Open Access to Frisian Language Material

Eduard Drenth, Pieter Duijff, Hindrik Sijens

Fryske Akademy, Box 54, 8900 AB Leeuwarden (NL)

E-mail: edrenth@fryske-akademy.nl; pduijff@fryske-akademy.nl; hsijens@fryske-akademy.nl

Abstract

The Fryske Akademy has a long history—since 1938—of developing printed Frisian dictionaries and word lists, usually with Frisian and Dutch or Dutch and Frisian as the source and target languages, respectively. In the 1990s, the Akademy also began working on digital language resources for Frisian: a language database, various digitized dictionaries, a digital preferred vocabulary for Frisian and an Online Dutch–Frisian translation dictionary.
This paper briefly describes the available digital language resources and how access to them can be improved by means of a yet-to-be-developed application programming interface (API). The Fryske Akademy has three primary user groups in mind: language users, linguists and developers. A list of superlemmas will be compiled to link the information in the different systems.
Several examples are used to illustrate the requirements demanded of the API. Underpinning all this are the questions that might be asked by the three user groups of the language resources. Sections 5 and 6 describe the work and projects that are required to implement the API. The final section outlines a roadmap for potential future developments.

Keywords: linguistics; API; service; corpora; dictionaries

1. Introduction

The Fryske Akademy (FA) has a number of digital language resources, but these are largely independent of one another. In addition, some cannot be accessed by the public from outside the FA, despite the Akademy's aim to make its products available through open access wherever possible. Taking the needs of its target groups as the starting point, the FA plans to use an application programming interface (API) to provide access to data in the language resources. This paper aims to show how the FA will serve its target groups via the API. The API will not be discussed in detail here; instead, we will use examples to demonstrate how the API can be used to retrieve information from different data sources in a coherent way. Key principles for the API are standardization of the interface, and ease of access and service provision for users.

Before discussing the technical provisions and requirements that the API must satisfy, we first describe the language functionalities at the FA that will underpin the development of the API. We then identify the target groups we need to serve: language users, linguists and developers. We describe how these groups are currently utilizing our resources and the options we will offer in the future for making digital language material accessible for language users, researchers and developers.

2. Current language resources

2.1 Preferred vocabulary

Frisian, the second national language of the Netherlands, is a minority language with a limited written tradition, even within the province of Friesland where it is the native language. Frisian spelling was officially established for the first time in 1879 and it was not until after the Second World War that these spelling rules were officially adopted—in a slightly modified form—by the province of Friesland. Standard Frisian did not develop until the latter half of the nineteenth century; much later than, for example, Dutch. The standard language has been recorded in dictionaries and teaching resources during the past 120 years. A preferred vocabulary, which is essentially a list of standard forms (Taalweb.frl), has been made available online by the FA since 2015. For a detailed description of the preferred vocabulary, see Duijff (2016).

Since the development phase of Frisian, it has been common practice in written Frisian to accept different dialect variants alongside one another. Even though increasingly fewer variants are to be found in Standard Frisian dictionaries and vocabularies, standard Frisian continues to display greater variation than Dutch (Breuker, 2001; Duijff, 2008; 2016; Duijff & Van der Kuip; 2017). This variation also applies both to dialect forms and spelling variants in the preferred vocabulary. Because Frisian has acquired a growing role within education and as a written language, this has sparked a need for a list of standard or preferred forms, which the provincial government subsequently commissioned. In 2014 the FA created a database of preferred forms, in which the different variants are linked to the respective preferred forms (see Figure 1).

The database underpinning this vocabulary currently contains 96,146 lemmas, whose sources are the lists of lemmas for various Frisian dictionaries, supplemented by recent material from a range of sources. Of the 96,146 lemmas in the database, 85,730 can be labelled as standard forms (89.2%) and 10,146 (10.8%) as variants of these forms. In addition to lemma forms, the database provides word information in the form of word type, paradigm information and hyphenation. This database of standard forms is already being used in an application, namely a spelling checker (see Sijens & Dykstra, 2013: 96-99).

The preferred vocabulary is stored in an access database, which comprises several tables that are linked via IDs. The main tables are ‘lemma’ and ‘paradigm’. In addition to a column with the lemma form, the lemma table has columns with part-of-speech information and preferred form marking, etc. The paradigm table contains, in addition to a column with the paradigm forms, a column with hyphenation forms and a column where the form can be marked as the preferred form.

