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. 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, 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 MNW². 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 Initiative³). It is not only widely used for online publishing of dictionaries (Grimm⁴, Mittelhochdeutsche Wörterbucher im Verbund⁵, Anglo-Norman Dictionary⁶), 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 words⁷. 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).

	eken	Uitgebreid zo	ZOCKCI ZOCKCI	n in bronnenl	ijst nese	taat Help & Info
тс	DTAAL GEV	ONDEN: 7 - <u>VMM</u>	<u>IW: 1 - WNT: 6</u>	RESULTATEN: 1 T/M		
Nr.	Nr. Wdb Trefwoord Originele spelling Woordsoort Frequentie Concordantie					Concordantie
1	VMNW	/ water	WATER	znw.v.,o.	1	ementen, t.w. aarde, water, vuur en lucht. De verhouding tus
2	WNT	water	WATER	znw.(o.,v.)	21	(Zich) voor water en vuur weten te bewaren e.d. Zie ten enz, zooals land, lucht, vuur e.d. In deze toep, is de gee elementen of krachten, inz. vuur en wind. Vaak in toep, waa eentonig liedeke boven het vuur zingt, snieders 15, 52 (ed. Vadren, Belgeni streén met vuur. Met water, honger, pest; e
	WNT taat - Conc cordanti		WATERBAD (I)			
Result	taat - Conc	es binnen h	et artikel 'WA (Zich)	ATER' in het	: WNT (Frequencies)	aatzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoo entie: 21) bewaren e.d. Zie DI. XXIII,
Result	taat - Conc	es binnen h es tinnen k es tgov. a	(Zich) (Zich) Idementen enz., zoo ndere elementen o el zijn eentonig liec	ATER' in het) voor water en bals land, lucht, v of krachten, inz. v deke boven het	: WNT (Frequence) wuur weten te wuur e.d. In de wuur en wind. wuur zingt, sni	aatzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoo rentie: 21) bewaren e.d. Zie DI. XXIII, ze toep, is de gedachte aan /aak in toep. waarbij water ders 15, 52 (ed. 1926) (1863).
Result	taat - Conc	es binnen h es binnen h e tgov. a waterket	(Zich) (Zich) lementen enz., zoo el zijn eentonig liec Uw Vadren, Belg Water in de	ATER' in het voor water en bals land, lucht, of krachten, inz. deke boven het deke boven het vater en Als water en e eene hand en	: WNT (Frequer vuur weten te vuur e.d. In de vuur en wind. vuur zingt, sni vuur, Met wat vuur, Met wat vuur in de anc	aatzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoo entie: 21) bewaren e.d. Zie DI. XXIII, ze toep, is de gedachte aan /aak in toep. waarbij water ders 15, 52 (ed. 1926) (1663). r, honger, pest; en toch zag omen, heelemaal niet overeenkomen. er e hand dragen e.d. Zie DI.
Result	taat - Conc	es binnen h es binnen h tgov, a waterket t	(Zich) lementen enz., zo el zijn eentonig liec Uw Vadren, Belg Water in de Water in de Vanneer men wate eurt in de aarden 1	ATER' in het oals land, lucht, if krachten, inz. Jeke boven het enl streën men e eene hand en v r in een pot op test het kooltje	: WNT (Frequential of the second of the seco	atzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoo entie: 21) bewaren e.d. Zie DI. XXIII, ze toep, is de gedachte aan Vaak in toep, waarbij water eders 15, 52 (ed. 1926) [1863]. r, honger, pest; en toch zag men, heelemaal niet overeenkomen. ere hand dragen e.d. Zie DI. thet warm, het zet zich uit ef gekocht voor een halven
Result	taat - Conc	es binnen h es binnen k tgov.a waterket L - Ir vaak	(Zich) lementen enz., vz. lementen enz., vz. el zijn eentonig liec Uw Vadren, Belg Water in de Water in de Wanneer men wate eutri in de aarden in spreekw. enz. wa als tegenstrijdige :	ATER' in het ovor water en oals land, lucht, i forachten, inz. deke boven het els water en e eene hand en ir in een pot op etst het kooltje iaarin water met zaken. Vgl. ook	: WNT (Frequ vuur weten te vuur e.d. In de vuur zingt, sni vuur, Met wat vuur vereenk vuur in de anc vuur zet, worc vuur zet, worc vuur di ze he vuur (of brand vuur (of brand vuur (), III). vuur ontzegge	atzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoor entie: 21) bewaren e.d. Zie DI. XXIII, ze toep, is de gedachte aan Aakin toep, waarbij water eders 15, 52 (ed. 1926) [1863]. r, honger, pest; en toch zag men, heelemaal niet overeenkomen. ere hand dragen e.d. Zie DI. ch et warm, het zet zich uit ef gekocht voor een halven an e.d.) genoemd wordt, vaak n. Zie DI. XXIII, 1401.
Result	taat - Conc	es binnen h es binnen h tgov. a waterket t t u t vaak	(Zich) dementen enz., zoz ndere elementen el zijn eentonig lieg Uw Vadren, Belg Water ind Water in de aarden i s spreekw. enz. wa als tegenstrijdge arvallen te worden, — De een	ATER' in het voor water en voor water en forachten, inz. deke boven het en streën met Als water en e eene hand en rin een pot op test het kooltje taskn vetremet gaf schrickelijk schreeuwt om	: WNT (Frequencies of the second seco	atzen gezien, hoedanig het Vuur van deze brand uit de Äard adenste water bevat, op het vuur, dan zal het bovenste, zoor entie: 21) bewaren e.d. Zie DI. XXIII, ze toep, is de gedachte aan /aak in toep, waarbij water eders 15, 52 (ed. 1926) (1863). r, honger, pest; en toch zag omen, heelemaal niet overeenkomen. ere hand dragen e.d. Zie DI. tet warm, het zet zich uit eft gekocht voor een halven an e.d.) genoemd wordt, vaak

