Text Typology and Selection Criteria for a Balanced Corpus: the Integrated Language Database of 8th-21st-Century Dutch

K.H. van Dalen-Oskam, D.J.G. Geirnaert, J.G. Kruyt
Instituut voor Nederlandse Lexicologie
P.O. Box 9515
2300 RA, Leiden
The Netherlands
karina.van.dalen@niwi.knaw.nl
geirnaert@inl.nl
kruyt@inl.nl

Abstract

The Institute for Dutch Lexicology is compiling the *Integrated Language Database of 8th-21st Century Dutch*, which will consist of three components: a text component, a dictionary component and a lexicon component. In this paper we describe the work done up till now on the text component. This will contain a balanced, diachronic corpus of texts. Paragraph 2 shows the role of the text typology, helping the designers to build a corpus that is a good representation of the Dutch language, and enabling users to make many different subcorpus selections. We describe how the text typology for the *Integrated Language Database* was designed, with the primary aim of the texts as governing principle. Paragraph 3 presents additional classification features that will give the user still more possibilities for subcorpus selection. The selection criteria for individual texts are discussed in paragraph 4, dividing texts into 'originals' and 'editions' and presenting the rules by which to choose from the available originals or editions. Paragraph 5 concludes this paper with a short description of the next empirical step: the building of a prototype.

1 Introduction: the *Integrated Language Database*

Recently, the Institute for Dutch Lexicology started preparations for the compilation of a large integrated language database of 8th-21st-century Dutch, the so-called *Geïntegreerde Taalbank* (*Integrated Language Database*). This is a long term project funded by the Dutch and Flemish governments. Work on the *Integrated Language Database* is still in the conceptual phase [Kruyt 2000] and is executed by a team of two historical lexicographers, three (historical and modern) linguists, two computational linguists and supporting personnel. After a prototype (see 5), parts of the database will become available for consultation over the Internet about every three years.

The *Integrated Language Database* will consist of three components: a text component, a dictionary component and a lexicon component. The dictionary component will consist of an electronic version of the main scholarly dictionaries, describing the Old, Middle and Modern Dutch language periods. The lexicon component will contain other types of lexica, some of them historical lexica containing entries not accounted for in the scholarly dictionaries. Links will be established between data in the various components, for example links between words in the texts and their entries in the dictionaries or historical lexica, and links between quotations in the dictionaries and their occurrences in the source texts.

In this paper we will focus our attention on the text component. One of the purposes of this part of the database is to enable researchers to question a balanced corpus of Dutch texts for lexicographical, lexicological and other linguistic phenomena, and, to a lesser degree, historical, literary, and other cultural information. In order to guarantee a balanced corpus in the text component, we had to design a text typology. We also had to develop ways to maximise the possibilities of subcorpus selection for the users of the *Integrated Language Database*. Furthermore, we had to define selection criteria which could enable a valid choice of the texts and/or text editions for each text type and each time period. We will present an overview of the work we have done up till now.

2 A Balanced Corpus: Developing a Text Typology and Keeping Count

There are two important reasons for using a text typology when building a corpus, one from the point of view of the designers and the other from the point of view of the users. The designers can make use of a text typology in order to build a balanced corpus. Their cataloguing (or labelling) of each text will of course be documented in the database. The users will then be able to select subcorpora on the basis of this information. We will first concentrate on the designer point of view and return to the user in the next section of this paper.

Early on, bibliographic research demonstrated that the already existing diachronic corpora for the Germanic and Latin languages all have a more or less different text typology [Depuydt & Van Dalen-Oskam 2000]. None of them, however, seemed to work for the Dutch corpus we have in mind. Differences in the covered time period (no other corpus covers as many centuries as the Dutch one will do) or differences in the surviving texts and text types from the older periods may have contributed to this situation.

We wanted to follow a primary guideline that would be applicable to texts from every century of the history of written Dutch. We decided to test whether the *primary aim* of a text might serve as that guideline. On the basis of this hypothesis, we designed a text typology. The process was emphatically empirical: we based our ideas on our knowledge of a large number of different texts, and continuously checked the results by looking at still more texts to find out whether these would fit in or not. This resulted in a text typology with two primary aims, which consist of six and three sub-aims respectively.

Primary aims: (A) Texts that have as their primary aim to present a factual representation of knowledge and information; they appeal to the desire for factual knowledge and information - 'Information'; (B) Texts that have as their primary aim to present a creative representation of knowledge and information; they appeal to the desire for creativity and for amusement - 'Imagination'. This corresponds with the subdivision in the *British National Corpus* [http://info.ox.ac.uk/bnc/what/writ_design.html].

