United in Diversity: Dutch Historical Dictionaries Online

Katrien Depuydt
Jesse de Does
Instituut voor Nederlandse Lexicologie

The Integrated Language Database of Dutch *(ILD) is a project of the Institute for Dutch Lexicology in Leiden, which integrates corpora, computational lexica and dictionaries describing the Dutch language from ca. 500 until the present. In 2007, the dictionary component was released, already containing two major historical dictionaries of Dutch, the* Woordenboek der Nederlandsche Taal *(WNT, Dictionary of the Dutch Language, 1500-1976) and the* Vroegmiddelnederlands Woordenboek *(VMNW, Dictionary of Early Middle Dutch, 1200-1300). When, by 2009, the* Middelnederlandsch Woordenboek *(MNW, Dictionary of Middle Dutch, ~1250 - 1550) and the* Oudnederlands Woordenboek *("ONW", Dictionary of Old Dutch, a current project at INL, to be finished in 2008, ca. 500-1200) will have been added, researchers of Dutch will have access to dictionaries covering the complete history of the Dutch language. The choice of a single application, integrating the dictionaries so that a user might query one or more dictionaries simultaneously, was a logical step because of the complementary nature of the dictionaries. The challenge was not only providing the user with optimal access to the dictionary information, but also doing so without compromising the uniqueness of each individual dictionary. We sketch the principles underlying the application.*

1. Introduction

The *Integrated Language Database of Dutch* (ILD) is a project of the Institute for Dutch Lexicology in Leiden in which corpora, computational lexica and dictionaries describing the Dutch language from ca. 500 until the present will be integrated [1]. In 2007, the dictionary component was released, already containing two major historical dictionaries of Dutch, the *Woordenboek der Nederlandsche Taal* (WNT, Dictionary of the Dutch Language, 1500-1976) and the *Vroegmiddelnederlands Woordenboek* (VMNW, Dictionary of Early Middle Dutch, 1200-1300). When, by 2009, the *Middelnederlandsch Woordenboek* (MNW, Dictionary of Middle Dutch, ~1250-1550) and the *Oudnederlands Woordenboek* ('ONW', Dictionary of Old Dutch, a current project at INL, to be finished in 2008, ca. 500-1200) will have been added, researchers of Dutch will have access to dictionaries covering the complete history of the Dutch language.

Choosing for one application, integrating the dictionaries so that a user can query one or more dictionaries simultaneously, was a logical step to take because the dictionaries complement each other. The challenge was not only to give the user optimal access to the dictionary information, but also to do so without compromising the uniqueness of each individual dictionary.

2. Dictionary data model

All dictionaries were already available in digital form. We started by first analysing the content, then the level of encoding and finally the applied encoding system. A thorough analysis of the content of each dictionary revealed that, in spite of obvious differences, they are very similar as to their macrostructure: headword, section with linguistic information at entry level, section with semantic analysis of the headword and section with related entries. They differed however greatly as to their level of encoding. In the original WNT data, the sense hierarchy of the article is encoded, but

[1] For information on the project, see Kruyt (2004). When the project started in 2000, the assumption was that the oldest Dutch material would not go back further than the 8th century. The editors of the ONW (Dictionary of Old Dutch) discovered later on that there is Old Dutch material dating from around 500.

individual citations only sporadically. The opposite is true for the MNW[2]. As for the VMNW and ONW, the situation is close to ideal: virtually every information category is distinguishable, either as a table in a relational database (VMNW) or encoded in the XML of the article (ONW). Since for each dictionary, the encoding system was different, and there was no compelling reason to use any of them, we chose to standardize the data by converting it to the XML version of TEI P4 (Text Encoding Initiative[3]). It is not only widely used for online publishing of dictionaries (Grimm[4], Mittelhochdeutsche Wörterbucher im Verbund[5], Anglo-Norman Dictionary[6]), but application to our data was pretty straightforward. More important to us was the fact that it enables both fine-grained and coarse-grained encoding. We decided to convert all available encoding in each dictionary to TEI, and we established a minimal level of encoding required for all dictionaries. Thus we did not need to impose one dictionary structure and level of encoding upon the others, but were still able to have simultaneous retrieval on the dictionaries. Achieving the minimal level of encoding implied a lot of data work for some of the dictionaries. Some additional data development was done for the sake of simultaneous retrieval. We have added a Modern Dutch (equivalent) lemma to each headword, so as to deal with the different headword spellings each dictionary has according to the language period it describes. And we mapped the indication of part of speech in each dictionary to a uniform one. Finally abbreviated variant forms, compounds and derivates were expanded. The latter two also received encoding as a headword, so that it does not matter how a particular word is treated in a dictionary: as headword or related entry.

