ED 413 750 FL 024 771

AUTHOR	Kruyt, J. G.; van Sterkenburg, P. G. J.
TITLE	A New Dutch Spelling Guide.
PUB DATE	1995-00-00
NOTE	10p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computational Linguistics; Computer Software; *Discourse Analysis; *Dutch; Foreign Countries; *Language Research; Lexicology; Linguistic Theory; *Spelling; Uncommonly Taught Languages
IDENTIFIERS	*Language Corpora; Netherlands

ABSTRACT

This paper describes the development of two new corpus-based Dutch spelling guides using language data gathered by the Institute for Dutch Lexicology, a research institute subsidized by the Dutch and Belgian governments. The guides were produced in 1990 and 1995. The guides are based on two earlier ones, published in 1866 and 1954, but attempt to resolve problems of orthography. The 1990 edition contains the word list of the 1954 guide and about 30,000 new entries in a preferred spelling only, with preferred and "allowed" orthography provided in a separate list. The 1995 guide is an official revision of the 1954 guide, using new, only slightly modified principles for spelling revision. Both of the new guides draw on empirical language data unavailable for the earlier guides. The corpus of 50 million words used for the 1990 edition was limited by the technology available for compiling it; substantial additional text was available for the later version. Analysis of the language data for the guides was hampered by lack of encoding standards and limits in the retrievability of data. Some of the corpora used are now available on the Internet. Contains 10 references. (MSE)

A New Dutch Spelling Guide

J.G. Kruyt, P. G. J. van Sterkenburg

Institute for Dutch Lexicology INL
P.O. Box 9515
NL-2300 RA Leiden
Tel.: +31 71 527 2270
Fax: +31 71 527 2115
E-mail: kruyt@rulxho.leidenuniv.nl

1. Introduction

The *Institute for Dutch Lexicology (INL)* is a research institute subsidized by the Dutch and Belgian governments. Apart from corpus-based lexicography, INL is active in the field of the compilation and (semi-)automatic linguistic annotation of text corpora, and the development of retrieval systems which allow the user to consult the corpora along various parameters. Corpus development at INL dates from the mid-seventies. Up to 1990, the INL text corpora were developed for lexicographical purposes mainly. Presently, they are used for a broad variety of research and applications (cf. Kruyt 1995, Van Sterkenburg & Kruyt 1996).

The language data available at INL have proven to be of interest for external users as well. The majority concerns academic users. The more limited interest in the data from the part of commercial companies may be explained by investments being relatively large and the market restricted, due to the rather small Dutch speaking language area and a minor interest of computational linguists in the Dutch language. Still, INL is presently cooperating with two commercial publishing houses in the framework of product development. A product recently developed in cooperation with *AND Electronic Publishing* (Rotterdam, The Netherlands) is a CD-ROM containing the historical dictionary of the Dutch language *Woordenboek der Nederlandsche Taal (WNT)*. The present paper reports on the development of two corpus-based Dutch spelling guides, in which the INL data have been crucial for both the list of entries and the values (actual word forms) for the information categories per entry. Partners were the publishing house *Staats Drukkerij en Uitgeverij (SDU)* (The Hague, The Netherlands) and, as for the second one, the Dutch-Belgian government body *Nederlandse Taalunie (NTU)* as well. The spelling guides are characterized in Section 2, the relevant INL data in Section 3. The final section discusses the need for more harmonization in the development of and access to data, so as to make product development more efficient and feasible.

Rather than on proper spelling issues, this paper focuses on the conditions in which the spelling products were developed. For detailed information on Dutch spelling, we refer to Molewijk (1992) and De Vries, Willemyns & Burger (1993).

2. Dutch Spelling Guides

The two corpus-based spelling guides compiled by INL date from 1990 and 1995, respectively. The guide of 1990 includes an earlier one, published in

1954. The latter was preceded by the first spelling guide commonly accepted, published in 1866. A concise characterization of the two earlier guides serves as a background for the products of 1990 and 1995.

2.1 Spelling Guide 1866

The *Word List for the Spelling of the Dutch Language* (1866) was compiled by the founders of the historical dictionary of the Dutch language *"Woordenboek der Nederlandsche Taal" (WNT)*, M. de Vries and L.A. te Winkel. In the framework of dictionary design, they devised a spelling system based on etymological principles. In 1882, the Dutch government implicitly acknowledged this spelling system as the official one, by applying it in the criminal code. A slightly simplified version of the spelling system got an explicitly official status in the next century, in 1947.