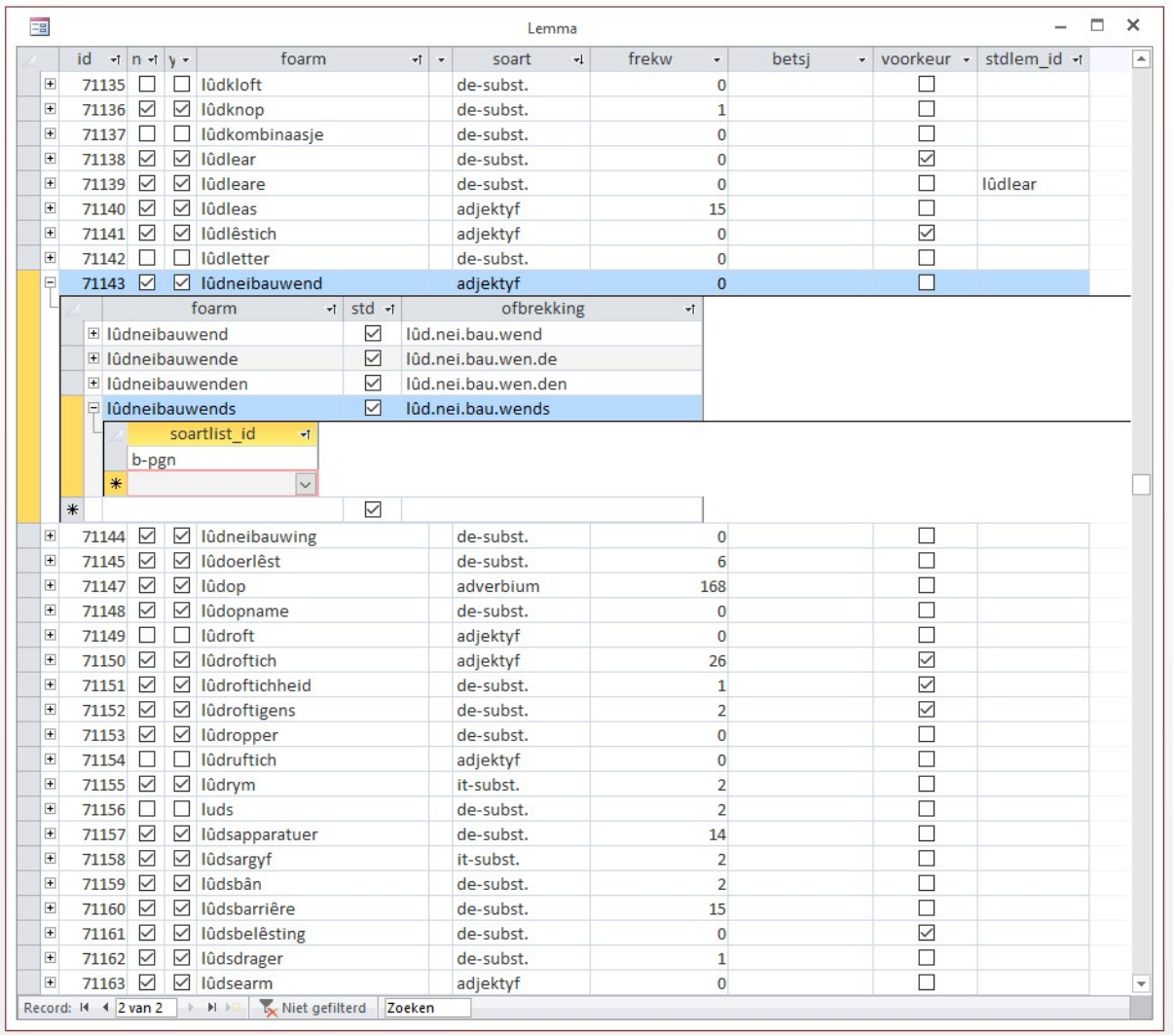

Figure 1: Screen of the preferred vocabulary database

2.2 Digital dictionaries

Since its establishment, the FA has compiled several dictionaries of Frisian. They are almost all bilingual, with mostly Frisian and Dutch alternating as the source and target languages. The most frequently used and most comprehensive translation dictionaries are still Zantema (1984), with 55,000 lemmas, and Visser (1985), with 45,000 lemmas. The historical/academic dictionary WFT (1984-2011) is also a bilingual dictionary, in the sense that Dutch is used to describe the Frisian language material. The most recent comprehensive desk dictionary with 70,000 lemmas is the monolingual Frysk Hânwurdboek/FHW (2008). Together with other dictionaries, these desk dictionaries can be consulted online at *Taalweb.frl*. The WFT can be consulted and searched online at Gtb.inl.nl (Depuydt et al., 2017). All these dictionaries were first developed as paper dictionaries and were only later made available online to language users.

2.3 Online Dutch–Frisian Dictionary

To meet the need for a modern, contemporary Dutch translation dictionary, the FA has begun compiling the Online Nederlands–Fries Woordenboek ('Online Dutch–Frisian Dictionary'/ONFW). The ONFW is an online production dictionary that takes modern standard Dutch as its source language and the standard Frisian equivalent as its target language. The dictionary will present not only the meaning and use of words and phrases, but also grammatical information. The dictionary will appear in parts from 2018 to 2022, after which it will continue to be updated and expanded (Duijff & Van der Kuip, 2017). For the source language, the ONFW will draw on the language corpus of the Algemeen Nederlands Woordenboek (ANW), an online dictionary of contemporary standard Dutch in the Netherlands and Flanders that describes Dutch vocabulary since 1970 (Schoonheim & Tempelaars, 2010: 718).