In (A), sub-aim (1), 'The presentation of lexicographical information', is in fact an aspect of sub-aim (2). We have given it a separate place because of the focus of our Institute on corpus-based lexicology and lexicography, and because of the fact that lexicographical works constitute a systematically-quoted category of source texts in the scholarly dictionaries of the dictionary component.

sub-aim	text type	subcategory of text type
(1) The presentation of lexicographical information.	Dictionaries and wordlists	-descriptive dictionaries and wordlists -bilingual or multilingual dictionaries and wordlists
(2) The presentation of factual knowledge	texts about subjects pertaining to contemporary affairs	-reports about current affairs -advertisements -texts presenting opinions meaning to influence public opinion or current affairs
	texts about subjects pertaining to arts and science	-general reference books (encyclopaedia) -texts about languages, humanities, arts -texts about science -texts about social sciences
	texts about subjects pertaining to non-scientific subjects	-texts about the practical organisation of society -texts about crafts and practical professions
(3) The presentation of information through administrating, accounting for and regulating affairs concerning businesses and public/political institutions	texts with an official status (from and to corporations, businesses, and government institutions)	-administrative texts -correspondence
(4) The presentation of information about personal and private affairs	'egodocuments'	-addressed to oneself or to an undefined audience -addressed to an addressee
(5) The presentation of information as a guideline for the development of one's philosophy of life or attitude to life (in the spiritual sense)	spiritual and religious texts texts meant to guide the reader to psychological well-being and certain social behaviour	
(6) The presentation of information as a guide or manual for other affairs than spiritual	manuals and instruction booklets	

Table 1: Primary aim (A) 'Information'

sub-aim	text type	subcategory of text type
(1) Texts that are meant to be read (to someone)	narrative texts	
(2) Texts that are meant to be staged	dramatic texts	
(3) Texts that are meant to be heard (in the mind)	lyrical texts	

Table 2: Primary aim (B) 'Imagination'

(A) is subdivided into six groups, (B) into three. These sub-aims (column 1 in both tables) in their turn are represented by different text types (column 2), which are, in table 1, usually divided into further subcategories (column 3). Each cell in the text typology table will be divided into nine subcells, designating the following time periods: (1) until 1300, (2) 1301-1400, (3) 1401-1500, (4) 1501-1600, (5) 1601-1700, (6) 1701-1800, (7) 1801-1900, (8) 1901-2000, (9) 2001 - . We have tested this text typology extensively from the opposite point of view from the one we had while designing it, namely by trying to find texts for every cell and subcell in this table. Although this resulted in a few minor adaptations, the major structure was left unchallenged.

As for the proportions, on the basis of bibliographic research we decided for a ratio of 66.6% 'Informative writings' to 33.3% 'Imaginative writings'. We still have to define the percentages for the sub-aims and the text types.

3 Maximum Flexibility: Subcorpus Selection on the Fourth Level

As is presented above, the text typology implies three hierarchical levels of text cataloguing or labelling: primary aim, sub-aim, and text type (including subcategory of text type). An overview of these labels will be presented to the user and will provide him or her with a lot of possibilities for selecting the exact subcorpus needed for specific research. Subcorpus selection on the basis of time periods is also possible. By giving each text still other labels designating aspects that do not relate to the text typology, we have created a fourth level on which subcorpora can be selected. These additional classification features are: topic or subject of the text, author's gender, region of origin, domicile, etc.

4 Drawing the Line: Criteria for the Selection of Texts and Representations of Texts

For the selection of texts for the *Integrated Language Database* we will follow the text typology and the additional labels described above. Following CES, we define 'text' as 'a piece of human language communication in the broader sense, that one has reason to consider as a whole'. As we described, each cell in the text typology table will be divided in nine subcells, designating time periods. For each subcell we have to make a well-founded choice from the existing texts.

In reality, however, there will be many cases in which we have to choose from different *representations* of a text: a certain version, imprint, edition, etc. We define 'representation of

a text' as follows: 'a material transcription of a text. It can be a paper printing, an electronic form, (...) etc.' [CES]. When it comes to the medieval period, it can be a hand-written version of a certain text (and a second hand-written version or copy is another representation of the same text). The consistency of the *Integrated Language Database* also depends on well-founded and clear selection criteria for these aspects. In our view, the 'representations of a text' can be divided into two groups: 'editions' and 'originals'. An *edition* is a representation of a text that is constituted through scholarly research by an editor; an *original* is a representation of a text that has been published without a scholarly intermediary.

