Language Web for Frisian

Hindrik Sijens, Anne Dykstra

Fryske Akademy, Doelestrjitte 8, 8911 DX Ljouwert / Leeuwarden, The Netherlands
E-mail: hsijens@fryske-akademy.nl, adykstra@fryske-akademy.nl

Abstract

The Fryske Akademy has developed a web portal, the Frisian Language Web, which consists of an online spell checker, a machine translator ('Oersetter') and a dictionary portal. These three applications will make a unique language tool for native speakers and learners of Frisian to help them to write proper Frisian. An important part of the Language Web will be a standardized word list of Frisian. The standard word list will be incorporated into a database underlying the spell checker. This database also contains a list of non-standard words that are linked to standard forms. This makes it possible to use the spell checker to guide users who write non-standard Frisian towards use of the standard language. The 'Oersetter' service will be a statistical machine translator based on a bilingual Dutch–Frisian parallel corpus. The dictionary portal will consist of existing, relatively recent, lexicographic material. Paper dictionaries and terminology lists were digitized, xml-parsed and linked to each other. The dictionary portal gives access to a wealth of information not (easily) accessible in the paper counterparts of the various dictionaries. By the different nature of the individual dictionaries, the user can draw on a wealth of lexical material. He has access to the portal through two languages, Dutch and Frisian. The dictionary portal will be the starting point of a new project: an online bilingual dictionary Dutch–Frisian.

Keywords: Standard wordlist; Language Web; Dictionary Portal; Frisian Language

1. Introduction

One of the themes of the fourteenth Euralex congress, which was held in Leeuwarden in August 2010, was the lexicography of lesser used and non-state languages. Keynote speaker Anne Popkema reported on a survey that was conducted by the Fryske Akademy about the state of the art of the lexicography of these languages (Popkema, 2010). The questionnaire contained several questions about the lexicographic output in a given region. With regard to the level of diversity of lexicographic output, West Frisian[1] got eight out of the maximum of ten points. The fact that Frisian had a monolingual dictionary and several bilingual ones yielded a good classification. There was some reason to be proud to see Frisian nestled between much larger minority languages such as Catalan and Basque. But in terms of the level of use of modern technology in lexicographical practice, the situation was not so rosy. Frisian and

[1] The term West Frisian is used to distinguish it from North Frisian, the variant of Frisian which is spoken in northern Germany (Schleswig-Holstein). In the remainder of this paper, we will use the term Frisian, to refer to West Frisian.

Friesland received poor results in terms of the availability of online dictionaries, but had satisfactory results in terms of dictionary writing software and the use of an electronic corpus. Frisian clearly lagged behind compared to languages like Catalan, Basque, Friulian and Welsh. It is absolutely desirable that the Frisian language performs better in terms of use of modern technology in lexicographical practice. The first steps to reach a higher level were taken in the past two years: the development of 'Taalweb', a language web for Frisian. This facility consists of an online spell checker, a machine translator called 'Oersetter' and a dictionary portal.

The core of 'Taalweb' is a newly created standard wordlist of the Frisian language. In this paper we will outline 'Taalweb' for Frisian. Furthermore, we would like to introduce our next project: a new and sophisticated online Frisian–Dutch dictionary.

2. Friesland – Frisian Language

Friesland is a province in the Netherlands, with 650,000 inhabitants. It is a bilingual province, about 54% of its inhabitants have Frisian as their mother tongue and about 65% are able to speak the language. About 25% of its inhabitants have Dutch as their mother tongue. Several other vernaculars are spoken by about 10% of its inhabitants. Frisian is a Germanic language and historically it is the closest extant language to English. Recent figures on language use show that 85% of the inhabitants of Friesland are able to understand Frisian and about 75% of the inhabitants of Friesland are able to read it. Almost 64% claim they are able to speak Frisian well. Only 10% are able to write Frisian well, about 18% quite well, while some 70% are unable or almost unable to write in Frisian (Taalatlas, 2011).

