Wurdboek fan de Fryske taal/Dictionary of the Frisian Language Online: New Possibilities, New Opportunities

Hindrik Sijens, Fryske Akademy, Leeuwarden
Katrien Depuydt, Instituut voor Nederlandse lexicologie, Leiden

The Wurdboek fan de Fryske Taal (Dictionary of the Frisian Language, WFT) describes the vocabulary of the Modern West Frisian language and consist of 25 volumes of 400 pages each. The dictionary contains more than 100,000 entries. This paper is intended to show that an electronic version of the WFT, once the data have been converted to state-of-the-art standards and made available to the public by means of an advanced retrieval application, will be a modern lexicographical resource of significant value. Integrating the WFT into the dictionary component of the Geïntegreerde Taalbank Nederlands (Integrated Language Database of Dutch, GTB) of the Instituut voor Nederlandse Lexicologie is the obvious means to reach this goal. In order to create more ways of searching the dictionary entries, data accessibility has to be enhanced by explicit tagging of information categories which can be exploited by a retrieval application. The process of implementing the online version of WFT took place in several stages: First the existing database had to be repaired and optimised. Mistakes and inconsistencies had to be repaired. The logical structure had to be parsed and tagged with XML mark-up. Furthermore the newly created XML database had to be enriched with TEI encoding. And, finally the dictionary was incorporated into the GTB application. The WFT has been incorporated into the online dictionary application of the Dutch language bank, and so is freely available to a large audience, allowing interested parties to search in one of the most complete Frisian dictionaries, and to explore the Frisian language in relation to Dutch.

1. Introduction

On the 9th of November 1984, the first volume of the *Wurdboek fan de Fryske taal* (WFT, Dictionary of the Frisian Language) was published. Editor in chief Klaas van der Veen presented the first copy to her majesty Queen Beatrix of the Netherlands. A volume has been published every year since. Twenty five years later, the dictionary is almost finished. The editorial phase has been completed; the 25th and final volume will be published in 2010. The printing of the final volume will mark the completion of a major dictionary project.

The WFT describes the vocabulary of the Modern West Frisian language from the period 1800-1976. More than 10,000 pages of lexicographic information about Modern West Frisian will be available to the professional linguist and the layman interested in the Frisian language, which with almost 400,000 speakers, is one of the minority languages of Europe.

The WFT is, to some extent, a language museum. It records a changing society and a changing language. Once all volumes have been published, will it be more than a museum piece, a nice collection of books on a shelf? This paper is intended to show that once the data have been converted to state-of-the-art standards and made available to the public by means of an advanced retrieval application, an electronic version of the WFT, will be a modern lexicographical resource of significant value.

2. The dictionary

The WFT is a scholarly, descriptive dictionary. It has its roots in the nineteenth century tradition of large dictionaries, and can therefore be compared with the *Oxford English Dictionary*, the *Deutsche Wörterbuch* and the *Woordenboek der Nederlandsche taal*. The dictionary contains 120,000 lemmas. The entries give information about the spelling of the headword, its part of

speech and its pronunciation. In addition, information is given about the flexion and etymology of the headword. The semantic section provides the user with information about the meanings of the headwords by means of definitions or translations into Dutch. All the meanings of a word are illustrated by citations, so the user is able to verify the lexicographer's work. Idiomatic information is given in the idioms section, which contains collocations, proverbs and figurative meanings. The final section of an entry describes compounds and derivatives belonging to the headword.

Five hundred copies have been printed of each volume, and some 400 subscribers receive a copy. They are language enthusiasts, professional linguists as well as university and public libraries. The WFT is a paper dictionary with restricted search possibilities. The alphabet is the only means by which the headwords and their descriptions can be accessed. The copious linguistic information in the dictionary deserves to be explored, not only by more people but also in a more exhaustive way.

The central question here is how to reach a broader audience and to provide them with as much linguistic information about Frisian as possible. To reach this goal, making the dictionary available online would be is an excellent option. In order to create more ways of searching the dictionary entries, data accessibility has to be enhanced by explicit tagging of information categories which can be exploited by a retrieval application.