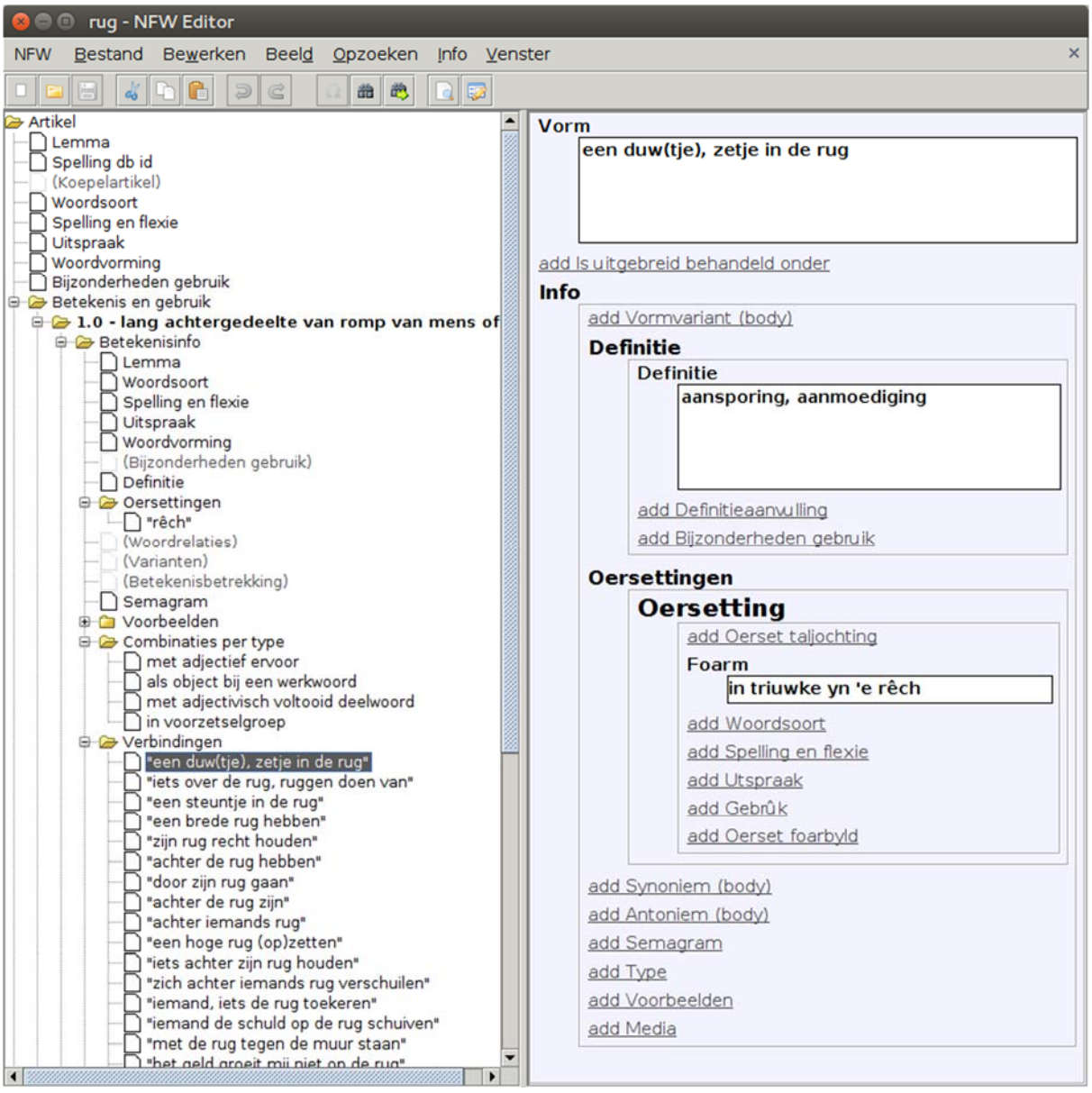

Figure 2: DWS screenshot of one page

The ANW, which is still being compiled, can be accessed at anw.inl.nl. The dictionary writing system (DWS) for the monolingual ANW has been modified so that it can be used for the bilingual ONFW. Figure 2 gives an idea of the DWS for the ONFW.

Using DWS enables editing of XML to conform a schema; below a snippet of the schema is shown.

```
<?xml version="1.0" encoding="UTF-8"?><xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema" version="1.0">
    <!-- xmlns:vc="http://www.w3.org/2007/XMLSchema-versioning" vc:minVersion="1.0"
vc:maxVersion="1.0" -->

    <xs:element name="Oersettingen">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="Taljochting" type="xs:string" minOccurs="0"/>
                <xs:element ref="Oersetting" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:element name="Oersetting">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="OersetTaljochting" type="xs:string" minOccurs="0"/>
                <xs:element name="Foarm" type="xs:string"/>
                <xs:element ref="Woordsoort" minOccurs="0"/>
                <xs:element ref="SpellingEnFlexie" minOccurs="0"/>
                <xs:element ref="Utspraak" minOccurs="0"/>
                <xs:element ref="Gebrûk" minOccurs="0"/>
                <xs:element name="OersetFoarbyld" type="xs:string" minOccurs="0"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
```

The bilingual DWS can also be used for other translation dictionaries that have Dutch as the source language. With some modifications, the system has already been made suitable for a yet-to-be-developed bilingual online dictionary of Dutch–Stellingwerfs. Stellingwerfs is a Saxon language variety spoken in the southeast of the Dutch province of Friesland and northwest of the Dutch province of Overijssel.

The ONFW will consist of a MySQL database and a Java application. The key feature of the database is a field with an XML that complies with an XML Schema. The database also contains user information, logging, status information for articles and workflow information. The XML can be edited using the Java application and the work status can be updated. The XML contains detailed information about Dutch entries and their Frisian translation, including spelling and inflection, word forms, pronunciation, combinations and fixed expressions.