Figure 1. KWIC-view

ΤΟΤΑΑ	L GEVOND	EN: 82 - <u>VMNW</u> : 4 - <u>)</u>	<u>WNT</u> : 78	RESULTATEN: 1 T/M 82	
Nr.	Wdb	Trefwoord	Originele spelling	Woordsoort	Betekenis
50	WNT	uil	UIL (I)	znw.(m.,v.)	.2.b (Overdr.) (Amst.) Weesjongen. Zoo genoemd naa de kleuren van de kleeding der weeskinderen in Amsterdam (zwart en rood) die overeenkomen met die van de een be, soort viinder: het weeskind (zie BREHM-HUIZINGA 3, 494 (1910)).
51	WNT	vijfwouter	VIJFWOUTER	znw.(m.)	1 Vlinder, kapel.
52	WNT	vleek	VLEEK	znw.(v.)	4 Opm. De bij KIL. [1599] en in enkele latere wdb. vermelde bet. vilinder ("Vieke, vleken. sax. sic. j. pepel. Papillo") berust woch. op een verkeerde lezing van CHYTRAEUS: deze heeft Vieken, waarin de beginletter als u moet worden opgevat (DE SMET in ZS
53	WNT	vlender	VLENDER	znw.(m.)	Bijvorm van <i>flenter;</i> zie ald., en nog de volg. aanh. Daarnaast ook vlinder (TER LAAN).
54	WNT	vlerk	VLERK	znw.(v.,m.)	5.c Vand. van de ziel (onder het beeld van een vlinder).
55	WNT	vlichelter(e)	<u>VLICHELTER(E)</u>	znw.(m.,v.)	Een alleen bij KILIAAN aangetroffen benaming voor vlinder, die teruggaat op vivelter(e), onder invloed gekomen van vlichelen (vgl. PAUWELS in <i>H. Top. Dial.</i> 9 [1935], 344).
56	WNT	vliegwouter	VLIEGWOUTER	znw.(m.)	Een door KILIAAN en zijn navolgers vermelde etymologiseerende variant van wiewouter ' Vlinder ', waarnaast volgens versch, idiotica ook vilege(n)bouter schijnt voor te komen (SCHUERM. [1865-1870]; JOOS [1900-1904]; TEIRL.)
57	WNT	vliemel	VLIEMEL	znw.(m.)	Eenmaal aangetroffen als een der vele namen voor den vlinder . Misschien een combinatie van de eerste

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).