Dutch is the official language of the Netherlands. It is used as the first language in formal domains such as administration, education, commerce, and the media. Frisian is the second official language of the Netherlands, but the language is used more intensively orally than in writing. The main reasons for Frisians to use Dutch and not Frisian as a written language are that Dutch has a higher status, and the spread of writing competence in Frisian is insufficient.

3. Standard Wordlist

Like most lesser-used and non-state languages, Frisian encounters difficulties in developing a standard. Because there exists no long and extensive tradition in written language, and because of the fact that there are different coexisting dialects, Frisian has no fixed standard. Consequently there are quite a few frequent dialectical differences in the written language and therefore also in dictionaries. One example of this is the Frisian word *giel* (yellow) which is pronounced as /gi.əl/ in the northern part of Friesland and as /ge:l/ in the southern regions. As the word is pronounced and written in two different ways, there are two entries in the dictionaries: *giel* and *geel*. Another example of variation in the spoken language, which we also find in the

written language, is the paradigm of the irregular verb *gean* (to go). The first person past tense comes in three forms: *ik gie nei hûs*, *ik gyng nei hûs*, *ik gong nei hûs* (I went home). All forms occur in the written language.

To give an idea about how many possibilities there are for some frequently used adverbs, take for example the word *eigentliken* (actually, in fact, really). The existing dictionaries recorded twelve variants:

- ***eigentliken***
- ***eigentlik***
- *eigentliks*
- *eigenliken*
- *eigenlik*
- *eigenliks*
- *einliken*
- ***einlik***
- *einliks*
- *einken*
- *eink*
- ***eins***

On the basis of morphological principles and frequency counts we have chosen four of these forms to be included in the standard wordlist: ***eigentliken***, ***eigentlik***, ***einliks***, ***eins***.

But not only is this variation or these dialectical differences, such as like *gie*, *gong* or *gyng*, part of the language, but Dutch-isms are too. Frequently used Dutch words such as *lui* (lazy) or *gebeure* (to happen) are often included in the dictionaries, together with their proper Frisian equivalents *loai* and *barre*.

Of course, dialectical variation demonstrates the richness of a language, but also creates uncertainty for hesitant users and doubting language learners. What form should they choose: *giel* or *geel*, *ik gie*, *ik gong* or *ik gyng*, all correct Frisian forms? The same can be applied to the occurrence of Dutch-isms in Frisian like *lui* and *gebeure*.

It can be difficult to choose between sometimes obsolete but correct Frisian words like *loai* en *barre* and contemporary, frequently-used Dutch-isms *lui* and *gebeure*. This doubt regarding correct usage is rooted in a lack of education and routine in writing Frisian. Frisian only became a compulsory school subject in the second half of the last century. In addition, because this obligation applied only to primary schools and because written Frisian in daily life plays a minor role, most Frisian people are not proficient in writing their own language. Language learners, as well as native speakers, are insecure in their language use; they fear to make mistakes. Even language professionals such as journalists, editors, translators and novelists experience these kinds of problems. Since the lack of a standard is felt to be an

obstacle to the use of written Frisian, language professionals uttered a desire to standardize the language. At the same time, the policy of the provincial government of Friesland is to promote written Frisian. The desire to standardize Frisian is in accordance with provincial policy. The provincial authorities therefore asked the Fryske Akademy to compile a standard wordlist of Frisian.

The existing spelling system proved to be quite complex and inconsistent. The Fryske Akademy suggested a moderate spelling reform to the provincial parliament. With a solid description of the spelling rules as a starting point, the next step was to extract a list of words from the existing language corpus and dictionaries. This basic list of 145,000 lemmas had to be edited, because it contained duplicates, homonyms, dialect forms, misspelled forms, Dutch-isms and obsolete words. With the help of a specially designed database, which contains Frisian words with their morphological structures, the individual paradigms were automatically generated, with a considerable degree of success.

Another hurdle was to develop criteria to choose the standard forms. In order to create consensus, the standard forms are usually taken from the two main dialects of Frisian. In deciding which variant should be the standard, frequency plays a role, but frequency is not always decisive. In some cases we have chosen the historical lexicalized form of a word instead of the historically correct form. In other cases, the criterion distance played a role. The form most remote from the Dutch equivalent was preferred. For instance, in the case of *giel* versus *geel* mentioned above, the chosen standard form is *giel*, because this form is different from the Dutch form *geel*.

