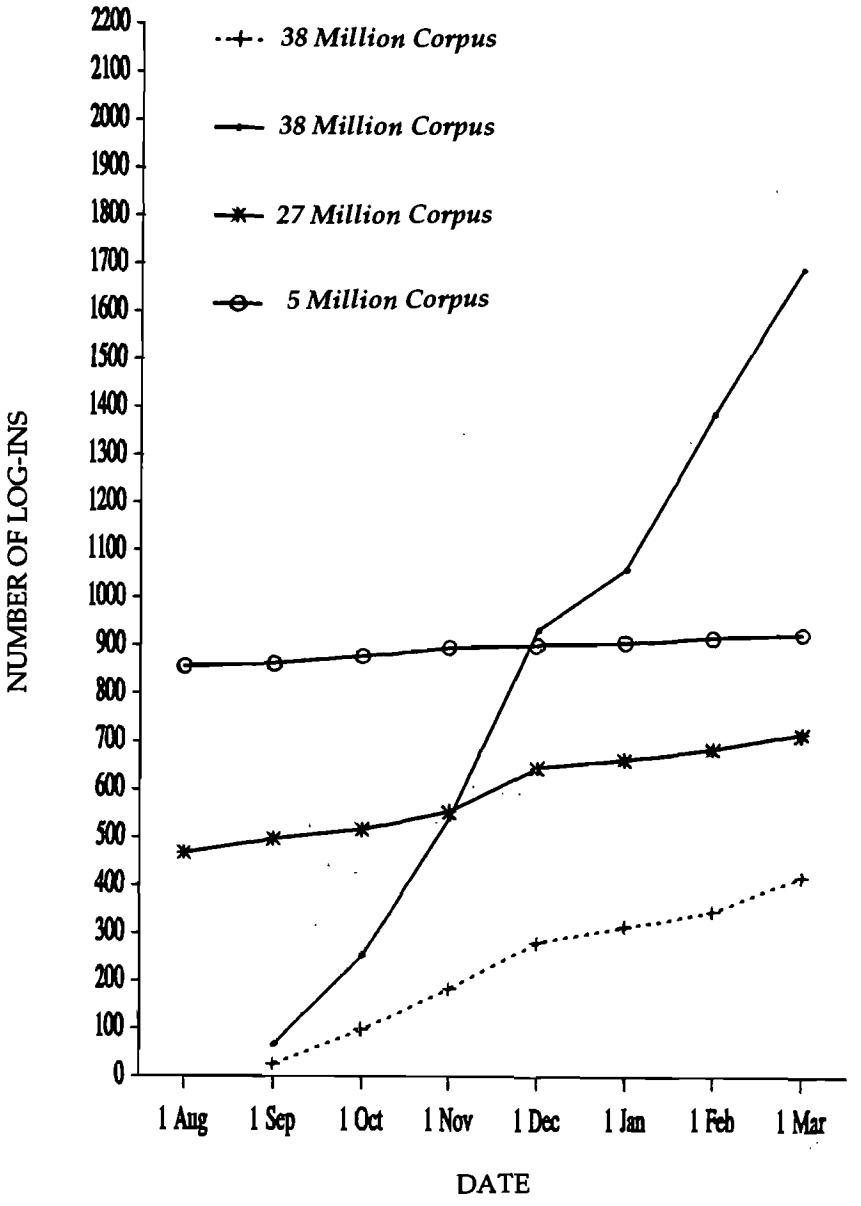
Vliegen (1996) studied complementation patterns of Dutch verbs of visual, auditory, olfactory and gustatory perception and verbs of verbal communication within the framework of the EC-funded project *DELIS* (LRE 61.034). This project aimed at methods and tools to build lexical entries based on evidence extracted from textual corpora, combining a corpus-based lexicographical approach and frame-based semantic theory. In Vliegen (1996), the *5 Million Words Corpus 1994* is mentioned among the corpora he consulted for this study. The INL user records show that he also consulted the *27 Million Words Newspaper Corpus 1995* and the *38 Million Words Corpus 1996* for his research on verbs of perceiving and verbal communication.

2.4.2 Use by individuals

Since the *38 Million Words Corpus 1996* was developed rather recently, the individual use of this corpus will be considered against the background of the use of the earlier INL corpora accessible via Internet (the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995*). Only external (i.e. non-INL) use will be discussed. Note that the figures presented reflect the momentary status on a particular date; figures change daily.