2.4 Language databases

The FA has various Frisian text corpora containing data from the period 500–2017. The main ones are the corpus of Old Frisian (500–c. 1550), Middle Frisian (c. 1550–1800) and Modern Frisian (1800–2017). In addition to these three databases of exclusively written material, there is also a corpus of spoken Frisian, compiled from the period 2002–2006. Additional contemporary spoken material is currently being collected and an old spoken corpus and will be made available once more. In this contribution, we confine ourselves to written Frisian.

2.4.1 Corpora

The Old Frisian corpus comprises texts from the entire Old Frisian period until about 1550. It is a closed corpus with about 323,000 tokens. The material is for the most part linguistically annotated (lemmatized and tagged with part of speech). The Middle Frisian corpus contains all surviving Frisian texts from the period 1550–1800. This closed corpus is linguistically annotated and contains 488,000 tokens and 19,000 lemmas.

The corpus of Modern Frisian comprises a selection of written texts from the nineteenth and twentieth centuries and contains approximately 25 million tokens. The nineteenth-century part contains a small selection of prose written at that time. Efforts have been made to ensure that the twentieth-century part of the corpus is as representative as possible to provide maximum coverage of the Frisian vocabulary (Dykstra & Reitsma, 1995: 63). The corpus is not linguistically annotated.

2.4.2 Web interfaces

There are various interfaces that provide access to the corpora. Three of them can be used via the internet, and one can only be used internally within the FA for copyright reasons. The interfaces were all developed at different times, using different techniques and with different aims. The oldest interface gives the option of searching the various corpora by word form, possibly with the help of wildcards. Figure 3 gives an impression of a search with results in the oldest corpus. The results are presented in a concordance that can be ordered alphabetically by the word occurring to the left or right of the keyword.

Searches can also be made in sub corpora. A distinction is made between the three different language phases for Frisian: Old, Middle and Modern Frisian. Frisian in the period 1900 can in turn be broken down into different periods distinguished by clearly identifiable spellings. This interface was developed in 1998, primarily for the lexicographical projects that the FA was, and continues to, work on.