From the outset, the spelling system was criticized. It was considered too complicated, particularly for the children at primary school, being obliged to correctly apply the spelling system.

2.2 Spelling Guide 1954

In order of the governments of the Netherlands and Belgium, a new spelling guide was compiled by a committee of twelve Dutch and Belgian (Flemish) experts in the field. They developed a more simplified spelling system. The compilers additionally had to bridge the gap between the Dutch and Flemish views on spelling issues. The Flemish traditionally wished to distinguish themselves from the French, and hence, had a preference for the character "k" over "c", e.g., "publikatie" rather than "publicatie" ("publication"). The Dutch on the other hand, did not appreciate German-like spellings at the time (a few years after the Second World War), and they preferred "c" over "k": "publicatie". The problem mainly applies to loan words. As a compromise, the *Word list of the Dutch language*, published in 1954, often lists two spelling forms for a word, a "preferred" one and an "allowed" one.

The double spelling forms evoked much criticism and the spelling was still considered too complicated. Additionally, many new words came into use, which were not in the list. The governments of The Netherlands and Belgium felt responsible for a solution and installed many spelling committees. For political reasons, a decision remained forthcoming until 1994.

2.3 Spelling Guide 1990

Meeting the need for an orthography for many words that entered the Dutch language since the fifties, INL and the *Staats Drukkerij en Uitgeverij (SDU)* published an inofficial spelling guide in 1990, the *Revised word list of the Dutch language*. SDU is a privatized, commercial publishing house. The activities include the publication of books (language, art, history) and state products (passports, bank notes etc.), as well as database publishing. INL was responsible for the contents of the guide, SDU for its publication. The division of the revenues was established by contract. The guide was published in printed form only; a slightly encoded machine-readable version was available for internal use.

The guide of 1990 contains the word list of 1954 (ca. 65 000 entries), and additionally ca. 30 000 new entries. The list contains entries in the "preferred" spelling (2.2.) only; an appendix lists entries in "preferred" and "allowed" orthography. The information categories per entry (microstructure) include the entry (in "preferred" orthography), variant form (if relevant), hyphenation, indication of meaning (in case of homographs), genus for nouns, inflected forms (plural and diminutive for nouns; past and perfect participle for verbs; inflected form, comparative and superlative for adjectives), and reference (if relevant).

Essentially new with respect to the former guides is that the selection of new entries, their orthography, and the values (the actual word forms) for the information categories per entry have an empirical basis. Evidence came from a broad coverage *50 Million Words Corpus* at INL (ca. 1600 sources, mainly 1970-1990). Both entries and word forms for an entry were included in the guide only if they met the criteria of frequency and coverage (distribution over text sources). The guide is principally corpus-based (Van Sterkenburg 1991). This implies that, for example for a particular noun, a diminutive or plural form is not listed if too few occurrences in too few sources were found in the INL corpus.

2.4 Spelling Guide 1995

An official follow-up of the guide of 1954, *Word list of the Dutch language*, is to be published by the end of 1995. In 1994, the Dutch and Flemish Ministers of Education and Culture decided on new but not too radically changing principles for a spelling revision. Among other things, the internal consistency within the spelling system should be improved. The Dutch-Belgian government body, *Nederlandse Taalunie (NTU),* whose task is the promotion of the Dutch language in the broadest sense, requested INL to supply the main entries and to compile the new guide under the authority of NTU; and asked SDU to publish it in both printed and electronic form. Commissions and financial revenues were established by contracts between NTU and INL, and NTU and SDU, respectively. INL and SDU have the right to negotiate together concerning a CD-ROM publication and other spin-offs.

