TEI-encoding for the Integrated Language Database of 8th – 21st-Century Dutch

Katrien Depuydt, Tilly Dutilh-Ruitenberg
Instituut voor Nederlandse Lexicologie
P.O. Box 9515
NL-2300 RA, Leiden
The Netherlands.
depuydt@inl.nl - dutilh@inl.nl

Abstract

A major project of the Institute for Dutch Lexicology is the Integrated Language Database of 8th-21st-Century Dutch (henceforth ILD). The ILD will consist of three components: a dictionary corpus component, a lexicon corpus component and a text corpus component. Its object is to facilitate synchronic and diachronic research on different aspects of the Dutch language and culture, and, in particular, to facilitate lexicological and lexicographical research. The three corpora will be interlinked and be made accessible by means of a retrieval system. To guarantee optimal retrieval facilities, extensive encoding is required of the entries in both the dictionary and the lexicon corpus. In the text corpus, each text will have to be fully tagged with PoS and lemma. The texts will also need encoding of the text structure, the typography and some other textual elements. We will discuss the ILD encoding proposal for the text structure (3) and the typography (4) of the text corpus component. Our focus is on the guiding principles that have determined this proposal (2, 3.1 and 4.1).

1 Introduction

A major project of the Institute for Dutch Lexicology is the Integrated Language Database of 8th-21st-Century Dutch [Kruyt 2000]. The Integrated Language Database (henceforth: ILD) will consist of three components: a dictionary corpus component with the major comprehensive dictionaries of the Dutch language (the Dictionary of the Dutch Language *WNT*, the Dictionary of Early Middle Dutch *VMNW* and the Dictionary of Middle Dutch *MNW*), a lexicon corpus component containing lexica of historical and present-day Dutch, and a text corpus component with carefully selected texts, the majority of which is used as source material for the above-mentioned dictionaries [Van Dalen et al. 2002]. The object of the ILD is to facilitate synchronic and diachronic research on different aspects of the Dutch language and culture, and, in particular, to facilitate lexicological and lexicographical research. The three corpora will be interlinked and will be made accessible by means of a retrieval system. To guarantee optimal retrieval facilities, extensive encoding is required of the entries in both the dictionary and the lexicon corpus. In the text corpus, each text will have to be fully tagged with PoS and lemma. The texts will also need encoding of the text structure, the typography and some other textual elements. We will discuss the ILD encoding proposal for the text structure (3) and the typography (4) of the text corpus component. Our focus is on the guiding principles that have determined this proposal (2, 3.1 and 4.1).

2 Basic TEI-encoding of the Integrated Language Database

To encode the ILD we have chosen to use TEI. Apart from being an international standard, it has proven to be very suitable for encoding historical material, the major part of the ILD.

The TEI guidelines provide encoding instructions for many text types and text forms, including dictionaries and lexicons. Also important to us is the 'controlled flexibility' of the TEI encoding system, since it has specific guidelines for making additions or changes. The encoding possibilities in the TEI-guidelines, however, are too extensive to be fully applied to every single text in the ILD. We have therefore decided to determine a subset for each major component. Since feasibility is an important factor, the encoding should be minimal but with maximum efficacy. This can only be accomplished by focusing on the function of the encoding. The encoding should enable a user to select a subcorpus, to define a particular search domain, to search a particular text unit (e.g. the title of a chapter) or information located within a TEI-encoded search domain, as well as to easily switch between lexicographical units and texts. For the encoding of large text corpora, there is already a standard available, the Corpus Encoding Standard [CES]. This standard was not an option, since it was designed for corpora used as a resource in language engineering, which is not the main purpose of the ILD.

3 ILD-encoding of the structure of a text

3.1 Guiding principles

What could be the function of the encoding of the structure of a text? When a user wants to query a text corpus, he will first determine the type of information he is looking for. He will further decide whether or not to use a subcorpus. He might also want to determine a particular context he wants to query. For example: in a subcorpus of newspapers, a user may decide to look for the headings of articles of a particular year. Or, in a subcorpus of 13th-century spiritual writings, he may want to search the prologues for the word 'God'. Without the presence of structural encoding, these types of queries could not be answered easily.

By the structure of a text, we mean the logical structure of a text, i.e. chapters, paragraphs, etc., as opposed to the physical structure of a text, by which we mean the pages, columns, etc. of the medium in which the text was originally printed. The ILD encoding proposal for the structure of a text has to meet the following requirements:
- There should be a minimum of manual encoding.
- A minimum level of encoding should be guaranteed for every text in the text corpus component.
- The minimum level includes the TEI-tags that, according to the guidelines, are 'required' or 'mandatory when applicable'.

The text material is so diverse in text type and time-period that we are forced to split the encoding into two levels: level-one encoding, which is general and applies to every text from every period of the text corpus and level-two encoding (cf. 3.4), which is specific for a certain text type or for a certain period.