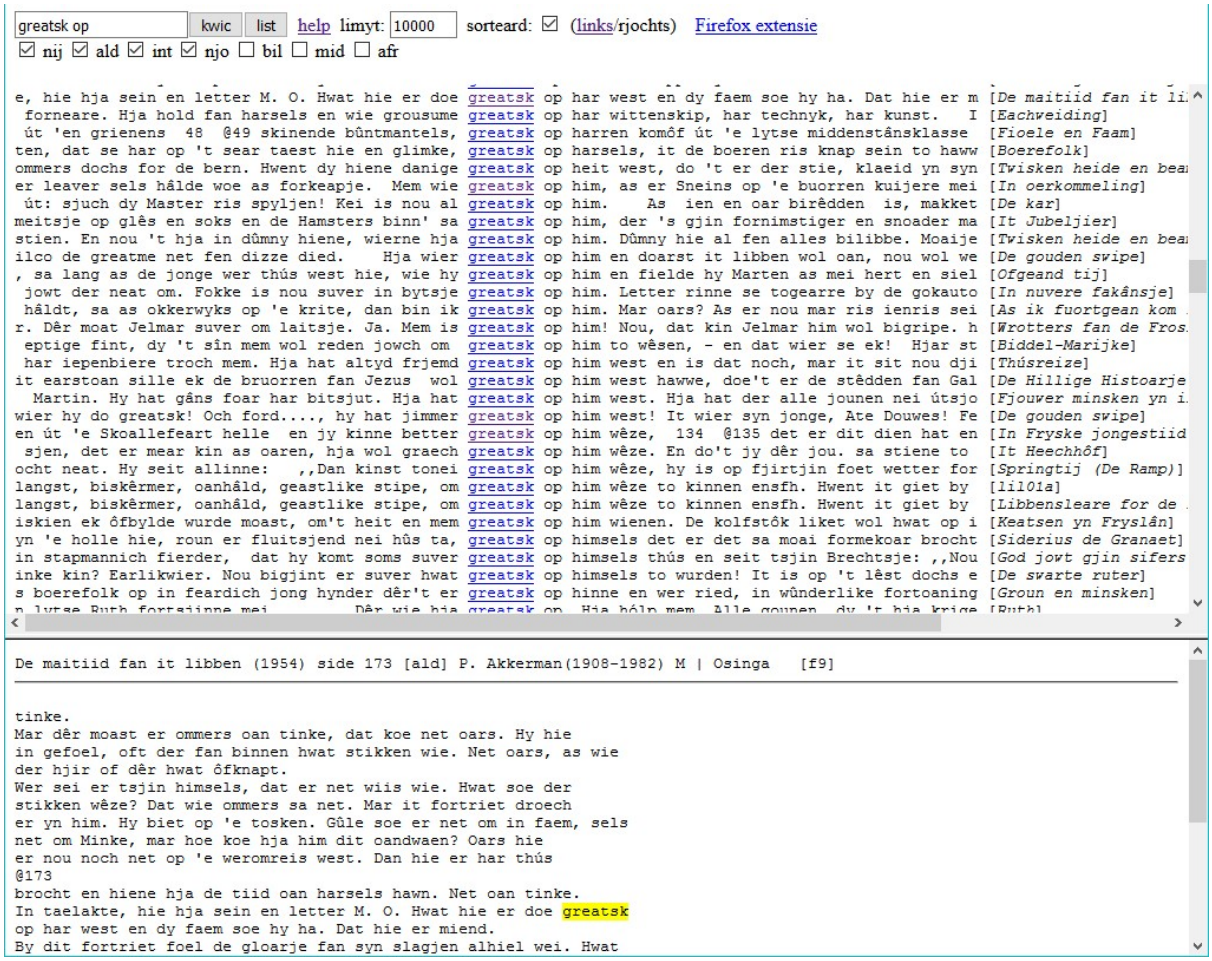

Figure 3: Search and results in the oldest corpus

The second interface provides access to the linguistically annotated corpus of Middle Frisian from the period c. 1550–1800, combined with corpus material from the earlier and later periods. Users can search by lemmas and word forms, possibly with the help of wildcards. They can opt to have the results presented in a concordance or in a list of word forms. This corpus is linked to a bibliography of secondary literature. Another special feature is that geographical information that is linked to lemmas can be downloaded. With a designated account, the database is freely accessible via http://pc245.fa.knaw.nl:8020/tdbport/. It was developed in 2002 to give easy access to Middle Frisian material, and with the option of adding more corpora. This interface continues to fulfil a need, namely searching by word form, or by morphological, diachronic and paradigm information.

A third interface, developed in 2009, makes the Old, Middle and Modern Frisian material accessible to a wider audience. With this interface, users can access sources directly or can search by lemmas. The results can be shown in KWIC (keyword in context) view, with an option to show the sources. Zantema (1984) is also integrated and his bibliographies are linked to this interface, which is freely accessible at

http://tdb.fryske-akademy.eu/tdb/. Unique in the language database are the integrated scans of medieval manuscripts, the integrated Old Frisian dictionary (Hofmann & Popkema, 2008), clickable words in the corpora and linked secondary literature. In the interface, it is not possible to search on word form or with wildcards; only lemmas are accessible. The offered language information in the results is restricted to word type and period.

None of the existing language databases can be searched by linguistic information or by other meta-information that is present. All versions are interactive and there is no interface to conduct searches from other applications.

3. Target groups

We have identified three different user groups for FA's digital language resources: professional and non-professional language users, linguists and developers.

3.1 Language users

The preferred vocabulary is used in education—in schools and within adult education—and serves as a foundation for the creation of teaching materials. Journalists, authors and publishers use it as a reference work when editing and correcting publications. Officials, lawyers and staff in public sector institutions use it to assist in document writing. In all these instances, the vocabulary serves as a lexicographical resource to enable users to write Standard Frisian. Users can check which variants are acceptable. The vocabulary also provides information about the basic inflection and hyphenation of lemmas. The preferred vocabulary is the basis of a spelling checker for spelling errors and typos, and to check for standard forms and Dutchisms. The ONFW is a lexicographical Dutch–Frisian translation resource for a target group made up of language learners and native speakers of Frisian. Language learners are primarily interested in finding translations and grammatical information, while native speakers also use the dictionary for text production, such as searching for the right word forms, collocations and idioms. Users will consult the language databases to find contexts for a particular word form, information about Old Frisian manuscripts, etc. Interested individuals can look at facsimiles of manuscripts.

3.2 Linguists

Linguists utilize the digital language resources of the FA, although the three resources offer differing possibilities.

The preferred vocabulary gives researchers only a limited range of options. It presents a preferred form for Standard Frisian. Researchers who want to find out how standardization has developed can view the vocabulary as the modern-day final stage

in this process. For example, they can investigate whether and to what extent the vocabulary differs from the one in current Frisian dictionaries. They can also explore which Frisian dialects have contributed preferred forms to the standard language, or they can use the vocabulary to check which articles go with nouns, since each noun is accompanied by the correct article.

The database underlying the preferred vocabulary provides researchers with more options. The grammatical information included with each entry, for example, is an invaluable source of information. Researchers studying inflection variation in spoken Frisian can check which inflections verbs take compared to the standard. This variation is on the increase, mainly as a result of the dominance of Standard Dutch, particularly among younger generations. The database also contains many grammatically correct variants of the standardized inflection.

The language database is used for a wide range of linguistic research. Firstly, the Old, Middle and Modern Frisian texts in the database can be used to compile lexicographical resources. To date there is no lexicographical access to the Old and Middle Frisian language material. The language database can be used to describe word forms and the grammatical and semantic properties of lemmas. The link between KWIC and manuscripts or text editions means that it is easy to illustrate the lemma descriptions with text fragments linked to the source. The language database offers almost unlimited opportunities for the study of Frisian grammar. Because images of the Old Frisian text sources are linked to the texts in the language database, philologists can work on text editions. Thanks to the availability of texts from all three stages of the Frisian language, the database can be used to conduct detailed research on Frisian language change over the centuries. An example of one such study is Versloot (2008), which describes vowel reduction in fifteenth-century West Old Frisian, on the basis of material in the language database.