The guide of 1995 contains ca. 110 000 entries: the word list of 1954 from which ca. 15 000 obsolete words have been removed (ca. 50 000), the new words from the list of 1990 (ca. 30 000), and ca. 30 000 new entries. The orthography of all entries is according to the new spelling rules. The distinction between "preferred" and "allowed" spelling (cf. 2.2.) is abolished, with the exception of a few pronunciation variants only. The information categories per entry in the printed guide are essentially the same as those in the guide of 1990 (2.3.). As for the genus for nouns, the original 19 categories (combinations of male, female, and neuter) are reduced to 9 categories, accounting for the increasing tendency to consider female and male as one genus category combined with the article "de", versus neuter as an another, combined with the article "het". The database also contains the obsolete words and some additional fields per entry, among which "spelling 1990", "hyphenation 1990", "morphological category" (free vs. compound vs. derivation), "Flemish" (yes vs. non), "loan word" (yes vs. non), "year of first publication" (1954, 1990, 1995), "obsolete" (yes vs. non), and some administrative fields.

Like the guide of 1990, the guide of 1995 has an empirical basis with respect to the selection and orthography of the entries as well as the values for the information categories per entry. Empirical data were used for, for example, the choice between conflicting orthographies, such as product vs. produkt; **cadeau** vs. **kado**, scène vs. scene, mafia vs. maffia, know-how vs. knowhow, context vs. tekst. Evidence came from the broad coverage *50 Million Words Corpus,* a *5 Million Words Corpus 1994* of recent texts (17 sources, most of them dating from 1989–1994), a *27 Million Words Newspaper Corpus 1995* (1 source, 1994–1995), and other machine-readable sources available at INL[1]. Main criteria were again frequency and coverage. For some cases, the government body NTU decided on a deviating outcome, based on political considerations.

3. Spelling Guides and INL Language Resources

The INL language resources used for the spelling guides have a different status. The *50 Million Words Corpus* was compiled in the eighties when texts in machine-readable form were hardly available. Optical character recognition (OCR) was applied for converting books into electronic form. The ca. 1600 sources, for the most part dating from 1970–1990, cover a broad variety of topics. The retrieval program developed by INL allows searches at the level of word form, and for a subcorpus of 15 million words, at the levels of lemma and part of speech as well. This implies that the retrieval of frequency and coverage data for the most part concerned individual word forms rather than head words with their corresponding inflected forms in sofar as they occur in the texts. This, of course, impeded efficiency.

Since 1992, INL acquires machine-readable texts (books, magazines, newspapers, etc.) from several publishing houses on a contract basis. Due to the use of different systems for text preparation by the publishing houses, the acquired texts have different formats. Some texts are rather clean, others have a dirty proper text (i.e., full of strange characters). The encoding, if present at all, is different with respect to both the number and the types of the encoded categories. The texts were to be converted, filtered for information not relevant to this application (e.g., usage codes), and formally harmonized to some extent, so as to make them appropriate as input for further processing and consultation. The different characteristics of the texts coming from the various publishing houses required the development of specific sets of software for handling the different text formats.

Part of these texts are included in the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995*. The *5 Million Words Corpus 1994* consists of several text types, most of them dating from 1989–1994, and covers a variety of topics. The *27 Million Words Newspaper Corpus 1995* covers one newspaper only, the editions dating from 1994 and 1995. These corpora have automatically been annotated for headword and part of speech, by a lemmatizer/POS-tagger developed at INL (Van der Voort et al. 1994). The retrieval program enables the user to formulate searches at the levels of word form, headword, and part of speech. Frequency and coverage data for the spelling guides could, hence, rather easily be retrieved. In order to get these data for the texts not included in the corpora (ca. 15 million words), the word forms needed still to be lemmatized and the texts to be made accessible for the purpose.

We can say that INL was equipped for the compilation of the corpus-based spelling guides by having appropriate text corpora and operational retrieval systems at the start of the spelling guide projects. Still, considerable efforts were needed so as to make the empirical basis as large and diversified, and hence, as reliable as possible. For more detailed information on the corpus compositions and the retrieval systems, we refer to Kruyt (1995) and Van Sterkenburg & Kruyt (1996).

4. Evaluation and conclusion

Much work and time could have been saved with a higher level of harmonization in several stages of product development. In Section 3, the additional work due to the lack of standards in text preparation by the publishing houses was mentioned. Although there is a tendency towards text encoding according to the *Standard Generalized Markup Language (SGML)* standard, it will take some time before all publishers will have changed their infrastructure.