The ILD encoding proposal is primarily based on the TEI's default text structure, the base tagset for prose, the core tags, the base tagsets for poetry and drama and the additional tagset for encoding printed dictionaries (Guidelines chapter 12). Thorough analysis of the front and back of 18 texts and 25 dictionaries was equally contributive. Other TEI-projects (CELT, Etext in Virginia and in the Netherlands ETCL and DBNL) have also been consulted.

3.2 Level-one encoding scheme for text structure

In the encoding scheme below (in this paper an excerpt), a distinction is made between tags on division-level (div-level) and tags on lower level (paragraph level). The tags are first presented for the front and back matter, and then for the body. The tags for the body are presented per text form.

Front and back (all text forms):

Div-level: Numbered div's: div1-div7. Type values: titlepage, poem, letter, list, table and for editions: 'editorial matter' *For drama (additionally)*: <prologue>, <epilogue>, <performance>, <castList> (<castItem>, <role>) Paragraph level: See the lower level encoding of the body of the different text forms.

Body:

Prose:	**Poetry**
Div-level: Numbered div's: div1-div7 Type values: volume, part, chapter Paragraph level: <head> <epigraph> <argument> <byline> <opener> <p> <table> <list> <figure> (<head>; <p>; <text>) <closer>	Div-level: Numbered div's: div1-div7 Type values: volume, book, canto, poem. Paragraph level: <head> <epigraph> <lg> <l> <figure> (<head>; <p>; <text>)

Table 1: excerpt of the encoding scheme

We have decided to work with numbered div's, of which a particular selection will be specified by the type attribute. In the front and the back, there is a large variety in content and form of the divisions. The above-mentioned analysis of the front and back of several texts and dictionaries indicated that content-based type-specification is not feasible, particularly when it should be applied to the entire text corpus material. Form-based type-specification, on the contrary, can be determined fairly unambiguously. Therefore, only the div's with a form different from default prose will be specified. Form-based type-specification is also very practical for automatic tagging of the paragraph-level structure. The encoding of drama forms an exception, since the TEI prescribes some specific, required content-based tags. Since not all divisions will be specified for type, the encoding of the title of a div (<head>) in the ILD-tagset is 'mandatory when applicable'. The retrieval system will

allow a search on in combination with <head>, so that a user can get some insight into the contents of a div through the title. Titlepages will be identified as such (div*n* type=titlepage) but will not receive further encoding, since the necessary information (title, author, etc.) will be available in the header. Lists and tables will also be identified as such. We have not decided yet whether we will apply further encoding. For poetry we will tag lines without giving explicit information as to whether the line is complete or not. That could only result from content-based analysis. In lexicographical sources in the text corpus component, only the entries will be tagged. We have opted for the flexible element <entryFree>, because of the variety of form of the entries in the different lexicographical sources of the ILD. A lot of texts are a mixture of different text forms. A speaker in a drama, for instance, can recite a poem, a character in a novel can do the same. In such cases, the poem will be encoded as a simple quotation: <q><l></l><l></l> ... <q>.

3.3 Special cases

Medieval 'rhyming prose', e.g. *Der naturen bloeme* by Jacob van Maerlant or *Sente lutgart* (a hagiography) – both 13th-century texts – ,will be encoded as if it were poetry, because of its verse structure. This will facilitate possible research on rhyme. It is also a way to limit the search domain. The logical structure of the hagiography of the *Sente lutgart*, for instance, is two books, divided into chapters without any paragraph structure. Encoding the text as prose would mean having a complete chapter as a search domain instead of a single line.

Editorial matter will also receive special treatment. For the historical texts, we will often have to use text editons. These editions contain the source text and a lot of editorial matter in the front, in the back and in the notes. For the ILD, the editorial matter is of minor importance. Only editorial notes containing source text material will be encoded. Other editorial matter will be separated from the source text by putting it into one div (div type='editorial matter') without further encoding of the lower-level structure. The editor's system of transcription, however, will be made retrievable for our users.

3.4 Level-two encoding

Level-one encoding applies to all texts and the proposed encoding is either required or mandatory when applicable. Any further encoding belongs to level two, i.e. encoding that is specific for a certain period or text type. It is, for instance, feasible to tag sentence units (<s>) in present-day Dutch texts, using the punctuation. Punctuation in medieval texts, on the other hand, has a totally different function. Therefore, the encoding of sentences would have to be done manually. This is not feasible, and goes against one of our guiding principles. That is why <s>-encoding of modern texts belongs to level two. Another example of level two is the encoding of the entries of the lexicographical sources in the text corpus. These could then be encoded in more detail: form, sense, grammatical information, etc. Further research is needed on this subject.

