

The Integrated Language Database of 8th – 21st-Century Dutch

J. G. Kruyt

Institute for Dutch Lexicology INL
NL-2300 RA Leiden, The Netherlands
[Kruyt@INL.NL]

Abstract

The Institute for Dutch Lexicology (INL) has a long-standing tradition in corpus-based lexicography. The results include electronic scholarly dictionaries of Dutch covering the vocabulary from 1200 up to 1976, linguistically annotated electronic text corpora of historical and present-day Dutch, and computational lexica. Added value to these data is given in an on-going long-term INL project, the Integrated Language Database of 8th–21st-Century Dutch (ILD). The aim is to create a flexible linguistic research instrument by linking the dictionaries, a balanced diachronic text corpus and lexica of historical and present-day Dutch. We will link part of our data with data collections stored at other institutes, creating a supra-institutional research instrument. The paper reports on the overall ILD design and the user's perspective. Focus is on the ILD prototype which, when finished, will function as a demonstration model to verify and assess user needs. It now functions to test the design empirically for its applicability to 'real data', as well as to obtain figures on workload, etc. The conclusion is that the latter function proved the prototype to be an indispensable pilot for the ILD.

1. Introduction

The Institute for Dutch Lexicology (INL) has a long-standing tradition in corpus-based lexicography. As a result, the INL now has electronic scholarly dictionaries of the Dutch language covering the vocabulary from 1200 up to 1976, and text corpora covering mainly (Early) Middle Dutch and present-day Dutch. In a European context (EC funding), the Dutch PAROLE corpus and the Dutch PAROLE/SIMPLE lexicon were developed (see www.inl.nl/eng/europe/projects.htm).

Three linguistically annotated corpora of present-day Dutch have been widely used for various research purposes in the fields of linguistics and social studies, for lexicography and lexicon building, for academic teaching, and for the delivery of customized data, since they became Internet-accessible in 1994 (Kruyt, 1998). The Dutch PAROLE corpus will soon be accessible for similar purposes (Van der Kamp & Kruyt, 2004). The follow-up is a bi-national, long-term INL project: the Integrated Language Database of 8th–21st-Century Dutch (ILD). The aim is to provide a flexible instrument for a wide range of synchronic and diachronic research into the Dutch language (and culture) throughout the centuries. For the purpose of flexible retrieval and navigation, various data types within the ILD will be linked. We will also link part of our data with data stored at other centres, creating a supra-institutional research instrument. See for projects with common features: Gellerstam, Cederholm & Rasmark (2000), Fournier (2001), Ruus (2002).

This paper reports on the overall ILD design (2). Then the user's perspective is considered (3). Focus is on the ILD prototype, which will function as a demonstration model to verify and assess user needs. In the current phase of the project, it functions to test the design empirically for its applicability to 'real data', to develop efficient procedures, and to obtain figures on workload for future plannings (4). The conclusion is that the latter function proved the prototype to be an indispensable pilot for the ILD (5).

2. The Overall ILD Design

2.1 General design principles

The ILD will have two dimensions. One is the diachronic dimension; data cover 8th- to 21st-century Dutch. The

other is the linguistic dimension; for each time period, various types of linguistic data are available: encoded dictionary data, linguistically annotated texts, and lexicon data. The diachronic dimension calls for 'general' design criteria that are applicable to different types of sources dating from different centuries; reliable diachronic research would otherwise be virtually impossible. We also distinguish 'criteria specific for a particular period or text type' if they have too much scholarly importance to be ignored. General criteria prevail over text- or period-specific criteria, in the sense that specific criteria are subordinate derivatives of the general ones. Another principle is that international standards are applied where possible. A sound and well-considered design is important for the quality of the ILD as a research instrument.

2.2 Contents

The ILD will consist of three mutually linked components: a dictionary component, a balanced diachronic text corpus component, and a component with lexicons of historical and present-day Dutch.

The dictionary component will comprise the Dictionary of Early Middle Dutch VMNW (4 printed volumes), the Dictionary of Middle Dutch MNW (10 volumes) and the Dictionary of the Dutch Language WNT (43 volumes), and in the longer term the dictionaries of Old Dutch and present-day Dutch (on-going INL projects). These dictionaries are the most comprehensive dictionaries of the Dutch language, compiled according to scholarly principles, and eventually covering the Dutch vocabulary from the 8th up to the 21st century. For these reasons, they are considered a separate component of the ILD (along with some smaller supplementary dictionaries). They are available in machine-readable form, albeit with a different extent of encoding.

