The Online Dutch-Frisian Dictionary - Workflow

Frits van der Kuip and Hindrik Sijens
Fryske Akademy, Leeuwarden
fvdkuip@fryske-akademy.nl, hsijens@fryske-akademy.nl

1   Short description of the project

The online Dutch-Frisian dictionary is a bilingual dictionary, with modern standard Dutch as its source language and modern Frisian as its target language. The dictionary will replace the outdated paper dictionary published in 1985. It provides non-Frisian-speaking people with a tool for translating Dutch into Frisian; its main goal, however, is to provide Frisian-speaking people with a tool that helps them to write correct Frisian that is less coloured by the dominant Dutch language. Most Frisian speakers lack an education in the Frisian language, they are used to writing in Dutch, and to some extent they are illiterate in their mother tongue.

As an emancipating language, Frisian is still acquiring new functions and penetrating into new societal domains, or into domains that were traditionally reserved for the dominant language. Therefore, it needs new terminology. Dictionaries may play an important role in reinforcing a language, in language codification and language standardisation (Popkema 2010). A pregnant question for lexicographers, for instance, is how to choose between a loanword and a purism (Duijff 2000, Van der Kuip 2010).

To serve its users, not only will the dictionary contain more words than its 1985 predecessor, it will also pay more attention to their use, by including multiple examples, collocations and idioms. Because the dictionary will be published on-line in stages, this will provide an ideal opportunity for crowdsourcing (Gouws 2009). The idea is to open a mailbox for comments, suggestions and additions.

At this point, we are still in the preparatory stages of the project. By the end of 2014, we hope to present a demo version of these tools. Data analysis and editing will be carried out over the next five years. The project will be carried out by three lexicographers (1,8 fte), a software engineer (0,4 fte for the time being) and two lexicographical assistants (0,9 fte).

2   Lexicographical workflow

2.1 Preparation

In 2013, we drafted the conceptual framework of the dictionary, and wrote several lexical test entries in preparation. As we will need a Dutch corpus for our dictionary, we sat down with the Institute for Dutch Lexicology (INL) in Leiden. As a result of our discussions, the material from the *Algemeen Nederlands Woordenboek* ('General Dutch Dictionary', ANW) will be our point of departure, and we will partially adopt the structure of the articles in the dictionary. Since the ANW is a monolingual explanatory dictionary, and the Dutch-Frisian dictionary is a bilingual dictionary, the INL has adapted its editor software (see Niestadt 2009) according to our demands. The editors are currently working on writing test entries in order to test the editor and to optimise it for editing this dictionary.

2.2 Data acquisition

The Dutch-Frisian dictionary is based on the data provided by the ANW: lexical entries, meanings, examples, collocations, expressions and proverbs. The translations are based on Frisian data generated from the Frisian language database. The database contains 24,000,000 tokens, however, it is not easily accessible: the data still have to be lemmatised and postagged. It is possible to construct KWIC ('keyword in context') concordances at this point. The Dictionary Writing System can be used to link to grammatical information and to prior lexicographical work, especially the 1985 version. All this information is available from online portals (*De Geïntegreerde Taal-Bank, Taalweb, Taalportaal*). Data mining of the internet may yield new information on frequency or meaning.

Our greatest challenge will be finding Frisian equivalents for Dutch neologisms. It is not inconceivable that we will have to come up with some of these equivalents ourselves.

2.3 Computerisation

The source language data are provided by the ANW. In addition, we will copy the INL editor and ANW database, which will then be adapted to the demands of the editors. To every field with elements from the source language, the Dutch-Frisian dictionary will add a field for the translation into the target language. The data will be entered into an XML schema, stored in a MySQL database with metadata (cf. Tiberius & Schoonheim 2014, section 2.3.1).

2.4 Data processing

The data for the Dutch-Frisian dictionary are provided by the ANW: the lexical entries, meanings, examples, collocations, expressions and proverbs. The 'word family' field from the ANW (a list of compounds and derivatives of the headword) will be copied as well. The ANW material will be assessed and selected when the separate entries are processed.

2.5 Data analysis

The fields on meaning, examples, collocations and expressions copied from the ANW will be supplied with a Frisian equivalent. Information on phonetics, inflexion, grammar and hyphenation will be added to one-word translations, because there are but few electronic lexicographic facilities providing this type of information. Moreover, this Dutch-Frisian dictionary will not have a Frisian-Dutch counterpart in the near future.

At this stage, the lexicographers will match the Dutch words with their Frisian equivalents. The assistants will add additional linguistic information. Then, the lexicographers will complete the lexical entry by providing the Dutch examples, collocations, idiom, and word families with equivalents or translations. Finally, the assistants will check the spelling and references in the entries and add multimedia information, if any. The editors will discuss the edited entries before publication.

2.6 Preparation for online release

We will take the approach applied by the ANW (Tiberius and Schoonheim 2014, section 2.5.). With approved entries, the column marked 'ready for online' in the database is ticked. Every three months, a new version of the dictionary will be released, and changes can be made in between.

Newly edited entries will be added to the dictionary every three months. Based on the feedback submitted to the mailbox, published entries may be completed or adjusted. The altered entries will be put online as well.

2.7 Afterlife

We propose keeping one editor available for additions and improvements of the dictionary after completion of the project. We have not decided upon the way we will handle version management yet.

3   Time span of the different phases

Include a table indicating the time spans for the individual phases.

Phase	Duration
2.1 Preparation	2013-2014
2.2 Data acquisition	2014
2.3 Computerisation	2014
2.4 Data processing	2014-2019
2.5 Data analysis	2014-2019
2.6 Preparation for online release	2014-2019
2.7 Afterlife	2019

Table 1 Process phases of dictionary project and their time span

4   References

Duijff, Pieter (2000), *Juridisch Woordenboek Nederlands-Fries*. Groningen: Martinus Nijhoff, Ljouwert: Fryske Akademy.

Gouws, Rufus H. (2009), 'Dictionaries as Innovative Tools in a New Perspective on Standardisation'. In: Henning Bergenholz et al (eds.), *Lexicography at a Crossroads*. Bern: Peter Lang AG, pp. 265-284.

Kuip, Frits van der (2010), 'Dictionaries and their influence on language purification in minority languages. The case of Frisian.' In: A. Dykstra and T. Schoonheim (eds), *Proceedings of the XIV Euralex International Congress*. Ljouwert: Fryske Akademy-Afûk, pp. 1497-1504.

Niestadt, Jan (2009), 'De ANW-artikeleditor: software als strategie'. In: Egbert Beijk et al (eds.), *Fons verborum. Feestbundel voor prof. dr. A.F.M.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie*. Leiden/Amsterdam: Instituut voor Nederlandse Lexicologie/Gopher BV., pp. 215-222.

Popkema, Anne Tjerk (2010), 'State of the Art of the Lexicography of European Lesser Used or Non-State Languages.' In: A. Dykstra and T. Schoonheim (eds), *Proceedings of the XIV Euralex International Congress*.Ljouwert: Fryske Akademy-Afûk, pp. 65-98.

Tiberius, Carole en Tanneke Schoonheim (2014: to appear), 'The *Algemeen Nederlands Woordenboek* (ANW) and its Lexicographical Process'. In: Vera Hildenbrandt (ed), *Der lexikografische Prozess bei Internerwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/2014).