Like the language database, the ONFW can be used for grammatical research. The inclusion of grammatical information with the Frisian translations is a feature of the ONFW. This information is generated from the preferred vocabulary database. The bilingual Dutch–Frisian dictionary will enable researchers to make lexicological and semantic comparisons of the two languages. Because the dictionary includes many examples of idiomatic usage, it is an ideal tool for studying the use of idioms in Frisian and the differences with Dutch.

3.3 Developers

In the future, the idea is that stakeholders within education or culture, for example, will be given opportunities to develop applications on the basis of the API, such as massive open online courses (MOOC) or apps for mobile devices. Examples are apps with lexicographical applications (translating or looking for definitions), apps that check and assess texts for style or grammar, or apps with spelling exercises and

language games (puzzles, Scrabble-type games). Other applications involving the API include serious games or applications in healthcare (care robots that understand and speak Frisian).

4. API

4.1 CLARIN

CLARIN offers solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. With the API, we hope to achieve at FA level what CLARIN is seeking to achieve at supra-organizational level: standardized access to digital language material for teaching, research and other purposes. The API will also serve as a springboard for the development of services within the CLARIN infrastructure.

4.2 Design

Figure 4 below shows what the API will look like, with the main data sources, the target groups and the subdivision into editing and production environments.

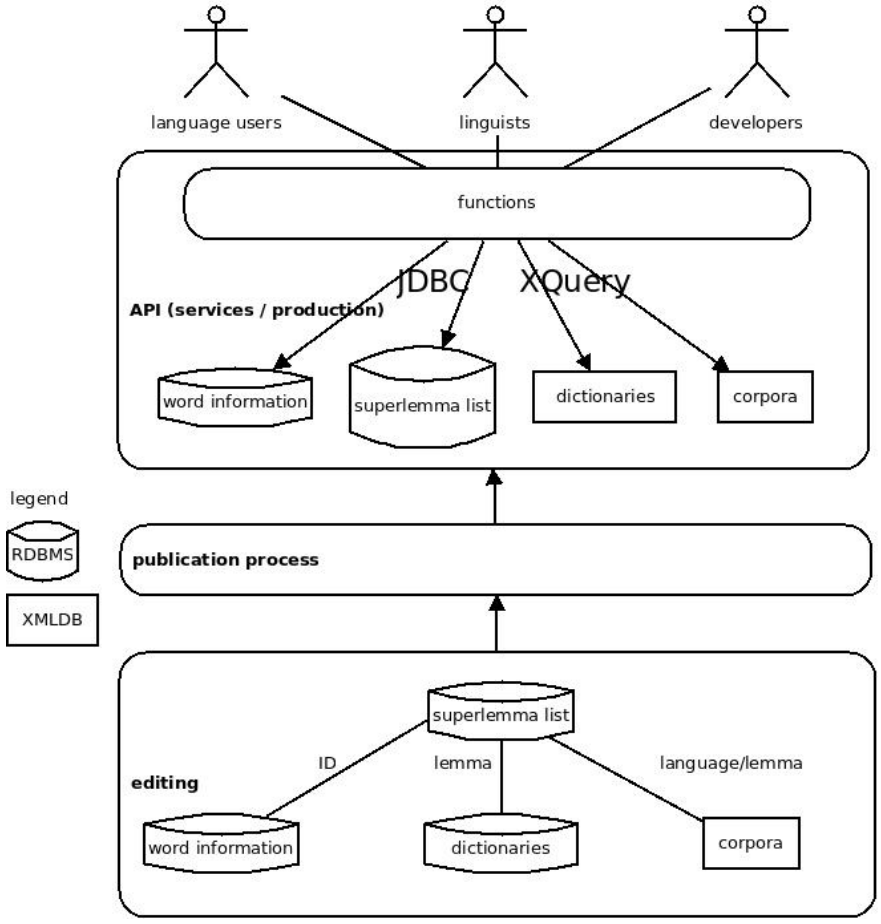

Figure 4: Diagram of the API

The principles are: information in one place, the separation of editing and production, and good support for work on the data.

Word information: Information at word level, including paradigm, morphology, preferred forms, word type.

Dictionaries: translation dictionaries from Dutch, based on the ANW.

Corpora: TEI-encoded texts with numerous possibilities for linguistic coding

Superlemma list: List of superlemmas with corresponding lemmas in a language category (Old, Middle and Modern Frisian, etc.)

4.2.1 Superlemma list

The superlemma list will play a key role. 'Superlemma' refers to an abstract lemma form in modern Frisian spelling to which the Old, Middle and Modern Frisian forms/lemma forms are linked. For example, the lemmas *sjitte* (Modern Frisian), *sjiette* (Middle Frisian) and *skiâta* (Old Frisian) are linked to the superlemma *sjitte* ('to shoot').

The aim is to arrive at, via a superlemma from the systems, information in another system. Lemmas in a particular language category are included under each superlemma. Superlemmas can be searched via a modern Frisian lemma or a lemma in another language category, together with that category. Each superlemma is also assigned an ID so that it can be selected directly. The available language categories will make up a list (for example: runen, old_frisian, bildts) to be published in the API.

Superlemmas for extinct words from older language phases will be reconstructions based on etymological patterns.

4.2.2 Links

The systems will be managed separately. The links between the systems are the information they contain that also appears in another system. See Table 1.