3. Integration

Integrating the WFT into the dictionary component of the *Geïntegreerde Taalbank Nederlands* (Integrated Language Database of Dutch, GTB) of the *Instituut voor Nederlandse Lexicologie* is the obvious means of reaching this goal. The GTB integrates corpora, computational lexica and dictionaries describing 15 centuries of Dutch language. The online dictionary component contains four Dutch dictionaries: the *Oudnederlands Woordenboek* (ONW, Dictionary of Old Dutch, ca. 500–1200), the *Vroegmiddelnederlands Woordenboek* (VMNW, Dictionary of Early Middle Dutch, 1200-1300), the *Middelnederlandsch Woordenboek* (MNW, Dictionary of Middle Dutch, ~1250 – 1550) and the *Woordenboek der Nederlandsche Taal* (WNT, Dictionary of the Dutch Language, 1500-1976). Subscription to the dictionary application is free of charge. By the end of 2009, more than 74,000 subscribers were registered. This is in addition to the access via IP address granted to schools, universities, libraries and other large organisations.

The WFT and the Dutch dictionaries were developed in the same lexicographical tradition. Integration is feasible because of the similarity in the structures. The advantages of linking the Frisian dictionary with the online Dutch historical dictionaries are many. Linking the dictionaries enhances the possibilities for synchronic and diachronic analysis of both languages. Which words appear in both languages, which are specifically Frisian or Dutch? What are the phonological and morphological differences between the two languages? What is the influence of the Dutch language on Frisian and vice versa? An additional value is that etymological information about Frisian words can be derived from one or more of the Dutch linked dictionaries.

In order to link the WFT to the GTB, a list of search options had to be drawn up. The starting point was the existing application and the possibilities of the tagged WFT data. Because of the similarity in the structures, the basic criteria, and combinations thereof, for searching for

dictionary entries, word senses, quotations, collocations in the dictionary application are also relevant for increasing the accessibility of the WFT. On the other hand, it was possible to link most of the information categories in the WFT to the application's existing search options, for example variants of the headword, words in collocations, idioms and proverbs, or languages mentioned in etymology field.

4. Data curation for the online WFT

The process of implementing the online version of WFT took place in several stages: First, the existing database had to be repaired and optimised. The logical structure had to be parsed and tagged with XML mark-up. Then, the newly created XML database had to be enriched with TEI encoding. And, finally the dictionary was incorporated into the GTB application.

Correcting errors

The original data for the print edition of the dictionary were stored in a database. Since the early 1990s, this has been a BRS/search database. BRS/Search is a full-text database and information retrieval system which uses a fully-inverted indexing system to store, locate, and retrieve unstructured data. The only metadata added to a dictionary entry are *Word* and *Desc*. *Word* refers to the headword of the dictionary entry, and *Desc* to a section devoted to de description of a particular word sense within the full text of the entry. No other information categories were tagged explicitly. The data were stored in Windows cp1252 format and marked with layout codes that are used by scripts to convert the database text to rtf documents. The entries of the dictionary were accessible with a search and input interface and a simple text editor.

Before the data could be added to the GTB, mistakes and errors, identified in the printed dictionary had to be corrected. For instance, the structure of the entry *astrant* 'cool, cheeky' suggested that there had to be more than one meaning, but only one meaning was given. Somewhere in the printing process something had gone wrong, and the second meaning 'smart, clever, attentive', never entered the dictionary, although the editor did edit it:

as'trant, adj. & adv. [astrOnt] a. 1869 R; **ek'strant**, º [EkstrOnt] 1860 R. Etym.: Fr. *assurant*. Lit. R W.J. BUMA, Wurk 47 [1969]. Astrant. **1**. aanmatigend, vrijpostig, brutaal. , A s t r a n t , *op een onbeschaamde wijze aanmatigend*. HALB., *Lex.* 127 [a. 1869]. E k s t r a n t , *'in de Wouden'*. *Fr.W.* [1900]. – De goede Pier, dy troch de extrante frou onder de tafel haldenwjar. T.G. v.d. MEULEN, *fen* 84 [1860]. As jy my de wjirheid freegje, den binne jimme meienoar wol hwet 'ekstrant'. J. HEPKEMA, *jouke* 17 [1894]. Dy astrante joad hie de posleinkas al waech en wiid iepen. *Fr.W.* [1900]. (It) wienen … dan ek mar in pear / astrante rabbelskûten. S. SPANNINGA, *rattelmansreau* 10 [1962].	**as'trant**, adj. & adv. [astrOnt] a. 1869 R; **ek'strant**, º [EkstrOnt] 1860 R. Etym.: Fr. *assurant*. Lit. R W.J. BUMA, Wurk 47 [1969]. Astrant (in bet. 1). **1**. aanmatigend, vrijpostig, brutaal. , A s t r a n t , *op een onbeschaamde wijze aanmatigend*. HALB., *Lex.* 127 [a. 1869]. E k s t r a n t , *'in de Wouden'*. *Fr.W.* [1900]. – De goede Pier, dy troch de extrante frou onder de tafel haldenwjar. T.G. v.d. MEULEN, *fen* 84 [1860]. As jy my de wjirheid freegje, den binne jimme meienoar wol hwet 'ekstrant'. J. HEPKEMA, *jouke* 17 [1894]. Dy astrante joad hie de posleinkas al waech en wiid iepen. *Fr.W.* [1900]. (It) wienen … dan ek mar in pear / astrante rabbelskûten. S. SPANNINGA, *rattelmansreau* 10 [1962]. 2. schrander, snugger; oplettend. , A s t r a n t , *'vaak in de volkstaal: schrander, snugger, vlug van bevatting … Ook wel: oplettend'*. *Fr.W.* [1900]. – In astrante jonge. *ibid*. Hy is altyd like astrant yn 'e saken. *ibid*.