The diachronic text corpus should support a wide range of user needs (cf. 1). It will therefore cover many varieties of Dutch written language, dating from the 8th–21st century. As no existing corpus design turned out to be applicable to texts from so many centuries, we developed a new one (Van Dalen-Oskam, Geirnaert & Kruyt, 2002), which, after several empirical tests, has been applied in the prototype (4). The leading principle is 'the primary aim of a text', with two major divisions that more or less

correspond with fiction and non-fiction¹: a creative ('imagination') vs. a factual ('information') representation of knowledge and information. Within this framework, 22 text types have been distinguished. Criteria for text selection have also been developed. Apart from other texts, texts quoted in the dictionaries of the dictionary component will be selected so as to be able to link corpus and dictionary data (2.4). For the acquisition of digital texts, see section 4.

The lexicon component will consist of a rather restricted, well-considered selection of present-day and historical lexica. The main criterion for the selection of lexica is that they provide information that is not at all or much less elaborated in the dictionaries of the dictionary component. For present-day Dutch, the PAROLE/SIMPLE lexicon, although period-specific, is relevant. Historical lexica will contain headwords (and their paradigmatic word forms) found in dictionary quotations or in corpus texts, but not covered as an entry by the dictionaries in the dictionary component. Such a lexicon not only fills a gap in the lexicographical description of Dutch, but is also needed for the annotation of historical Dutch texts (cf. 4.4).

2.3 Annotation

For linking and retrieval purposes, the data will be encoded according to the TEI standard.

The dictionaries in the dictionary component will be encoded for the major information types within the entries (headword, etymology, quotation, meaning description, etc.). For the dictionaries MNW and WNT, this requires a substantial extension of the current encoding and an improvement of the dictionary files, due to inconsistency and lexicographical practices (cf. Kruyt & Van der Voort van der Kleij, 1992-93). Furthermore, present-day Dutch headwords are added to the (historical) headwords of all dictionaries for easy retrieval (3) and linking (2.4); this work have been finished for the VMNW dictionary and for about 90,000 headwords of the WNT.

The texts in the diachronic text corpus will be encoded at several levels. At text level, the text type and other metadata (still to be specified) will be encoded, as parameters for the selection of a user-defined subcorpus. Within the texts, the text structure, the typography and some other textual elements will receive basic encoding geared to retrieval purposes. The design is ready (Depuydt & Dutilh, 2002) and is now tested with prototype texts (4.3). We have adopted a 'database view' on text, which implies, among other things, a clear distinction between the actual text and its medium (such as manuscript, printed book, electronic file); see further section 4.

Work on how we should tag the words ('tokens') of the texts for Part of Speech (PoS) from a diachronic perspective is in progress. In a first, maximal approach, we used a slightly different version of the Dutch EAGLES/PAROLE tagset and we manually tagged the tokens of three historical prototype texts from different periods with both a 'lexical' and a 'functional' tag when applicable (cf. Dutilh & Kruyt, 2002). Decorte, Dutilh-Ruitenbergh & Kruyt (2004), however, conclude that this approach is not feasible, mainly due to lack of consensus among linguists about how to handle linguistic

phenomena such as transcategorisation, lexicalisation and grammaticalisation. See section 4.4 for the follow-up. Apart from PoS, all tokens will be lemmatized with a present-day Dutch headword, or an etymologically reconstructed one when there is no modern equivalent. Lexica need no annotation, as all information is explicit and unambiguous.

2.4 Linking

For user-friendly navigation, links will be established between data within a source and between data of different sources, including external sources. The linking functionality implies that a mouse-click leads the user from a particular point in a query result to related data elsewhere, within or outside the ILD. We will implement direct and indirect links, the latter offering the user several destinations to choose from. Links foreseen include a link from a dictionary entry to its corresponding entry in another dictionary (through the present-day headwords; 2.3); from a dictionary quotation to its equivalent in the original text in the corpus component (for more context); from a corpus word to corresponding entries in the dictionary component and, vice versa, from a dictionary headword to corpus instances (through the present-day headwords); from a corpus text to metadata; from an arbitrary word in the ILD to other occurrences in the ILD, or to the same word stored at an external centre.