Analogy as a criterion also played a role. The form *read* (red) is realized as *read* and with d-deletion: *rea*. However, in inflected forms like *reade flagge* (red flag) the /d/ is always written and pronounced. Therefore we have chosen *read* as the standard form.

This standard wordlist is a reliable tool for anyone who wants to write Frisian. It is a benchmark for the language and a basis for the language technology products that are part of 'Taalweb'.

4. Spelling Checker

The standard wordlist is the core of a new spelling checker tool for Frisian. The history of Frisian spelling checkers began in the early nineties of the last century. A word list derived from the then existing dictionaries and databases was implemented in WordPerfect, at that time the most common word processor. In the late nineties, the Fryske Akademy together with the Dutch language technology company Polderland, developed a spelling checker for Microsoft Word. Ten years later the same team created an electronic language assistant for Microsoft Office, consisting of two bilingual dictionaries, a spelling checker and an option to correct and improve

texts, called 'Taalhelp' (Language Help). The production of these spelling checkers and tools was supported by a grant from the provincial government of Friesland.

Due to the changes in new releases of Microsoft Word, the tools were no longer compatible with Office 2010. The need for a new spelling checker has since been increasingly felt. Because it is provincial policy to support the use of written Frisian, the province financed the development of a new spelling checker. The new spelling checker is a plugin compatible with Microsoft Word, but it can also be accessed online.

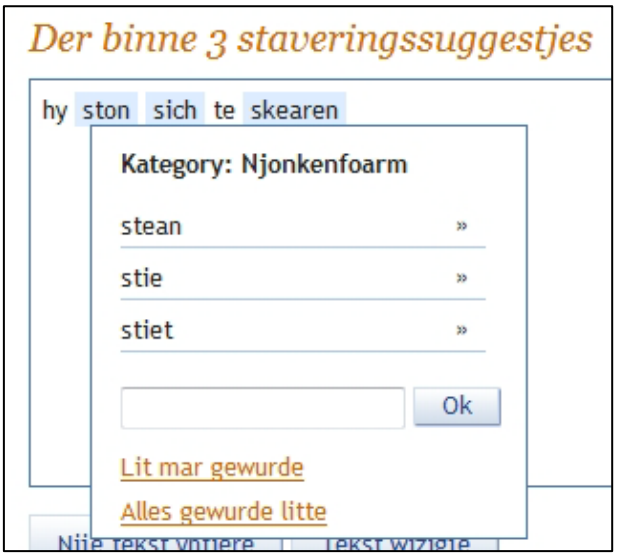

Figure 1: hy **ston** sich te skearen

Figure 2: hy ston **sich** te skearen

An example illustrates the design of this new tool. In the Frisian sentence *hy ston sich te skearen* (He was shaving himself), the spelling checker highlights the verb *ston*

and the pronoun *sich*. *Ston* is a dialect form of standard Frisian *stie* (stood), which is suggested by the spelling checker.

The reciprocal pronoun *sich* is marked as Dutch-ism. In Frisian the pronouns *him* (him) or *har* (her) should be used and in this context, *him* is the most likely.

On the back end of this unique language tool we have stored a complex system of alternative, non-standard forms and Dutch-isms, all linked to the preferred standard form. Whenever an incorrect or a non-standard form is encountered by the spelling checker, the author will receive suggestions to improve and correct his text.

However, solving one problem is creating another. The spelling checker correctly marks an alternative form like *ston* (pret. 1 sing.) as 'variant' of *stie*. The verb *wurde* (to get, to become) however has a standard paradigm form *wurde* in the present tense which is identical to a variant form in the past tense.