By March 1, 1997, 175 external users signed a personal user agreement for one or more INL corpora: 138 have access to the *5 Million Words Corpus 1994*, 98 to the *27 Million Words Newspaper Corpus 1995*, and 58 to the *38 Million Words Corpus 1996*. It should be noticed that ca. 30 subscribers had not consulted the corpora yet, and that 16 user accounts are reserved for students of the Free University of Amsterdam, who will follow a short-term course in corpus linguistics. Monthly user records show that the number of users of each of the corpora is steadily growing (two new users per month on the average for the earlier corpora; 3,5 per month on the average for the most recent one). The users are mainly (over 80%) from the Netherlands and Belgium; ca. 14% comes from Germany, the USA, the United Kingdom and South Africa. Other users are from Norway, Denmark, Austria, Slovenia, Latvia, Malaysia and Korea.

By March 1, 1997, the corpora were accessed 3337 times in total. "Accessed" means that a user made contact with ("logged in to") the INL computer so as to address one or more queries to a corpus. For each of the corpora, the diagram (see p. 238) shows the number of log-ins over the period August 1996 up to March 1997. The rates for the *38 Million Words Corpus* increase very fast. This can mainly be explained by its use by the lexicographers of the *RBN* project (see under section 2.4.1). The number of log-ins by the other users is represented by the dotted line. If this line is taken into account, the three corpora show steadily rising curves, although somewhat flatter for the older ones than for the latest one. The highest rate (926) is still for the oldest, the *5 Million Words Corpus*. The average number of queries per consultation is very different for the corpora and for the individual users. As a rough indication, the proportion can be fixed at four to five queries per consultation.



Number of log-ins in the period of August 1, 1996 up to March 1, 1997, for three INL corpora containing 5 million, 27 million and 38 million words, respectively. The dotted line shows the number of log-ins for the *38 Million Words Corpus*, excluding the users of the *RBN* project.

From publications and from an analysis of the queries over the past half-year (cf. Kruyt 1995a for earlier research with INL corpora), it can be concluded that the corpora are consulted for essentially two purposes: incidental looking up of particular words or phrases and research in the field of linguistics and social studies. Some examples of research are the following: Hoeksema and Klein (1996) investigated the usage of Dutch "even" (equally) as a comparative and as an adverb of degree within the framework of the *PIONIER* project *Reflections of Logical Patterns in Language Structure and Language Use.* Pollman (1996) used the historical distribution of dates in newspaper corpora for a psychological essay about memory and the systematics of collective historical consciousness. In her thesis, Cornelis (1997) investigated the passive construction in several Dutch text corpora. Corpus composition appeared to have an impact on the results (Cornelis 1997: 209). Other research topics appearing from the queries concern the orthography of geographical names, male and female variants of nouns, conjunctions, reciprocal pronouns, verbs with strong and weak inflection, inflection of separable verbs, specific verb and noun constructions, words with particular prefixes or suffixes, the vocabulary in the field of social legislation, fashionable words and neologisms.

3. Conclusion and discussion

By providing easily accessible instruments for corpus-based research, the three INL corpora on Internet have proven to meet external needs from the (inter)-national research community. The function of the *38 Million Words Corpus* in international lexicon projects demonstrates its relevance for lexicographical purposes in spite of its shortcomings with respect to corpus design and text classification (cf. section 2).

Our conclusion is that a less ideal corpus is apparently better than no corpus. This may apply particularly to a minority language such as Dutch, as there exist no other Dutch corpora comparable in size, coverage, linguistic annotation and easy access. In English, for instance, several large corpora are available (e.g. the *British National Corpus* and the *Cobuild Bank of English*). Rather than a series of steadily improving corpora (cf. section 1), the INL might at once have opted for an ideally representative general-language corpus composed according to the principles outlined by Biber (1994). In order to achieve representativeness, Biber proposes a cyclical method consisting of four stages: (1) pilot empirical investigation / theoretical analysis, (2) corpus design, (3) compiling a portion of the corpus with grammatical tagging (pilot corpus), and (4) empirical investigation on the pilot corpus by automatic language analysis. The results of stage (4) are used to confirm or modify the design parameters of stage (2), and the process is repeated until representativeness is reached. A critical factor with respect to this method, though promising, seems to be feasibility due to, among other things, labour-intensiveness and, for Dutch, the lack of machine-readable texts covering the various registers and copyright restrictions. The INL could

not have met the needs for corpus data, as has been done since 1994, if a method like the one proposed by Biber had been applied.