In the longer term, different dictionary headwords (also with different PoS) will be linkable at word sense level by using the SIMPLE lexicon and its ontology with semantic types and qualia roles (Pustejovsky, 1998).

3. The User's Perspective

The ILD data will be accessible by means of a retrieval system that will offer its users many more facilities than our present corpora (cf. 1), due to the various data types and the diachronic dimension within the ILD, and due to more advanced means for retrieval and navigation. An information-technological concept is developed by our EDP department. The PAROLE interface (Van der Kamp & Kruyt, 2004) functions as a model for the corpus component. In the EC-funded ELAN project, we participated in building a prototype retrieval system with access to geographically distributed data through one user interface.

To be geared to a broad user group, historical data will be accessible by use of a present-day Dutch headword (cf. 2.3). Etymologically reconstructed headwords will be presented to the user together with morphologically or semantically related modern Dutch headwords (e.g. reconstructed 'aanvaardigen' with modern 'aanvaarden'). Of course, specialists in historical Dutch can have access via historical forms as well.

Provided that the data are sufficiently annotated and linked, such a retrieval system will offer users many research facilities. Here follow some examples. A researcher who is interested in the history of words may ask the system: for the present-day Dutch word X, give me the corresponding headwords with their form variants and etymology sections from the dictionaries WNT, MNW and VMNW. A researcher can ask for more usages of a headword in the corpus texts if the quotations in the dictionaries are not satisfying. Someone interested in spelling may ask: list all variant forms of the headword Y

¹ Especially for old texts, the distinction between fiction and non-fiction cannot be drawn sharply.

with their text source and geographical location. A researcher interested in loan words may ask: list loan words from French attested in the dictionary WNT and in 18th-century narrative texts. And if relations from the SIMPLE lexicon can be used, a researcher interested in the vocabulary of the Industrial Revolution may ask: find words with word senses belonging to the semantic class of 'instrument' attested in 19th-century texts about science. If specific information is not available in the ILD, the researcher can navigate to an external database. The list of potential research options that make use of the annotated and linked data is virtually endless.

That is still in the future. The first step now is the ILD prototype (4), a demonstration model for future users to verify and assess user needs. In the prototype, the major retrieval and linking functionalities will be implemented, including links from some French-Flemish dialect headwords in the VMNW and MNW dictionaries to a dialect centre in Belgium. Not only the User Committee and Advisory Board connected to our project, but also our present corpus users will be requested to give feedback.

After the prototype, we will develop and make available subsystems of the ILD as intermediate products, rather than wait until the complete ILD has been realized. We will investigate whether it is feasible to give users access already in the development phase.

4. The ILD Prototype

4.1 Introduction

When the corpus design was ready, we started building an ILD prototype, a small-scale model of the contents and the retrieval functionalities of the ILD (cf. 2.4; 3). In the current phase of the project, it is used to test empirically the soundness and applicability of the conceptual ideas, to develop efficient procedures, and to measure workloads in view of future plans for intermediate products. These functions have turned out to be extremely useful, as particularly historical texts and their medium (cf. 2.3) have many unforeseen characteristics requiring solutions. We started with the prototype corpus component. Here follows a description of the results so far.

4.2 Text Selection and Acquisition

In principle, 224 text fragments of about 5 pages (carefully selected from front, body and back) were planned and selected according to the corpus design, covering the 8th-20th century represented by 8 periods. The proportion is 33% 'imagination' and 66% 'information' (cf. 2.2). In 31 cases, suitable texts could not (yet) be found, almost all of them for the period before the 15th century, due to the general problem that only few old texts have survived. 43 texts were acquired from digital repositories elsewhere, 150 text fragments were digitized by in-house scanning and correction (see also below). For text editions, we applied the criteria for measuring the editorial quality (Van Dalen-Oskam, Geirnaert & Kruyt, 2002), in order to choose the best one if more than one was available. In the meantime, we weigh all text editions, a report of which will be shown to the ILD user as a concise representation of the editor's transcription method. For the external text files, we developed a similar report indicating the quality of each file in terms of format, errors, type of encoding, etc.

For all texts, bibliographic and other metadata are now stored in an Access database; for the ILD, we foresee a more sophisticated database.