🕲 WNT: KRUID - Mozilla Firefox	
Eje Edit Vjew Higtory Bookmarks Tools Help	$\langle \rangle$
INU INL GTB III WNT: KRUID III	-
	^
KRUID	
RROID	F
Woordsoort: znw.(o.)	
Modern lemma: kruid	
znw. onz., mv. kruiden, kruien (vroeger ook kruideren). Mnl. cruut; ofrikrûd; os. mnd. krûd; ohd. mhd. krût, nhd. kraut. Vrij algemeen verwant geacht met gr. βρύειν, zwellen, uitspruiten, en βρύον, mos enz.; wellicht behoort het bij 't ww.	
Kruien dat oorspronkelijk duwen, stooten enz. beteekend heeft (zie UHLENBECK, in <i>P B B</i> 27, 126). Over 't algemeen wordt	
kruid voor een collectief begrip gebruikt; vandaar dat het voor een enkel voorwerp zelden anders wordt gebezigd dan met	
bijgedachte aan de soort waartoe dat enkele voorwerp behoort; vandaar ook dat het zoo vaak in 't mv. voorkomt.	
 Benaming voor planten met niet-houtigen stengel. a. Als naam voor eene plantensoort, of voor een enkel exemplaar dat als vertegenwoordiger van zijne soort 	
 a. Als haam voor eene partensoort, or voor een enker exemplaar dat als vertegenwoordiger van zijne soort wordt genoemd 	
Een seltsaem cruydeken wt Oost Indien, DODON. 1575 a [ed. 1608].	
Hebben oock ander excellent wit Lywaet, seer lustich ghewrocht, uyt een cruyt, t'welc sy spinnen als garen: Dit garen wort geheeten cruyt van Bengalen, waer mede sy bestickten de Bedt-dekens, Pavellioenen <i>enz.</i> , v.	
LINSCHOTEN, Itiner. 21 b [1596].	
'Tcruyt oft bladt Betele oft Bette ghenoemt, OI. e. WI. Voyag. 2, 59 c [1598]. Daer 't Bietje suycker vint, juyst uyt dat eygen kruyt Daer suyght de vuyle Spin vergiftigh voetsel uyt, CATS 1,	
45 a [1618].	
Hoe langer hoe liever is een kruid met purpren bloemen, BEETS 3, 301 [1862].	
 Soms zelfs, zonder lidw., meer of min als stofnaam. Verg. bij d). 	
 — Spr. (in ZNederl.). Kwå kruid en bederft niet, in denzelfden zin als onkruid vergaat niet (zie Dl. X, 1711). b. Als naam voor een afzonderlijke plant, zonder dat aan de soort wordt gedacht. Thans in 't enkelv. niet in 	
gebruik.	
Een schoone Bloem is een çieraet van 't kruyt , VISSCHER, Brabb. 41 [c. 1600].	
Als 't teere kruidje werd vertreeden, Dan reist zyn geartje van beneden, LUYKEN, Bykorf 103. De lentedagen Daar hier een knopie zwelt, en gints een kruitie lacht, SCHERMER 256 [c. 1710].	
c. Het meerv. kruiden (kruideren) heeft vaak de onder a) genoemde bet., of wel 't wordt in collectieven zin	
gebruikt voor: al wat er aan kruidachtige planten is (in 't algemeen, of in zekere streek b. v.), soms zelfs min of meer als stofnaam; verg. hieronder bij d). Slechts zelden heeft 't den onder b) genoemden zin, maar niet altijd is 't	
meet als stomaarn; verg, heronder bij oj. siechts zeiden neert t den onder bij genoemden zin, maar met augus t onderscheid tusschen de genoemde beteekenissen met zekerheid te maken.	
Dat de crudekins zoet Byden Zoeten Reyn huer roke vut scieten, EVERAERT 101 [1525].	
Als dat het van vele stucken ende lappen by een ghenaeyt, ende van velerley Bloemen ende Cruyderen geraept is, so dat het dieshalven te rechte een Byecorf is genaemt, MARNIX, Byenc. 6, 6 (bl. 255 a).	
Alle soorten van cruyden , aldaer binnen den lande wassen met heurlieder ayren ende knoopkens, DE ZARATE,	~
Done	11

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 platform⁹ (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 database¹⁰ for the storage and retrieval of the search results, and the open source Lucene search engine¹¹ (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.