3. Dictionary application model

Since the dictionary application is freely accessible, we will only go into some general underlying principles. As mentioned before, we did not want to integrate the dictionaries by mere extensive linking; we wanted to enable integrated searches, respecting each dictionary's own information categories. In the application, a user can select one or more dictionaries and for instance search for a headword in the selected dictionaries simultaneously. By using the Modern Dutch lemma as the search key, this can be done without knowledge of the historical spelling. Since the dictionaries share a minimal level of encoding, simultaneous searches on other information categories within the dictionaries are also possible. When a search is not applicable to one of the dictionaries, the search field is greyed out when only that particular dictionary is selected, or, in case of combined dictionary searching, no results from that particular dictionary are obtained. Another important issue for us was to approach the dictionaries in a more corpus-like fashion, meaning that we aimed at providing the user as much relevant information as possible without forcing him/her to read through a complete dictionary article. This was necessary because of the length of the dictionary articles: an article like *water* in the WNT for instance contains over 5700 citations and more than 300 senses and subsenses, totalling 144,450 words[7]. The corpus-like approach is visible in several places. When looking for a word or words in a sense, a citation or in a full article in the simple search option, the result will be displayed as concordances with the number of results per article (cf. fig. 1).

---

[2] Both WNT and MNW first appeared on CD-ROM (not available anymore).

[3] *http://www.tei-c.org*

[4] *http://germazope.uni-trier.de/Projects/DWB*: the online version of the *Deutsches Wörterbuch* of Jacob and Wilhelm Grimm. The dictionary is also available on CD-ROM.

[5] *http://germazope.uni-trier.de/Projects/MWV*: the online version of the *Mittelhochdeutsches Wörterbuch* of Georg Friedrich Benecke,Wilhelm Müller and Friedrich Zarncke, the *Mittelhochdeutsches Handwörterbuch* by Matthias Lexer, the *Findebuch zum mittelhochdeutschen Wortschatz* of Kurt Gärtner, Christoph Gerhardt, et. al and of the *Nachträge zum Mittelhochdeutschen Handwörterbuch von M. Lexer.* There is also a version of this application on CD-ROM.

[6] *http://www.anglo-norman.net/:* the online version of the *Anglo-Norman Dictionary* of William Rothwell, Stewart Gregory, William Rothwell, David Trotter et al.

[7] An average novel like *Pride and Prejudice* has about 124,000 words.

If the concordances do not provide enough information, or more context is needed, the user can always view the article the concordances were found in. Within advanced search, the user can adapt the display of the search results to the kind of research he or she wants to do. Either the result is a list of articles, like in classic dictionary applications, or he can opt for a list of dictionary meanings, citations, collocations, or head sections of a dictionary article (cf. fig. 2).

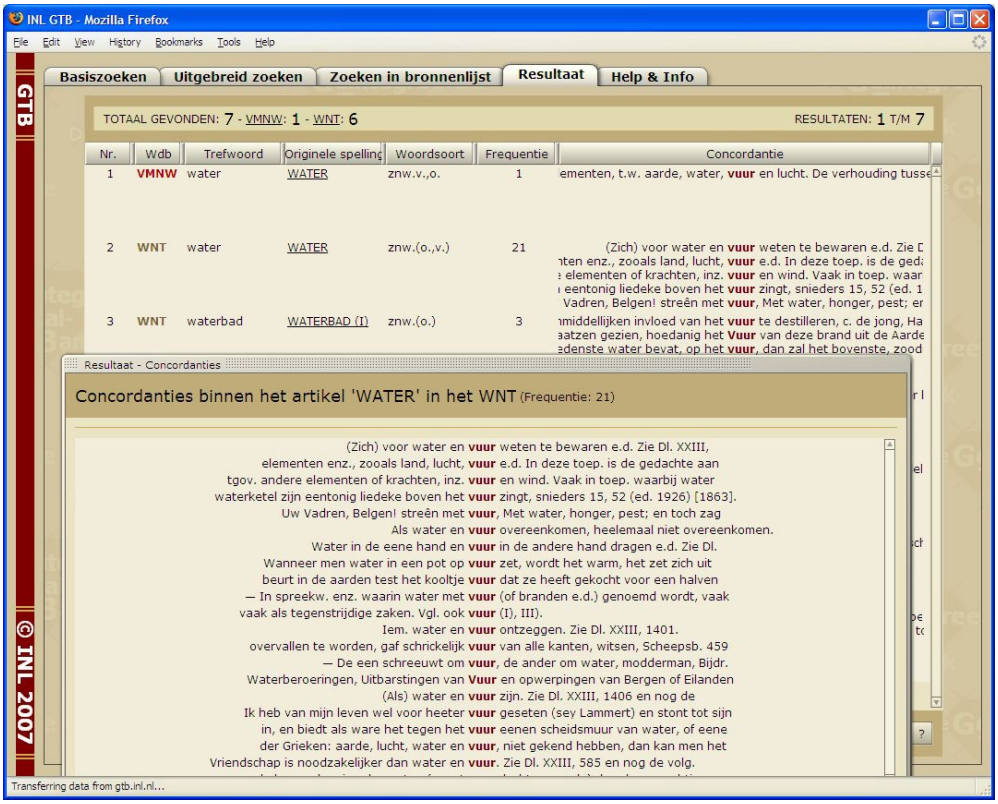

Figure 1. KWIC-view

Figure 2. Result from advanced search for definitions containing *vlinder* (butterfly).