We started with instructions for digitizing that were aimed at a rather detailed representation of textual characteristics. This was common practice in many other projects using TEI and, due to organisational factors, we did not yet know at the time which degree of detail would be necessary for our TEI encoding of structure and typography. The experience we gained from digitizing and encoding so many historical texts, with so many unexpected peculiarities, has changed our view on future digitizing for the ILD, which will be less detailed (in line with our current database view; 2.3), and focused on the actual text rather than on the characteristics of the text medium, such as certain decorative features. Furthermore, due to the knowledge of TEI encoding acquired through the prototype, it will become possible, to a large extent, to merge the processes of digitizing and TEI encoding. This will lead to a much more efficient procedure in the future.

4.3 Encoding of Text Structure and Typography

There are two major issues relevant to the encoding of text structure and typography: our database view on text (2.3) and the notion of what we consider 'the text to be encoded'.

The database view implies that we will abstract from the original typography and font, and display equal structural text elements in a uniform rendering on screen. We still need to define which rendering we will use. Due to the detailed method of digitizing (4.2), we will have to remove the encoding that has become superfluous according to the database view.

As for the notion of 'the text to be encoded', we give priority to the original text selected according to the corpus design, irrespective of its publication in a text edition or as part of a larger entity (an anthology, for example). Consequently, when applicable, the text is isolated from the text around it and the encoding does not account for the place of the text in the overall structure of the complete publication (whether a comprehensive printed work or an electronic file). We only retain the editor's transcription method and the editorial notes, which offer essential information to the user. As a practical consequence, we do not need to digitize more text than intended for our purpose.

We nearly finished the TEI encoding of the 150 in-house digitized prototype text fragments. After some automatic conversion and validation procedures, 'pre-TEI' tagged XML files of the texts have been encoded manually with the aid of a purpose-built editorial tool. So far, the design has only needed some minor adaptations, though some issues are still to be decided on. For example, we consider extending the form-based type specification of particular div's (e.g. letters), in view of a refined retrieval or subcorpus selection. The application of the design presented us with three major practical problems, as 'real' texts show much more variety than TEI accounts for. One was that TEI sometimes does not provide satisfying solutions, resulting in rather contrived encoding. The second was the choice of a suitable TEI tag when a structural text element approaches more than one TEI definition. The third was the development of criteria for consistent and transparent solutions when more than one solution is TEI-acceptable.

As for the files derived from external repositories, we investigated their characteristics and differences with the in-

house digitized files, and we will soon start to encode them, adhering to the principle that all files will receive basic encoding according to the design, if feasible.

4.4 PoS Tagging and Lemmatizing

After our first experience with PoS-tagging (2.3), we elaborated a more modest approach, starting from a reduced tagset and applying a lexical tag method only. We will investigate to what extent we can compensate for the less refined tagging by offering predefined complex queries in the interface, which can be customized by the user, i.e. a functionality similar to the one called 'patterns' in the PAROLE corpus retrieval system (Van der Kamp & Kruyt, 2004). We need a substantial amount of data to be able to define such patterns. As PoS tagging and lemmatization are related issues, we recently started tagging and lemmatizing prototype text fragments from all periods. A tool was built to make this manual work as efficient as possible. Our approach to tagging and lemmatizing probably needs to be customized gradually, depending on the empirical results. The outcome will be a linguistically annotated prototype corpus and a prototype historical lexicon. As a separate activity, we are currently developing a historical lexicon of Middle Dutch by automatically matching the MNW headwords with their paradigmatic word forms attested in the quotations. In spite of all spelling variants, our program matches over 92% for the alphabetic sections A to K.

5. Conclusion

Rather than following the 'opportunistic' approach, we aim at a scholarly sound design in view of its importance for the quality of the ILD as a research instrument. The design should also guarantee the homogeneity of the complete ILD, which will be composed of joined intermediate subsystems (cf. 3).

All design aspects were thoroughly discussed with experts from our User Committee and Advisory Board. It still appears to be hardly possible to predict the variable characteristics and complexity of 'real data', which require sound solutions. The prototype, originally intended primarily as a small-scale demonstration model for the user, has also proven to be an indispensable pilot for the ILD.