The success of the INL corpora can be understood from the efforts needed for corpus-building. The development of the corpora required several man-years per corpus. Additionally, the INL has the technical infrastructure and the specialists in different disciplines (lexicographers/corpus linguists, computational linguists and information scientists) needed for large-scale, annotated corpora. For researchers, a corpus is an instrument rather than an end in itself. But even in the lexicographical projects referred to above, the use of an available, easily accessible corpus was preferred to building a corpus specifically for the purpose. In these projects particularly, the option of defining subcorpora has been applied.

From an internal point of view, the development of the three corpora, particularly the last one, has yielded much experience and insight with respect to the procedures to be followed, their routing, the time needed for the various phases and the problems to be solved. This experience will be very useful for the planning of new dictionary projects to be started at the INL after the completion of the current ongoing dictionary projects. The corpora developed so far may function as pilot corpora in the sense of Biber's stage 3.

For the near future (1997-1998), the INL aims at expanding and enhancing the research instruments, both for internal and external use. Within the framework of the *LE-PAROLE* project, a corpus is being prepared in which text structural elements are encoded in *TEI* format, an international standard for the encoding and interchange of electronic text for research purposes, developed in the past years by the *Text Encoding Initiative* (*TEI*) (Sperberg-McQueen 1994). This corpus will be the basis for a large, syntactically annotated corpus, which will enhance retrieval functionalities. Furthermore, the INL intends to offer the research community on-line annotation facilities which enable the researcher to annotate his own texts by use of linguistic software developed by the INL. Due to the external importance of these facilities, the Netherlands Organization for Scientific Research NWO will cofinance the required hardware.

Until the early nineties, the INL developed corpora for lexicographical purposes only. By broadening its scope (cf. section 1), a much broader experience in corpus-building has been acquired which is indispensable for new internal lexicographical projects and for the INL Integrated Language Database of 12th-21st Century Dutch (cf. Kruyt 1995b). In view of the need for corpus data and the efforts needed for building corpora, corpus builders may, from the outset of a new dictionary project, consider the possibility of giving access to their data to external users and establishing the legal conditions to realize this. For the INL, this has proven to be profitable for all parties.

Notes

1. The *LE-PAROLE* project (LE2-4017) is a project cofinanced by the European Commission. The aim of the project is the development of corpora and lexica for 12 Western European languages, which are comparable with respect to linguistic background, contents, text representation (in *TEI*) and access. They will be used for (multilingual) research and language technology products. The *PAROLE*-specifications for the contents of corpus and lexicon are based on European standards developed in *EAGLES* and related projects and have been specified in the earlier *PP-PAROLE* project (MLAP63-386). *PP-PAROLE* was preceded by the *NERC* project (Calzolari *et al.* 1996), a feasibility study into a Network of European Reference Corpora (cf. Kruyt 1995a).
2. Lemma and basic POS category have automatically been assigned by the lemmatizer / POS-tagger DutchTale, developed by the INL (Van der Voort van der Kleij *et al.* 1994) in the framework of the European *NERC* project (see note 1). Improved versions have been developed by S. Raaijmakers (Kruyt *et al.* 1995). These improvements particularly addressed the encoding of POS and headword for word-forms that were not found in the lexicon, and the disambiguation of word-forms that were assigned more than one POS and/or headword on the basis of the lexicon. For information on the lexicon component, see Van der Voort van der Kleij and Kruyt (1997).
3. The *MECOLB* standard has been developed in the framework of the European project *MECOLB* (MLAP93-21), sponsored by the European Commission and coordinated by R. Neumann, Institut für Deutsche Sprache, Mannheim. The *MECOLB*-tag set for Dutch morphosyntactic annotation was developed in cooperation with the TOSCA Research Group (University of Nymegen), under the direction of Prof. dr. J. Aarts. The *MECOLB* POS-encodings in the corpus have been partially disambiguated by use of a neural network which was trained on a corpus developed in cooperation with the TOSCA Research Group.
4. Access to the INL corpora is provided free of charge for noncommercial research purposes. For each corpus, a separate personal user agreement is to be signed. An electronic user agreement form can be obtained from the INL mail server

   Mailserv@Rulxho.LeidenUniv.NL

   or by request from the INL helpdesk

   Helpdesk@Rulxho.LeidenUniv.NL

   A hard copy of the agreement form must be made, a copy kept, and a signed copy returned to the Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden, The Netherlands, fax. 31 71 527 2115. After receipt of the signed user agreement, the applicant will be informed of his/her user name and password.
5. See note 1.