4 ILD-encoding of the typography of a text

4.1 Guiding principles

There are two possible views on text material: a database view and an editorial view. For the ILD, the database view on the material is the most important. We do not want to reproduce

the exact typographical features of a text. It is optimal accessibility to the content of the texts that we aim to offer our users. In a text, printed or published electronically, typography has different functions. Aesthetics is one of them. It often determines the choice of a particular colour of ink, of a particular font or type page. Typography, however, is also used to mark content elements of a text: titles of chapters, words or passages in a foreign language, etc. These are usually deviations from the default mark-up of a text, such as changes in font, or changes in margin size, etc. Only significant deviations will be encoded, i.e. when they give information about the structure or indicate other relevant textual aspects.

To come to our proposal for typography, we have consulted, apart from the TEI-guidelines, the Parole project, the CES, the Women Writers Project [WWP] and Text Encoding in Libraries (Library of Congress).

4.2 Encoding proposal for the typography

Following the TEI-guidelines, the global atttribute 'rend' is used to encode typographical information in text that has already been interpreted by a TEI-tag, for instance typographical information on the title of a chapter which has already been encoded by <head>. The element <hi> with the global attribute 'rend', on the other hand, is used for text needing only typographical specification.

For determining the values of the rend attribute, we apply a certain level of abstraction. It is, for instance, indicated that certain text is boxed but without giving further information as to the size of the lines, the form, etc. The size of a letter is defined in relation to the default size as small, very small, extra small, large, etc.

According to the TEI guidelines, an attribute can, strictly speaking, contain only one value. For rendition that is a problem: a title, for instance, can be in a type x letter, in size y, and bold: three different values for three different types of typographical information. In the Parole project, and in the CES, this is solved by using an abbreviation for each value, linked by a hyphen [Parole] or separated by a space [CES]. A more systematic solution, however, is to use rendition ladders, as does the WWP. The value of the rend attribute is a rendition ladder, consisting of a series of one or more keywords, usually followed by a value or series of arguments, delimited by parenthesis. Default rendition is specified in the header. The deviation is specified in the rendition tag. Example: for a title, in the default font, printed in italic and in bold, the encoding will be: <head rend='weight (BO) slant (IT)'>Nooit meer slapen.</head> ('weight' and 'slant' are keywords, 'BO' and 'IT' values).

The typography in the ILD will be encoded using the following keywords and values (in this abstract a selection):

Font	explanation
NDEF1	*non default 1*
NDEF2	*non default 2*
NDEF3	*non default 3*
GOT	*gothic*
CIV	*civilité*

Weight	
BO	*bold*
Slant	
IT	*italic*
RO	*roman*
Underlined	
no argument	*underlined*
Size	
XS	*extra small*
VS	*very small*
SM	*small*
LA	*large*
VL	*very large*
XL	*extra large*

Table 2: excerpt of the typographic encoding

5 Concluding remarks

Since the basic TEI-tagset for the ILD was developed before the release of the TEI P4 guidelines (XML edition), the encoding proposal is in SGML, not in XML. We will carefully review the benefits and downsides of using XML before making a final decision. We have presently reached a stage where we can apply the basic ILD-encoding on the planned prototype of the ILD [Van Dalen et al. 2002]. Practical experience will undoubtedly lead to further refinements of the present ILD-tagset.

Acknowledgements

Previous discussions on the subject with K.H. van Dalen-Oskam, J. de Does, D.J.G. Geirnaert, J.G. Kruyt and J.J.W. van der Voort van der Kleij formed a major contribution to this paper.

References

[CELT] Corpus of Electronic Texts (http://www.ucc.ie./celt/)
[CES] Corpus Encoding Standard (http://www.lpl.univ-aix.fr/projects/multext/CES/CES1-1.html)
[DBNL] Digitale Bibliotheek voor de Nederlandse Letteren (http://www.dbnl.nl/)
[Van Dalen et. al. 2002] Dalen-Oskam, K.H. van, D.J.G. Geirnaert and J.G. Kruyt, 2002. Text Typology and Selection Criteria for a Balanced Corpus: the Integrated Language Database of 8th-21st-Century Dutch, to be submitted to: *Proceedings EURALEX 2002.*
[ETCL] Electronic Text Centre Leiden (http://etcl.leidenuniv.nl/)
[Etext] Electronic Text Centre of the University of Virginia Library (http://etext.lib.virginia.edu/)
[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.
[Parole] (http://www.inl.nl/eng/index.htm) (European projects)
[TEI] Text Encoding Initiative (http://www.tei-c.org/)
[Text Encoding in Libraries] (Library of Congress) (http://www.indiana.edu/~letrs/tei/#level2)
[WWP] Women Writers Project (http://www.uic.edu/orgs/tei/trc/nwi/exwwprnd.html)