A possible consequence of these separate links is that systems could become 'out of sync'. This is particularly true of the superlemma IDs. Checks will be built into the management environments to help prevent links from no longer being valid.

The various data sources will be managed separately in the management environment. There will be a minimal relationship between the systems, just enough to gather information in the production environment. 'Word information' and 'superlemma list' will be linked via the ID that uniquely identifies each superlemma. 'Corpora' and

'superlemma list' will be linked via language category and lemma, while 'dictionaries' and 'superlemma list' will be linked via the new Frisian lemma.

word information	superlemma list	Each superlemma has an ID that can be included with an entry in 'word information'
corpora	superlemma list	Lemma annotation together with an annotation for language category can be found under a superlemma
dictionaries	superlemma list	A Frisian lemma in a dictionary can be found in a superlemma
dictionaries	word information	A Frisian lemma can be found in a dictionary with an entry in 'word information'

Table 1: Overview of links between systems.

4.2.3 Publication

Information will enter the production environment through a publication process, whereby data in the systems—with the exception of the dictionaries—will in principle be transferred one to one. Because of optimizations, users may choose to save certain data in the production environment twice. In the case of the dictionaries, the XML of the articles is removed from the database field in question and put into eXist-db. The publication process is also the place where transformations to standardized formats, etc. will be made.

4.2.4 Service

In production, the service offers functionality for the development of applications. See Section 4.3 for some detailed examples of functions within the service. The service uses standard technical links—JDBC and XQuery—to access information from the underlying systems. These technical interfaces provide access to all the information in the data sources and offer expressive query options. The technical interfaces can be used directly, but this requires extensive knowledge of the interface and the underlying data. The service offers a more user-friendly portal to information in the data sources. Filtering, sorting, pagination and other important functionalities that users require when querying data sources are built into the API.

4.3 Functions

In the use cases below, we will demonstrate how questions from target groups will be answered by means of a set of functions in the API. The functions will be defined in such a way that they can be used in different use cases. To maintain the focus on functionality, we have not included filtering, sorting, pagination and other general functionalities such as error handling in the examples.

4.3.1 Language users

Use Case: translation

For a Dutch word and its Frisian translation, a user also wants to find the inflection and pronunciation for that translation, as well as examples of contexts in which the Frisian translation is used. For this, the API offers the following functions.

Firstly, it must be possible to translate text from a language (in this case Dutch) into Frisian via a function. Input characters, possibly with wildcards, are used to search for matching Frisian lemmas. The result is a list of found Frisian lemmas, in which each found lemma is accompanied by the ID of the associated superlemma, the word type and the description. First, a search is made in dictionaries (ONFW in this case) for the Frisian translation of a text. This translation is then used to search in 'word information':

Signature: FrisianLemma* translate(text, language category)
Input: Text with wildcard support * and ? and a language category (Dutch in this case)
Output: 0 or more FrisianLemmas, with the superlemma ID, word type and description
Data used: 'word information' and dictionaries (ONFW)

Second, a function for retrieving inflection information on the basis of the superlemma ID:

Signature: Inflection getInflection(ID)
Input: superlemma ID
Output: Inflection
Data used: 'word information'

Third, a function for retrieving pronunciation information on the basis of the superlemma ID:

Signature: Pronunciation getPronunciation(ID)
Input: superlemma ID
Output: Pronunciation

Data used: 'word information'

Finally, a function is needed to show context information (KWIC) on the basis of a Frisian lemma. Searching for context information can be confined to a particular language category:

Signature: KWIC* getKWIC(lemma, language category)
Input: Frisian lemma, language category
Output: 0 or more KWIC showing text before the searched lemma, the lemma itself (or word forms of that lemma) and subsequent text.
Data used: corpora

Use Case: corpora

While searching the language database (TDB), a user finds a word form in the Old Frisian corpus that he cannot place. He therefore wishes to find a Dutch lemma for the word form. For this, the API offers the following functions.

Firstly, a function is needed to find superlemmas on the basis of a lemma in a particular language category (Old Frisian in this case). The principle here is that the word form in the corpus is annotated with the associated Old Frisian lemma. In the superlemma list, superlemmas are searched on the basis of the Old Frisian lemma (lemma + language category Old Frisian):

Signature: superLemma* findSuperLemma(lemma, language category)
Input: lemma and language category
Output: 0 or more SuperLemma, with ID and associated lemmas
Data used: superlemma list

The superlemma now has to be searched in dictionaries (ONFW) to find Dutch translations:

Signature: DutchLemma* translate(FrisianLemma)
Input: FrisianLemma, the Frisian lemma used to search for Dutch lemmas
Output: 0 or more DutchLemma, with meaning
Data used: dictionaries (ONFW)

The superlemma can be used to retrieve the new Frisian inflection in 'word information', for example, to compare it with the Old Frisian (and possibly Middle Frisian) inflection.

4.3.2 Linguists

Usage

The integration of the databases offers extensive opportunities for comparative research across time and space. With the help of the superlemma, information about lemmas can be selected in 'word information'. Dictionaries can be searched for Dutch translations of the lemmas, and in the TDB searches can be made in the corpora, for example by word form and their dialect distribution, or by linguistic information.