References

Biagini, Lisa and Eugenio Picchi. 1996. INTERNET and DBT. Gellerstam, Martin, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Papmehl (Eds.). 1996: 47-53.

Biber, Douglas. 1994. Representativeness in Corpus Design. Zampolli, Antonio, Nicoletta Calzolari and Martha Palmer (Eds.). 1994: 377-407.

Calzolari, Nicoletta, Mona Baker and Johanna G. Kruyt (Eds.). 1996. *Towards a Network of European Reference Corpora, Report of the NERC Consortium Feasibility Study.* Pisa: Giardini Editori e Stampori.

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: 41-61.

Cornelis, Louise H. 1997. Passive and Perspective. Van den Hoven, Paul and Wolfgang Herrlitz (Eds.). 1997: 1-295.

Dutilh, M.W.F. and J.G. Kruyt. 1992. *Feasibility Experiment Design Criteria. Investigation into Text Typological Classification Tools. NERC* Paper WP6-94. Unpublished report. Leiden: Instituut voor Nederlandse Lexicologie.

Gellerstam, Martin, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Papmehl (Eds.). 1996. *Euralex '96 Proceedings* I-II. Göteborg: Department of Swedish, Göteborg University.

Grefenstette, Gregory. 1996. Approximate Linguistics. Kiefer, Ferenc, Gábor Kiss and Júlia Pajzs (Eds.). 1996: 83-96.

Handboek van de Nederlandse pers en publiciteit. Gedrukte media. 1990. Schiedam: Nijgh Periodieken.

Hoeksema, Jack and Henny Klein. 1996. From Comparative to Adverb of Degree; Dutch *even.* Jonkers, Roel, Edith Kaan and Anko Wiegel (Eds.). 1996: 59-70.

Jonkers, Roel, Edith Kaan and Anko Wiegel (Eds.). 1996. *Language and Cognition* 5. Groningen: Centre for Language and Cognition, Department of General Linguistics, University of Groningen.

Kiefer, Ferenc, Gábor Kiss and Júlia Pajzs (Eds.). 1996. *Papers in Computational Lexicography COMPLEX '96.* Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Krishnamurthy, Ramesh. 1996. The Data is the Dictionary: Corpus at the Cutting Edge of Lexicography. Kiefer, Ferenc, Gábor Kiss and Júlia Pajzs (Eds.). 1996: 117-144.

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

Kruyt, J.G. 1995b. Technologies in Computerized Lexicography. *Lexikos* 5: 117-137.

Kruyt, J.G. and E. Putter. 1992. *Corpus Design Criteria. NERC* Paper WP6-129. Unpublished report. Leiden: Instituut voor Nederlandse Lexicologie.

Kruyt, J.G., S.A. Raaijmakers, P.H.J. van der Kamp and R.J. van Strien. 1995. On-line Access to Linguistically Annotated Text Corpora of Dutch via Internet. Rettig, Heike (Ed.). 1995: 173-178.

Kruyt, J.G. and P.G.J. van Sterkenburg. 1996. Corpus Design Criteria. Calzolari, Nicoletta, Mona Baker and Johanna G. Kruyt (Eds.). 1996: 57-72.

Neumann, Robert. 1995. *MLAP93-21 MECOLB Progress Report II.* WP5 Final report Quality Assessment. Unpublished report. Mannheim: Institut für Deutsche Sprache.

Noël Dirk, Bart Defrancq and Filip Devos. 1995. Considering Bilingual Dictionaries Against a Corpus. *Lexikos* 5: 57-81.

Noordman, L.G.M. and W.A.M. de Vroomen (Eds.). 1994. *Informatiewetenschap 1994, Wetenschappelijke Bijdragen aan de Derde StinfoN-conferentie.* Leiden: StinfoN.

Norling-Christensen, O. 1996. *Design and Composition of Reusable Harmonised Written Language Reference Corpora for European Languages.* Unpublished *PAROLE* paper. Copenhagen: Society for Danish and Literature (DSL).