Acknowledgements

The following colleagues contributed to the linguistic modelling of the ILD and the ILD prototype so far (colleagues who are no longer in the department are marked with an asterisk): M. van Bennekom, K. van Dalen-Oskam*, S. Decorte, K. Depuydt, J. de Does, T. Dutilh-Ruitenbergh, D. Geirnaert, W. de Ruyter, M. Veltkamp*, B. Verheij, J. van der Voort van der Kleij.

References

Decorte, S., Dutilh-Ruitenbergh, T. & Kruyt, T. (2004). Language change and linguistic annotation in the Integrated Language Database of 8th-to21st-Century Dutch. To appear in Proceedings 2. Freiburger Arbeitstagung zur Romanistischen Korpuslinguistik: Korpuslinguistik und Historische Sprachwissenschaft. Sektion A. Korpusprojekte, Sprachdatenverwaltung

- und Analysewerkzeuge. In a prefinal version at http://www.inl.nl/eng/pub/decorte_freiburg_eng.pdf
- Depuydt, K. & Dutilh-Ruitenbergh, T. (2002). TEI encoding for the Integrated Language Database of 8th–21st-Century Dutch. In Anna Braasch & Claus Povlsen (Eds.), Proceedings of the Tenth Euralex International Congress, Euralex 2002 (pp. 683--688). Also at <http://www.inl.nl/eng/pub/tei.htm>
- Dutilh, T. & Kruyt, T. (2002). Implementation and Evaluation of PAROLE PoS in a National Context. In Manuel González Rodríguez & Carmen Paz Suarez Araujo (Eds.), Proceedings of the third International Conference on Language Resources and Evaluation (pp. 1615--1621). Paris: ELRA. Also at <http://www.inl.nl/eng/pub/lrec2002.pdf>
- Fournier, J. (2001). New directions in Middle High German Lexicography: Dictionaries Interlinked Electronically. *Literary and Linguistic Computing*, 16(1), 99--111.
- Gellerstam, M., Cederholm, Y. & Rasmak, T. (2000). The bank of Swedish. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S. and Stainhaouer, G. (Eds.), Proceedings Second International Conference on Language Resources & Evaluation (pp. 329--333). Paris: ELRA.
- Kruyt, J.G. & Van der Voort van der Kleij, J.J. (1992-93). Towards a computerized historical dictionary of Dutch. *Acta Linguistica Hungarica* 41(1-4), 159--174. Budapest: Hungarian Academy of Sciences.
- Kruyt, J.G. (1998). Dutch written language resources, their users and uses. In A. Rubio, N. Gallardo, R. Castro and A. Tejada (Eds.), Proceedings First International Conference on Language Resources & Evaluation (pp. 959--963). Paris: ELRA. Also at www.inl.nl/eng/pub/grancon.htm
- Pijnenburg, W.J.J., Van Dalen-Oskam, K.H., Depuydt, K.A.C. & Schoonheim, T.H. (2000). *Vroegmiddelnederlands Woordenboek. Woordenboek van het Nederlands van de dertiende eeuw in hoofdzaak op basis van het Corpus-Gysseling*. Leiden: Instituut voor Nederlandse Lexicologie.
- Pustejovsky, J. (1998) *The generative lexicon*. Cambridge, Massachusetts: The MIT Press.
- Ruus, H. (2002). A Corpus-based Electronic Dictionary for (Re)search. In Anna Braasch & Claus Povlsen (Eds.), Proceedings of the Tenth Euralex International Congress, Euralex 2002 (pp. 175--185).
- Van Dalen-Oskam, K., Geirnaert, D. & Kruyt, T. (2002). Text Typology and Selection Criteria for a balanced Corpus: The Integrated Language Database of 8th–21st-century Dutch. In Anna Braasch & Claus Povlsen (Eds.), Proceedings of the Tenth Euralex International Congress Euralex 2002 (pp. 401--406). Also at http://www.inl.nl/eng/pub/euralex_dalen_eng.htm
- Van der Kamp, P. & Kruyt, T. (2004). Putting the Dutch PAROLE Corpus to Work. In Proceedings LREC 2004.
- Verwijs, E. and Verdam, J. (1885-1929). *Middelnederlandsch Woordenboek*. 's-Gravenhage: Martinus Nijhoff.
- Woordenboek der Nederlandsche Taal* (1882-1998). Leiden: Martinus Nijhoff. 's Gravenhage: SDU.