Optimisation
The data had to be optimised in other ways as well. For instance, abbreviations such as *Id.* en *ibid.* for same author and same source had to be resolved. In a set of compounds with a common first part, the abbreviation marks had to be expanded. Another job was checking the consistency of cross-references between entries.

The part of speech information of the headwords needed to be mapped to the tag set used for the Dutch online dictionaries. For instance, a search query for reflexive verbs in the application uses the standardised category label *Ww refl.* where the WFT has the label *v.* with the addition of the reflexive pronoun *jin*. Linking the Frisian label to Dutch *ww refl.* will enable the simultaneous retrieval of both Frisian and Dutch verbs in this category.

Adding Modern Dutch equivalents
For users who do have a command of Dutch but no knowledge of Frisian, it may be difficult to search for a Frisian entry. That is why a Modern Dutch equivalent lemma was added to a Frisian headword. Although Frisian and Dutch are related languages, the differences are substantial. It is safe to assume that cognates like Dutch *neus* 'nose' and Frisian *noas* are equivalent. Therefore, the modern Dutch lemma *NEUS* covers both entries.

Both languages also know the words *naad* ('seam') and it meaning 'seam'. But another meaning of *naad* in Frisian is 'ridge', which is not recorded for *naad* in the Dutch dictionaries. The Dutch word for this meaning is *nok*.

Frisian *naad* can be mapped to the Modern Dutch equivalent *NAAD*. Subsequently, all compounds and derivates with *naad-* can be translated in the same way. The equivalent of Frisian *naadfoarst* ('ridge tile') would therefore be the non-existent Dutch equivalent NAADVORST. No Dutch, non-Frisian user would use this morphologically correct term to search the dictionaries for the concept 'ridge tile'. So when a Frisian lemma has no Dutch equivalent, another strategy has to be used to find the correct Frisian entry. Since the definitions in the WFT are mostly Dutch synonyms, a user can enter this synonym in the 'definition' search field in the application.

The assignment of Dutch equivalents to the Frisian headwords was done automatically by using scripts, and subsequently correcting manually.

Sources and references
It is possible to search the list of citation sources used in the dictionaries. Therefore, a list of sources used in the WFT has been linked to the dictionary. Also, a list with references to linguistic literature was created and added to the GTB.

TEI encoding
In order to implement the WFT into the GTB application, the dictionary data had to be converted to the TEI annotation scheme for printed dictionaries. The existing online dictionary application, part of the Dutch Language Bank, allows for querying in one or more dictionaries simultaneously. At the time plans were drawn up, the challenge was not only to give the user optimal access to the dictionary information, but to do so without compromising the uniqueness of each individual dictionary. All Dutch dictionaries were available in digital form, however in a different encoding system and with a different level of encoding. Similarities in structure though

were: headword; the section with linguistic information at entry level; the section with semantic analysis of the headword; and the section with related entries. TEI encoding for printed dictionaries was chosen as a standard because it allows both fine-grained and coarse-grained encoding. Moreover, all encoding needed for the main Dutch historical dictionaries could be converted to TEI without modifying the encoding scheme, which is more than can be said of competing standards like LMF. A basic encoding scheme for the Dutch dictionaries was defined at INL. This scheme defines a minimum level of mandatory encoding for all dictionaries necessary for the integrated retrieval on the dictionary data. Apart from the basic level of encoding which applies to all dictionaries, the additional encoding present in each of the dictionaries has been converted into TEI. Consequently, there are some retrieval possibilities applicable to all dictionaries, whereas others are applicable to only one, or a smaller group of dictionaries, depending on the level of encoding. The application of the TEI dictionary encoding scheme to the Dutch historical dictionaries is documented in Depuydt (2010).