We want to make well-founded decisions between the following representations of a text: (1) more than one original; (2) one or more originals and one or more editions; (3) more than one edition. Where more than one original is available, we have decided to choose the oldest (authorised) one. We want to facilitate diachronic and synchronic language research in the *Integrated Language Database.* A consistent choice of the oldest original will guarantee a synchronic homogeneous language situation in a (representation of a) text, so that this text may safely be compared with texts of e.g. 40 years later. Therefore we will choose the first edition of a novel, and not the last (or any other) one the author revised.

When we have to choose between an original (or more than one) and an edition (or more than one), we will consistently choose the original (and the oldest one).

When we are dealing with editions only, as is the case for most medieval texts, the problems accumulate: we want the *best* edition. But how can we ascertain which one is the best? For a well-founded choice we have drawn up a list of thirteen different features a text edition can have, and we have validated all the values of these features with a mark ranging from 1 to 5, symbolising the range between 'very bad, very undesirable' and 'very good, very desirable' especially as regards linguistic, and in particular lexicological, research. One of these features is the representation of *u*, *v* and *w*. A lot of editors have normalised the medieval spelling of *u*, *v* and *w*; others have chosen to edit exactly what the manuscript reads. So this feature has two values. However, in the Dutch language it is still not clear how these graphemes have developed through the centuries. That is why we prefer editions that exactly represent the graphemes in the manuscripts, in order to facilitate future research of this linguistic aspect. That is why we assigned the first value 2 points, and the second one 5 points, the maximum. This system of 'weighing' text editions will be used primarily for comparing different editions of the same text, or editions of several texts that could be placed in the same subcell of our text typology. The edition with the highest score (the points for all features added) is the most likely candidate for inclusion in the text component of the *Integrated Language Database.* We have also defined which type of editions can be disqualified in advance and do not have to be weighed: re-spelled or unscholarly transliterated (older) texts, and text editions that are commonly discarded as unscholarly or unreliable by specialists in that field.

5 Conclusion: the Next Step

We have demonstrated how work on the text component of the *Integrated Language Database of 8th-21st-century Dutch* is progressing along clearly empirical lines. We designed the text typology taking into account as many texts as possible, and tested it by checking whether we could find texts for all cells and subcells in the text typology. We followed the same path in developing the criteria for individual text selection. We are currently working on the next step: the compilation of a prototype of the *Integrated Language Database*. This will contain a clearly defined part (in most cases about 5 pages) of about 200 texts, one for each subcell in category A, 'Information', and three for each subcell in B, 'Imagination', resulting in about 66.6% 'Information' and 33.3% 'Imagination'. This prototype will be used to demonstrate the possibilities of the *Integrated Language Database* as a whole to future users, among others. Furthermore, it is used as another test of the selection criteria and as a way to test the process of preparing the texts for incorporation: the scanning and keyboarding of texts, or the adaptation of digital texts, TEI-encoding [Depuydt & Dutilh 2002], adding PoS-tags and a Modern-Dutch lemma, etc. By carefully monitoring this next step and continually learning from our experiences, we will be able to take on the 'real thing' on a sound empirical and scholarly basis.

Acknowledgements

This paper has profited greatly from previous discussions with K.A.C. Depuydt (contact person; email depuydt@inl.nl), J. de Does, M.W.F. Dutilh-Ruitenberg, and J.J.W. van der Voort van der Kleij.

References

[British National Corpus] http://info.ox.ac.uk/bnc/

[CES] (Corpus Encoding Standard) http://www.lpl.univ-aix.fr/projects/multext/CES/CES1-1.html

[Depuydt & Van Dalen-Oskam 2000] Depuydt, K.A.C. & K.H. van Dalen-Oskam, 2000. Oude teksten online. Een overzicht van internetcorpora met historische teksten, on: http://www.inl.nl/pub/intweb.htm

[Depuydt & Dutilh 2002] Depuydt, K.A.C. & M.W.F. Dutilh, 2002. TEI-encoding for the Integrated Language Database of 8-th-21st-Century Dutch, in: *Proceedings EURALEX 2002.*

[Kruyt 2000] Kruyt, J.G., 2000. Towards the Integrated Language Database of 8th-21st Century Dutch, in: *Revue française de linguistique appliquée* 2000, V-2, 33-44.