Another stage is data retrieval. In this respect, the INL textual resources have a different status (Section 3). If various text corpora are accessible on equal linguistic parameters in a uniform way, frequency and coverage data will be retrieved more efficiently. A similar argument applies to other data, of course, particularly in a multilingual environment of comparable corpora of different languages, as envisaged by the EC-funded project PAROLE.

An additional complicating factor in the spelling projects concerned the structure of the electronic spelling guide data. The electronic spelling guide of 1990 was a slightly encoded text file (2.3.). The format of the electronic file of the 1995 guide consists of a number of attribute/value pairs for each entry (e.g., for the entry "demonstratie" (demonstration) among others the pairs: flexcat (attr.): plural (value); flexform (attr): demonstraties (value)). With respect to the guide of 1990, the electronic 1995 guide has an increased number of information categories (2.4). The information of the 1990 guide was to be extracted from the text file and inserted into the new format, which involved automatic identification of specific information, making the information more explicit, and restructuring the electronic data according to the new format. For the printed edition, the publisher SDU for his part requested a selection of information categories in another, deviating format, maintaining the attribute/value pairs but with different names for the information categories and with different delimiters between the pairs. A common data structure, used by both our institute and the project partners, would have saved much programming effort.

In conclusion, we can say that future cooperation can be supported and improved by more uniform standards at the levels of text preparation, data structure, and access to data. This does not alter the fact that the two spelling projects, as well as the earlier mentioned historical dictionary on CD-ROM (cf. 1.), demonstrate that data and facilities originally developed for internal purposes mainly may be a useful basis for product development in cooperation with commercial partners.

Note

INL offers the opportunity to consult the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995* by Internet, up to now for non-commercial research purposes only. In order to get free on-line access to the retrieval program developed for these corpora, 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" or "SEND [27MLN95]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 71 527 2115. After receipt of the signed user agreement, you will be informed about your user name and password. For the conditions for commercial use of INL resources, please contact the director of INL, Prof. dr. P.G.J. van Sterkenburg. If you need additional information, please send an e-mail to "Helpdesk@Rulxho.LeidenUniv.NL".

References

Herziene Woordenlijst van de Nederlandse taal. 1990. s'Gravenhage: SDU uitgeverij.

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

Molewijk, G.C. 1992. Spellingverandering van zin naar onzin (1200-heden). s'Gravenhage: SDU.

Sterkenburg, P.G.J. van. 1991. "Het groene boekje". Bennis, H., A. Neijt and A. van Santen (eds.). De groene spelling, 54–71. Amsterdam: Uitgeverij Bert Bakker.

Sterkenburg, P.G.J. van and J.G. Kruyt. 1996. "Dutch Electronic Corpora: their history, applications and future". Computers and the Humanities (in press).

Voort van der Kleij, J.J. van der, S. Raaijmakers, M. Panhuijsen, M. Meijering and R. van Sterkenburg. 1994. "Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem". Noordman, L.G.M. and W.A.M. de Vroomen (eds.). Informatiewetenschap 1994. Wetenschappelijke bijdragen aan de derde STINFON-conferentie, Tilburg, 181–194.

Vries, J. W., R. Willemyns and P. Burger. 1994. Het verhaal van een taal. Negen eeuwen Nederlands. Amsterdam: Prometheus.

Vries, M. de and L.A. te Winkel. 1866. Woordenlijst voor de spelling der Nederlandsche taal. s'Gravenhage en Leiden: Martinus Nijhoff, A.W. Sijthoff.

Woordenlijst van de Nederlandse taal, samengesteld in opdracht van de Nederlandse en de Belgische regering. 1954. s'Gravenhage: Staatdrukkerij- en uitgeverijbedrijf, Martinus Nijhoff.

Woordenlijst Nederlandse taal. 1995. Den Haag: SDU Uitgevers.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: TELRI - Proceedings of the First European Seminar:"Language Resources for Language Technology", Tihany, Hungary, Sept. 15 and 16, 1995

Author(s): Heike Rettig (Ed.)

Corporate Source:

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title:
Norbert Volz, M.A.
TELRI Project Manager

Organization/Address:
Institut für deutsche Sprache
R 5, 6-13 - 68161 Mannheim
Postfach 101621 - 68016 Mannheim

Telephone: +49 621 1581-437

FAX: +49 621 1581-415

E-Mail Address: volz(at)ids-mannheim.de

Date: 28/11/97

(over)