An specially designed highlighting mechanism makes sure that there is a one to one correspondence of of search results and hit highlighting in the article view (cf. fig. 3).

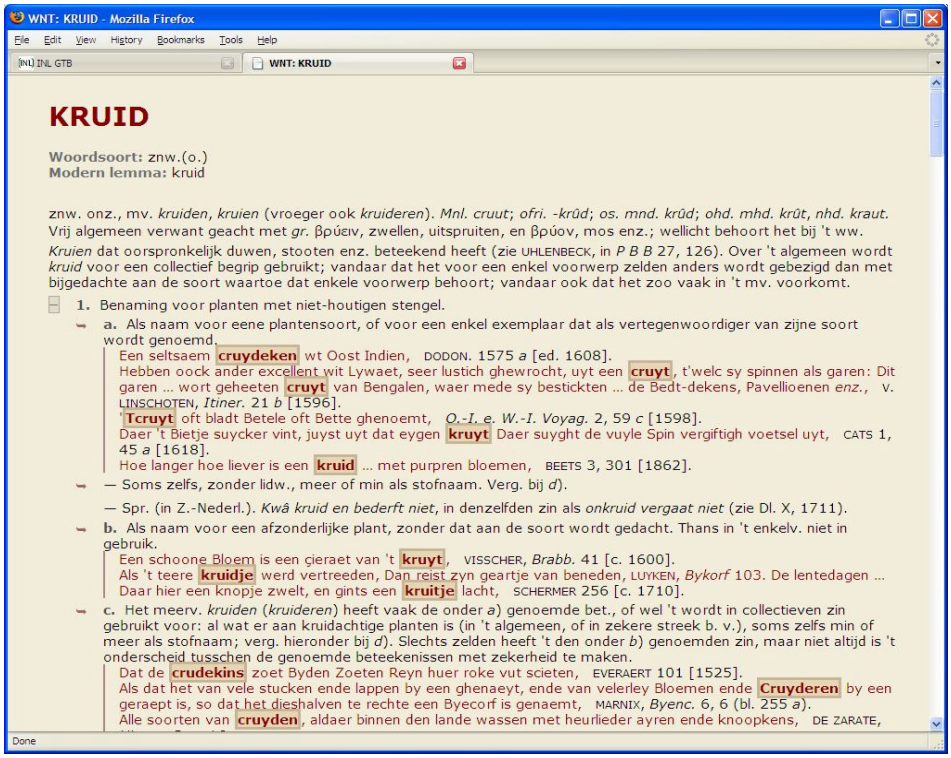

Figure 3. Highlighting of hits after regex search for variants of *kruid* (herb) in citations.

To make the dictionary application even more suitable for research purposes, all search results can be exported in HTML, XML or in CSV[8]. Searches with regular expressions are possible in every search field of the application.

4. User interface

We used the Openlaszlo platform[9] (release 3.3.1) with flash object code to ensure cross-browser compatibility of the user-interface.

The aim of platform-independency was achieved. Though the interface was implemented before the advent of Windows Vista or even Microsoft Internet Explorer 7 and the recent beta version of IE8, no changes in the Openlaszlo part of the user interface were necessary for these platforms.

5. Search engine

The search engine was implemented independently of the user interface. A preliminary investigation showed that neither XML-based nor relational databases were adequate to the task of combining the necessary full-text search features with structured querying.

We ended up using a combination of a MySQL database[10] for the storage and retrieval of the search results, and the open source Lucene search engine[11] (release 1.9) for full-text search functionality.

To ensure an efficient evaluation of a combined query on the entry, the sense and the quotation level, no preliminary computation of the partial results on each level is performed. Projected streams of

[8] HTML: HyperText Markup Language; XML: eXtensible Markup Language; CSV: comma-separated values.

[9] *http://www.openlaszlo.org*.

[10] *http://www.mysql.com/*

[11] *http://lucene.apache.org/*

results on each level (*spans* in Lucene terminology) are combined on the level of the specified granularity of the search result. The resulting search engine can handle complex combinations of queries without loss of efficiency.

6. Access

The dictionary application is accessible without payment after a simple one time registration, providing the user with username and password. The username and password have to be entered only once for each user on a single workstation. The application URL is *http://gtb.inl.nl*.

7. Future work

The next major milestone in the development of the Integrated Language Database will be the addition of the dictionary of the Middle Dutch Dictionary and the Dictionary of Old Dutch. Since 2007, work has also started on the lexicon component of the ILD, by the development of a large integrated lexicon of the Dutch Language, a diachronic lexicon for 6th – 21st century Dutch (the socalled "GiGaNT"-lexicon). With the help of this resource (which will of course also build on existing language resources), we hope to achieve the integration of corpus and dictionary material without necessarily lemmatizing the full corpus text. This means we will have to extend the use of modern search keys from dictionary lemmata to inflected forms in running text. Given the amount of orthographic and other linguistic variation in Dutch historical documents, this is a major challenge.

References

Kruyt, J. G. (2004). "The Integrated Language Database of 8th - 21st-Century Dutch." In Lino, M. T.; et al. (eds.). *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Paris: ELRA. 1751-1754.