	StandardLem	StandardPara	Variant1
2	stean		
4		stean	ston
5		stiest	stonst
6		stiet	ston
7		stean	steane
8		stie	ston
31	wurde		
32		wurd	
34		wurde	
35		waard	wurde
37		waarden	wurden
38	wurd		
39		wurdsje	
41		wurden	
42		wurden	

Figure 3: paradigm of verb *wurde*

Since the spelling checker is not a grammar checker, the non-standard form *wurde* cannot be marked in the same way as *ston* has been marked in the previous example. However, the user must be drawn to the fact that he may have typed a non-standard form. But as long as there is no grammar checker available, we use a practical solution for this shortcoming.

When a user types a sentence like *hy wurde lilk* (he became angry), using the variant form *wurde* instead of *waard*, the spelling checker marks *wurde* and alerts the user: this is a standard form, but it can also be a non-standard, dialect form. The form *waard* is proposed, but if the user deliberately chooses to write dialect forms, he can

ignore the suggestion. And of course, if the user has typed a sentence in the present tense, for instance *wy wurde lilk* (we become angry) he also can ignore the suggestion.

This problem does not occur only in verbs, but also with homonyms. The numeral *alve* (11, eleven) has a non-standard form *elf*. But *elf* is also a noun which refers to a figure that appears in fairy tales and fantasy films.

As already mentioned, the wordlist represents the standard language, but this does not imply that the non-standard forms are always incorrect. Non-standard word forms still can be Frisian word forms. And the author can deliberately choose to use variants because they belong to his dialect or personal language. But it is to be expected in future that the standardized words will increasingly displace non-standard forms in written Frisian.

Figure 4: *waard* and (non) standard *wurde*

It is not our intention to rebuke the Frisian writing people in a pedantic way, or to discourage them from writing in Frisian. One of our aims is to guide people from their own (local) variant to the Frisian standard language. Moreover, the standard will be prescribed in teaching and strongly recommended in official language. And it is our expectation that writers, editors and journalists will also use this new list.

5. Machine Translator

Another feature of 'Taalweb' is a statistical machine translation system called 'Oersetter'. The system is able to translate from Dutch into Frisian and Frisian into Dutch and is intended to help non Frisian speaking people to understand Frisian. It is also a nice and easy way to create a basic translation, which can subsequently be edited with the other features of 'Taalweb'.

'Oersetter' has been developed at the Radboud University Nijmegen. The translation system is built around the open-source, phrase-based SMT software Moses. The Fryske Akademy has compiled a Frisian–Dutch parallel corpus. After sentence-alignment, the corpus comprised a total of 44,503 sentence pairs, containing 701,782 words of Frisian and 673,277 words of Dutch, including punctuation marks. The monolingual corpus used to create a Frisian language model consists of 594,975 sentences and 10,043,516 words, making it considerably larger than the parallel corpus. The Frisian portion of the parallel corpus has also been included in the corpus that was used for the language model. The corpus contains texts from 1980 onwards. Though the FA tried to cover as many domains as possible, a major part of the corpus inevitably consists of literary texts. A more detailed description of the background of the machine translator can be found in Van Gompel et al. (2010). While testing the translation system, the results were satisfactory and encouraging. Frisian text generated with this translation system may be spell checked to see if it is in accordance with the standard wordlist.

6. Dictionary Portal

The third part of 'Taalweb' consists of a dictionary portal. As already indicated, the state of affairs concerning online lexicography in Friesland was not sufficient.[2] Back in 2010, the bilingual dictionaries Dutch–Frisian and Frisian–Dutch were partially available online. The interface provided only translations of headwords. Contexts, idioms, multi-word expressions and proverbs were absent. Unfortunately, the extensive monolingual dictionary, which was published in 2008, was not accessible online. The 25 volumes of the scholarly Dictionary of the Frisian Language were put online at the Euralex Congress in 2010.

The Fryske Akademy used custom-made dictionary writing software, which consisted of a simple text editor and a BRS/search database. It was a full-text database and information retrieval system which used a fully-inverted indexing system to store, locate, and retrieve unstructured data (Sijens and Depuydt, 2010). Furthermore, a non-tagged language database was available for dictionary compilation purposes. This corpus was established in the preceding decades and contained at that time some 24 million words. The available electronic lexicographic products were digitized versions of paper dictionaries. It is needless to say that we are not dealing here with proper electronic lexicography.

