Clarin-NL project WFT-GTB - 28 oktober 2010

Drs. H. Sijens
Lic. karien Depuydt

1. The WFT project

On July the 6th, her majesty Queen Beatrix of the Netherlands opened the online version of *The dictionary of the Frisian Language* or as it is called in Frisian: *Wurdboek fan de Fryske taal*. Hereafter I will use the abbreviation *WFT* for this dictionary. The presentation of the online version of the WFT was the final step in the WFT-GTB project, a project that was supported by Clarin-NL. This project involved a data curation and a demonstrator project and was carried out by two Clarin partners, the Fryske Akademy in Leeuwarden and the INL, the Dutch Institute for Lexicology in Leiden.

The *Wurdboek fan de Fryske Taal* describes the vocabulary of the Modern West Frisian language from the period 1800 until 1976. Next year the 25th and final printed volume will be published. 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 Dutch *Woordenboek der Nederlandsche taal*. The dictionary contains about 115,000 lemmas. The entries provides the user with 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 shows 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. The idioms section contains collocations, proverbs and figurative meanings. The final section of an entry describes compounds and derivations 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 was how to reach a broader audience and how to provide them with as much

linguistic information about Frisian as possible. To reach this goal, making the dictionary available online was an excellent option. By doing so, the digital surrounding enables extensive forms of free and structured search queries. More specifically, the database structure enables thematic search operations. And matching the WFT with Dutch historical dictionaries enables comparative studies with Dutch materials.
In order to create more ways of searching the dictionary entries, data accessibility had to be enhanced by explicit tagging of information categories which can be exploited by a retrieval application.

2. Integration

The technical embedding in the Integrated Language Database of Dutch, provides a useful application of existing technology and offers web users an integrated search tool. The Dutch Language Bank of the *Instituut voor Nederlandse Lexicologie* integrates corpora, computational lexica and dictionaries describing 15 centuries of Dutch language. The online dictionary component contains four Dutch dictionaries:

- the *Oudnederlands Woordenboek* (Dictionary of Old Dutch),
- the *Vroegmiddelnederlands Woordenboek* (Dictionary of Early Middle Dutch),
- the *Middelnederlandsch Woordenboek* (Dictionary of Middle Dutch)
- the *Woordenboek der Nederlandsche Taal* (Dictionary of the Dutch Language).

Subscription to the dictionary application was free of charge and by the end of 2009, more than 74,000 subscribers were registered. Since the language bank is part of the Clarin infrastructure, it is not necessary to subscribe anymore.

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 Dutch Language Bank, 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.

3. 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 Dutch Language Bank application.

Correcting errors

The original data for the print edition of the dictionary were stored in 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. Problem was that 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 the 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 text 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 interface and a simple text editor.

Optimisation

Before the data could be added to the Dutch Language Bank, known mistakes and errors,

identified in the printed dictionary had to be corrected. 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 adverbs ending with the Frisian suffix *-lik* and corresponding Dutch suffix *-lijk* in the application uses the standardised category label *bw.* where the original WFT has the label *adv.* Linking the Frisian label *adv.* to Dutch *bw.* enables the simultaneous retrieval of both Frisian and Dutch adverbs with *-lik/-lijk.*

Adding normalized 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 in the Dutch Language Bank. That is why for about 50.0000 lemmas a Normalized Dutch equivalent lemma was added to a Frisian headword. The assignment of Dutch equivalents to the Frisian headwords was done automatically by using scripts, and subsequently correcting manually. It is not easy to connect the remaining 65.000 lemmas. Due to etymological en semantic differences, not every Frisian word can be translated into a normalized Dutch lemma.

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 normalized Dutch lemma *NEUS* covers both entries.

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.

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.* This section serves the user with references to cognates and equivalents in other Germanic languages, but it can also contain morphological information such as 'denominatief van *noas* ('denominative of nose'), or just references to other languages.

Headword WFT	**Field *Etymology***
noas ‘nose’	Etym. → N. *neus*, D. *Nase*, E. *nose*.
noasje ‘notion’	Etym.: Fr., Lat.
noaskje I ‘to nose’	Etym.: denominatief van *noas*.

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

TEI encoding

In order to implement the WFT into the Dutch Language Bank 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, but 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.

4. The future

As already mentioned, to facilitate searches in Frisian for non-Frisian users, a first series of

approximately 50,000 Frisian lemmas have been connected to the normalized Dutch lemmas of the online dictionaries. This is the first step in the wider goal of connecting not only the different diachronic stages of Frisian and Dutch, but also to establish a cross-language correspondence.
The INL and Fryske Akademy proposed a new project for the second Clarin Call. The demonstrator part of this project will be an addition to the existing Dutch Language Bank-web application: a link to etymological articles in the Frisian Language Database in the ‘koppelingen’ frame of the application.
The articles will be stored in the Frisian Language Database of the Fryske Akademuy, but through a proper linking of lemmas and technical interoperability of the two database systems an access is given from within the Dutch Language Bank-web application. The Dutch Language Bank-application will be given access
to the Frisian Language Database, to conduct a search in its lemma list and to retrieve linked information, such as the scanned journal articles.

5 Results

The *Dictionary of the Frisian language* 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. The result: A modern language museum providing a service to all who are interested in the language but have been unable to find room on their bookshelves for the 25 attractive printed volumes.