Parsing

Writing parsing software in order to tag the logical structure of the dictionary entries caused some difficulties, due to the inconsistencies in the structure. For instance, the etymology section starts with the label *Etym*., but it can also contain morphological information such as ‘denominatief van *noas* (‘denominative of nose’). Proper etymological information consists of references to cognates and equivalents in other languages, or just references to other languages.

Headword WFT	Field *Etymology*
noas	Etym. → N. *neus*, D. *Nase*, E. *nose*.
noasje	Etym.: Fr., Lat.
noaskje I	Etym.: denominatief van *noas*.
noaskje II	Etym.: dim. van *noas*?
noaster	Etym. → ofri. *noster*, Gron. *nöster*, D. *Nüster*, E. *nostril*.
noat III	Etym. → *genoat*.
noatich	Etym. → *noat*?, *nuet*?
noatmuskaat	Etym. → N. *nootmuskaat*.
noatte	Etym. → *knotte*?
noazelje	Etym.: denominatief van *noas*.
noazem	Etym. → N. *nozem*.

In order to support specific queries, further analysis was needed to distinguish morphological and etymological information.

Morphology

One of the features of the GTB is searching for morphological elements of a headword and word formation types. This feature is also available for the WFT section of the GTB application. In the print edition of the dictionary, word formation is not a specific field in the microstructure. Although it is mentioned in the etymology section, most of the time it is absent. We have included the so called *morfo-database* of the Fryske Akademy in the application. This database contains detailed information about the morphological structure and components of most of the headwords in the dictionary.

Future

As far as the future is concerned, the etymology field could be made more explicit. Linking the Frisian entries to the Dutch Etymological dictionary *Etymologisch Woordenboek van het Nederlands* is one of the possibilities. The Fryske Akademy has collected an enormous amount of dialectological data of Frisian dialects. Including this information into the application would be of great value. Another new feature could be the connection of the dictionary with the Frisian language database, which is also project carried out by the Fryske Akademy.

5. Results

The online version of the WFT is now available to the public. The WFT has been incorporated into the online dictionary application of the Dutch language bank, and so is freely available to a large audience, allowing interested parties to search in the most complete Frisian dictionary, and to explore the Frisian language in relation to Dutch. The result: A modern language museum providing a service to all who are interested in the language but been unable to find room on their bookshelves for the 25 attractive printed volumes.

References

Depuydt, K.; Does, J. de (2008). 'United in Diversity: Dutch Historical Dictionaries Online'. In Bernal, E.; DeCesaris, J. (eds.). *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15-19 July 2008). Barcelona: IULA, DOCUMENTA UNIVERSITARIA 2008.

Depuydt, K.; Does, J. de (2010). 'TEI-structuurcodering van de woordenboeken in de woordenboekencomponent van de Taalbank Nederlands.'

Dykstra, A.; Schoonheim, T. (2010). 'Naar een onlineversie van het Woordenboek der Friese taal' (Paper read at the Frysk Filologekongres December 2008).

Dykstra, A. (1998). *A small language, a large dictionary* [on-line] Ljouwert: Fryske Akademy. http://www.fryske-akademy.nl/fa/3departments-and-disciplines/department-of-linguistics/lexicography-terminology/dictionary-of-the-frisian-language/wft-brochure [Access date 11 Mar. 2010].

De GeïntegreerdeTaalBank [on-line] Leiden: Instituut voor Nederlandse Lexicilogie http://gtb.inl.nl/ [access date 16 Mar. 2010].

Kruyt, J.G. (2004). 'The Integrated Language Database of 8th – 21st-Century Dutch'. In Lino, M.T, Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (eds.). *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 1751-1754. Paris: ELRA.

Wurdboek fan de Fryske Taal / Woordenboek der Friese Taal. Ljouwert/Leeuwarden: Fryske Akademy, 1984-2010.