There was a lack of online dictionaries and there was also the desire to establish a new Dutch–Frisian dictionary. The most recent bilingual Dutch–Frisian dictionary was published in 1985, so it is rather outdated. Besides that, it contains too few examples to provide good, accurate and modern translations in Frisian. In order to

[2] Data are taken from the questionnaire response for Frisian, cf. Questionnare, 2010.

fill the existing gap, a new project was conceived: a dictionary portal. This new online service contains several Frisian lexicographic products compiled in the years 1984 to 2008: a Frisian–Dutch dictionary (1984) with 56,000 entries, a Dutch–Frisian dictionary (1985) containing 53,000 entries, a juridical dictionary Dutch–Frisian (2000), which has 13,000 entries and finally a monolingual dictionary (2008) with 70,000 entries.

In addition to the dictionaries, the portal also contains a number of bilingual terminology lists ranging from administrative terms, through food terminology to terminology of school subjects such as geography, biology and physics. Newly compiled lists with terminology for these domains fill several lexical gaps.

The basic idea behind the dictionary portal was that linked information fields of the joint dictionaries and lists would provide much more useful information to the user than a stand-alone, digitized paper dictionary. For this purpose, the following information fields in the dictionaries were made searchable: headword, translations, idiom, synonyms and proverbs. Not every dictionary contains all fields, but that is hardly a problem. The bilingual dictionary Frisian–Dutch for instance lists more than 1,800 proverbs, a comprehensive list containing the most common Frisian proverbs. If a user is looking for a proverb, then what this dictionary provides will be sufficient to the user and the fact that the juridical dictionary does not deal with proverbs is no problem.

The recently published monolingual dictionary obviously lacked the field 'translation'; however, since 'translation' is the main objective of the portal, we had to add an extra field with Dutch keywords to that dictionary.

One of the functions of the portal is to bridge the gap between the old bilingual dictionaries and a new Dutch–Frisian dictionary. The 1985 outdated Dutch–Frisian dictionary often lacks modern words that belong to the domains of computer science, modern media and sports. The dictionary portal is a cluster of lexicographical products which covers a period of almost thirty years, from 1984 until 2008. Often, when the old dictionaries fail to give a translation for a modern concept, the more recent ones offer a complement. The 1985 Dutch–Frisian dictionary has an entry for *kompjûter* (computer) but not a single compound with *kompjûter-*. The 2008 monolingual dictionary additionally offers 23 compound words with *kompjûter*.

The dictionary portal provides more translations and examples than a stand-alone online bilingual dictionary. Where the Dutch–Frisian dictionary of 1985 has its limitations, the linked dictionaries of the portal offer much additional information and many more possibilities. Take for example the Dutch adverb *vliegensvlug* (very quickly, at top speed). This one-word expression has no one-word equivalent in Frisian. The Dutch–Frisian 1985 dictionary translates the headword *vliegensvlug* with three multi-word expressions:

- *fleanende hurd*
- *mei kûgelsfeart*
- *as de reek*

While searching the entire database, including the field 'idiom', yields more hits:

- *as de duvel - vliegensvlug*
- *op in giseldraaf rinne - vliegensvlug draven*
- *dat giet der koers troch - dat gaat vliegensvlug, razendsnel.*
- *gean, rinne, fleane, jeie as it spoar - vliegensvlug gaan, lopen, draven, rijden.*
- *it giet, rint as it spoar - het gaat vliegensvlug*

All these matches are from the Frisian–Dutch dictionary of 1985, taken from the field of idioms. When translating from Dutch to Frisian, these additional alternatives for *vliegensvlug* can help people who want to write in Frisian, to create more varied and better texts. The standardized wordlist is envisaged in 2013. The lexicographic products that are part of the portal contain many words and variants that are not part of the standard language. Therefore, these works all have to be adapted to the standard wordlist once it will be official.