Pollmann, Thijs. 1996. De wet van het uitdovend verleden. Wetenschapsbijlage, *NRC Handelsblad*, 11 January 1996: 1.

Raaijmakers, Stephan and Tilly Dutilh. 1995. Morphosyntactic Corpus Annotation for Dutch. INL Contribution to the *MECOLB* Project. Neumann, Robert. 1995: 10-21.

Rettig, Heike (Ed.). 1995. *Language Resources for Language Technology. Proceedings of the First European TELRI Seminar.* Mannheim: Institut für Deutsche Sprache.

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.* London / Glasgow: Collins ELT.

Sinclair, John and Jackie Ball. 1995. *Text Typology.* Draft document EAGLES Text Corpora Working Group, submitted to EAGLES Workshop Issues in Corpus Linguistics, Madrid 1996.

Sperberg-McQueen, C.M. 1994. The Text Encoding Initiative. Zampolli, Antonio, Nicoletta Calzolari, Martha Palmer (Eds.). 1994: 409-427.

Teubert, Wolfgang. 1995. Language Resources: The Foundations of a Pan-European Information Society. Rettig, Heike (Ed.). 1995: 105-128.

Van den Hoven, Paul and Wolfgang Herrlitz (Eds.). 1997. *Utrecht Studies in Language and Communication* 10. Amsterdam / Atlanta: Rodopi B.V.

Van der Voort van der Kleij, John, 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: 181-194.

Van der Voort van der Kleij, John and Truus Kruyt. 1997. Lexicon for Linguistic Annotation of Dutch Text. *TELRI Newsletter* 5: 32-35.

Vliegen, Maurice. 1996. Verbs of Perceiving and Verbal Communication in Dutch: Clausal Complementation. Gellerstam, Martin, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Papmehl (Eds.). 1996: 309-319.

Zampolli, A. 1996. Corpus Linguistics, the Use of Computers in the Humanities, Computational Linguistics. Calzolari, Nicoletta, Mona Baker and Johanna G. Kruyt (Eds.). 1996: XI-XXXIX.

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.

Zampolli, Antonio, Nicoletta Calzolari, Martha Palmer (Eds.). 1994. *Current Issues in Computational Linguistics: In Honour of Don Walker.* Pisa: Giardini editori e stampatori / Dordrecht: Kluwer Academic Publishers.

Appendix. Composition of INL 38 Million Words Corpus 1996

TOTAL CORPUS				38,164,250
VARIED SUBCORPUS				
TOPIC	**SUBTOPIC**	**MEDIUM**	**PERIOD**	**TOKENS**
MIXED				6,380,508
	—	1 NEWSPAPER	1995	2,607,579
	—	2 WRITTEN-tb-SPOKEN	1991-1995	3,772,929
HEALTH				413,967
	Health	5 BOOKS	1992-1993	242,114
	Psychology	3 BOOKS	1993-1994	171,853
LEISURE				
	Leisure	2 MAGAZINES	1992-1995	1,017,881
HUMANITIES				1,030,033
	Languages	1 BOOK	1992	69,628
		1 MAGAZINE	1991-1995	515,921
	Philosophy	6 BOOKS	1993-1994	444,484
SCIENCE				1,040,131
	Environment	2 MAGAZINES	1989-1995	781,261
		1 BOOK	1989	61,559
	Astronomy	1 MAGAZINE	1992-1993	197,311
SOCIETY				2,874,744
	Politics	1 BOOK	1982	245,482
		2 MAGAZINES	1992-1995	628,026
		1 MISCELLANEOUS	1991	36,491
		1 WRITTEN-tb-SPOKEN	1970-1986, 1988	36,188
		1 REPORTED SPEECH	Nov-Dec 1995	1,869,099
	Social Studies	2 BOOKS	1990-1991	59,458
TOTAL				12,757,264
NEWSPAPER SUBCORPUS				
TOPIC	**SUBTOPIC**	**MEDIUM**	**PERIOD**	**TOKENS**
MIXED				9,127,200
	—	1 NEWSPAPER	1992-1995	
LEISURE				3,305,237
	Sports	1 NEWSPAPER	1992-1995	
TOTAL				12,432,437
LEGAL SUBCORPUS				
TOPIC	**SUBTOPIC**	**MEDIUM**	**PERIOD**	**TOKENS**
SOCIETY				12,974,549
	Law	1 MISCELLANEOUS	1814-1989	
TOTAL				12,974,549