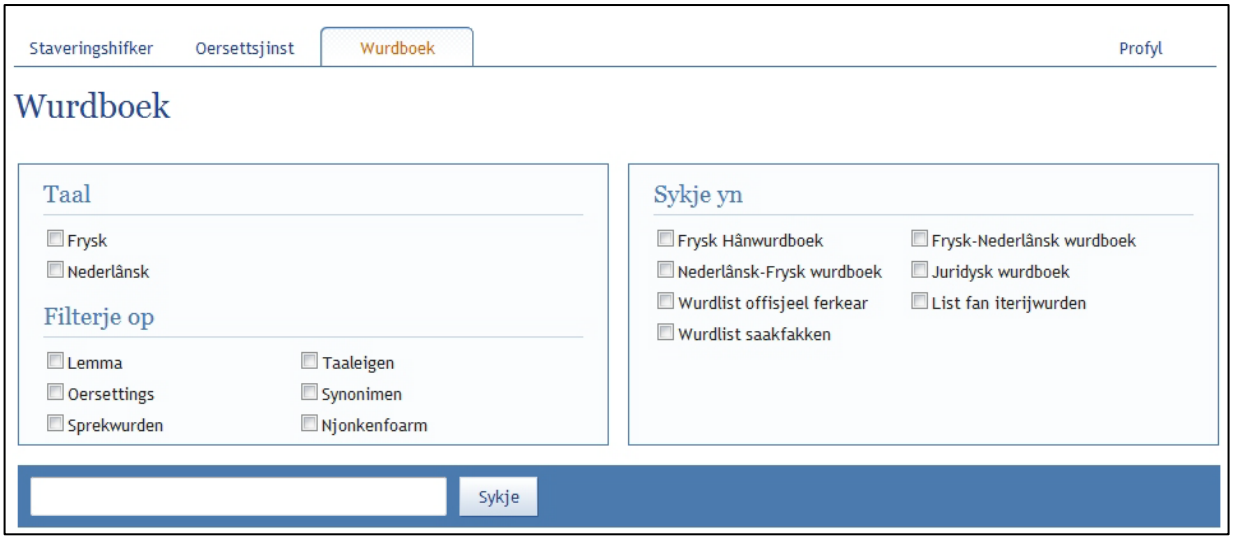

Figure 5: Opening screen 'Dictionary Portal'

7. Future Dutch–Frisian Dictionary

The state of affairs of internet lexicography for Frisian has greatly improved with the completion and launch of the dictionary portal. However much information and new possibilities the portal offers, it will not be able to offer the same as a completely new online Dutch–Frisian dictionary can. This new dictionary will be a lexicographical database tailored to the needs of the user. To this end we will perform a study of

recent literature on both online bilingual dictionaries and dictionary use.

The target audience for this dictionary consists of learners and native speakers. The needs of two user groups have to be satisfied: Firstly, non-Frisian-speaking people will use it for translating Dutch into Frisian. Secondly, for native speakers the dictionary will be employed as an aid to using proper Frisian. Frisian speakers often have had too little mother tongue education, resulting in a lack of knowledge of their own language. In order to serve both user groups, the dictionary will offer them about 70,000 Dutch headwords with their standard Frisian equivalents. Its microstructure will contain many examples, multi-word expressions, phrases and idioms that will enable the users to produce proper and varied Frisian.

This project provides new opportunities to compile an up- to-date lexicographic database with a user-friendly interface. The new dictionary will be part of the Frisian Language database, a database system intended to open up eight centuries of Frisian. A demo version of the Frisian Language database can be accessed at http://tdb.fryske-akademy.eu/tdb.

8. Conclusion

For a small language community like Frisian, it is difficult to create a good lexicographical infrastructure. In his Euralex keynote lecture, Anne Popkema stated 'Factors like magnitude of the language community and governmental recognition will be of influence on what medium a lexicographer chooses, since such factors for a considerable part determine the quintessential factor for any lexicographical endeavour: funds.' (Popkema 2010: 87). In some way, this also applies to Friesland. The Fryske Akademy has only a small lexicographical staff at its disposal. As a scientific research center the academy is required to conduct high-quality research. This has yielded a scholarly dictionary of Modern Frisian. At the same time the academy is required to use the acquired knowledge about lexicography for the benefit of Frisian society. Therefore it is obvious that the Fryske Akademy should produce dictionaries and tools for the community within which it is part. With the financial support of the provincial government, it is possible to develop the required lexicographical infrastructure in cooperation with fellow institutes, universities and language technology supplied by IT companies. The 'Taalweb' with the dictionary portal is a step forward on this path. We hope to integrate the various tools in such a way that, for example, a user will be able to go from a misspelled form to the correct one via (selected) dictionary information.

Or, when offered automated translations, the user will be able to call up the relevant dictionary entries, so as to improve the automated suggestions. We like to think that even in its present state the 'Taalweb' will offer much to the professional user of Frisian and that it will be a quite useful tool for language learners, whether or not in the context of a language course.

9. References

9.1 General:

Adamska-Sałaciak, Arleta (2013). Equivalence, Synonymy, and Sameness of Meaning in a Bilingual Dictionary. *International Journal of Lexicography* 26(3), pp. 329-345.

Fuertes-Olivera, Pedro A. and Sandro Nielsen (2012). Online Dictionaries for Assisting Translators of Lsp Texts: The Accounting Dictionaries. *International Journal of Lexicography*, 25(2), pp. 191-215.

Gompel, M. van, A van den Bosch, A. Dykstra (2013). Oersetter: Frisian - Dutch Statistical Machine Translation. In P. Boersma, H. Brand and J. Spoelstra (eds.) *Philologia Frisica anno 2012. Lêzings fan it njoggentjinde Frysk Filologekongres fan de Fryske Akademy op 13, 14 en 15 juny 2012*. Leeuwarden / Ljouwert: Afûk - Fryske Akademy (forthcoming).

Gouws, Rufus H. (2013). Contextual and Co-Textual Guidance Regarding Synonyms in General Bilingual Dictionaries. *International Journal of Lexicography*, 26(3) pp. 346-361.

Ilson, R. (2013). The Explanatory Technique of Translation. *International Journal of Lexicography*, 26(3), pp. 386-393.

Kwary, D.A. (2012). Adaptive Hypermedia and User-Oriented Data for Online Dictionaries: A Case Study on an English Dictionary of Finance for Indonesian Students. *International Journal of Lexicography*, 25(1) pp. 30-49.

Popkema, A.T. (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, (Leeuwarden, 6-10 July 2010)*. Ljouwert: Fryske Akademy - Afûk, pp. 65-98.

Questionnaire (2010). *Questionnaire concerning lexicography of European lesser used languages. Q026 West Frisian* (unpublished).

Sijens H. and K. Depuydt (2010). Wurdboek fan de Fryske taal / Dictionary of the Frisian Language Online: New Possibilities, New Opportunities. In A. Dykstra and T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, (Leeuwarden, 6-10 July 2010)*. Ljouwert: Fryske Akademy - Afûk, pp. 726-732.

Taaltatlas 2011. *De Fryske taalatlas 2011. Fryske taal yn byld*. Ljouwert / Leeuwarden: provinsje Fryslân.

9.2 Dictionaries:

Boersma, P. / K.F. van der Veen (1984-2011). *Wurdboek fan de Fryske Taal / Woordenboek der Friese Taal*. Ljouwert / Leeuwarden: Fryske Akademy. Accessed at: http://gtb.inl.nl.

Duijff, P. (2000). *Juridisch Woordenboek Nederlands - Fries, met een index Fries - Nederlands*. Groningen / Leeuwarden: Martinus Nijhoff / Fryske Akademy.

Duiff, P. en F.J. van der Kuip (2008). *Frysk Hânwurd- boek*. Leeuwarden: Fryske Akademy / Afûk.

Visser, W. (1985). *Frysk Wurdboek 2, Nederlânsk - Frysk*. Leeuwarden: A.J. Osinga Uitgeverij.

Zantema, J.W. (1984). *Frysk Wurdboek 1, Frysk - Nederlânsk*. Leeuwarden: A.J. Osinga Uitgeverij.