Researchers can also investigate, for example, the differences between separable verbs in Dutch and Frisian, in modern Frisian and Dutch and in older phases of these languages. A query in 'word information' will give a list of all separable and inseparable verbs with their paradigm, plus morphological information. The 'translation' field can be used to establish a link between this information and Dutch verbs in the ONFW. Paradigm and morphological information can also be retrieved from that database. Finally, the TDB can be searched for corpus evidence.

Functions in the API that support research are presented below. A function for finding all separable/non-separable verbs:

Signature: FrisianLemma* getVerbs(separable)
Input: Boolean separable
Output: 0 or more FrisianLemma, with the superlemma ID, word type and description
Data used: 'word information'

Next, a function for finding words with particular linguistic annotations. This function also supports separable verbs. The result contains the superlemma for the found words; this can be used to retrieve information in 'word information' and dictionaries. The linguistic annotations that are available for searches are published and updated in the API https://bitbucket.org/teibestpractices/linguistic-customization.

Signature: Result* find(text, linguistics*)
Input: Text with wildcard support * and ?; combinations of linguistic properties that are searched by
Output: 0 or more Results, showing found words in context, the superlemma for found words and metainformation on the corpus
Data used: corpora and superlemma list

As well as this function for retrieving results, there is also a function simply for counting:

Signature: CountResult count(text, linguistics*)
Input: Text with wildcard support * and ?; combinations of linguistic properties that are searched by

Output: The number of results and metainformation on the corpora
Data used: corpora and superlemma list

Researchers can also search corpora on the basis of information in 'word information'. An example is searching for neologisms, whereby 'word information' is searched for entries labelled 'neologism', possibly restricted to certain lemmas. The associated superlemmas are then retrieved. Under the superlemma are lemmas with a language category that can be used to search the corpora.

Signature: Result* findNeologisms(text)
Input: Text with wildcard support * and ?
Output: 0 or more Results, showing found words in context, the superlemma for found words and metainformation on the corpus
Data used: 'word information', corpora and superlemma list

5. Further development of data sources

5.1 Word system

The current access database for the preferred vocabulary will be transformed into a server database, such as MySQL. The database will be redesigned, bearing in mind the merging of information from the preferred vocabulary with information from other systems such as a morphological database. A management application will then be designed and built and a conversion will be written for converting data. In this conversion, linguistic terms will be converted into terms from linguistic-customization.

5.2 Online Dutch–Frisian Dictionary

The XML from the online dictionaries will be published to an XML database (eXist-db). This database will become the source in which searches will be made from the API via XQuery and/or REST. A website will also be generated for the ONFW so that people can engage interactively with the dictionary.

5.3 Language database

A new version of the language database is being developed. It is based on TEI XML, with a linguistic expansion based on universaldependencies.org (see linguistic-customization). We are thus opting for reputable, internationally supported open standards which enable digital publication with minimum effort and which offer a foundation for research. Tei-c.org makes this possible by choosing customization as a base. This occurs via One Document Does all (ODD), which will manage validation, support/editing support and presentation.

The XML contains information about the manuscript, such as author, repository, location, the manuscript text, linguistic annotations at word level and a reference to the superlemma list.

The Oxygen XML Editor is used for editing and offers support for TEI and the linguistic expansion.

eXist-db is used for storing and accessing the material. eXist-db offers the option of querying the manuscripts using the standard XQuery language. There are no restrictions here; all information present can be queried.

There is a need to generate a website with TEI Publisher for the corpora, where manuscripts, including scans, can be viewed, where the material can be searched by text, with KWIC results, and where manuscripts can be downloaded as PDF files.

The material in the language database is not always free of copyright. This will be taken into consideration, including technically via the availability element in TEI.

6. Implementation

Implementation mainly involves upgrading the current systems, setting up a management and production environment and publication processes, designing and building links (via superlemma) and designing and building the API. The steps in this implementation process will be set up as projects that will be assessed, prioritized and scheduled in relation to one another. At the very least, the building projects will involve versioning, dependency management and issue management. Ideally, we will also work with continuous build, with a test environment and with other solutions that are customary in a development process, for example Docker.

6.1 Projects

Table 2 provides an overview of the work required to implement the API. It does not include scope, prioritization, phasing, etc.

6.2 Service and support

An online help desk will be set up for language users, researchers and application developers. There will also be built-in options for reporting problems and suggestions and for monitoring their status. System monitoring will also be set up to maintain automated monitoring of system use.

Component	**Work**
Word information	Design and build data model in management screens
Word information	Migration and conversion of existing material from the preferred vocabulary and morphological database, etc.
Word information	Design and build technical interface
Dictionaries	Import information from the preferred vocabulary
Dictionaries	Design and build publication to eXist-db
Dictionaries	Design and build technical interface
Corpora	Design and build functions
Corpora	Design and build technical interface
Superlemma list	Design and build data model and management screens
Superlemma list	Design and build technical interface
API	Design and build functions, with input, output and error handling functionality
API	Build the implementation of the API
General	Set up publication processes, including automation

Table 2: Work to realize the API.

7. Conclusion

The information about functionality in this paper is based on information already contained in the databases. Meanwhile, the FA has a number of databases and language resources. Lexicographical tools are digitally available since the end of the twentieth century. In our paper the available databases of the Frisian language are described. The reader has been able to conclude that these databases can certainly be improved. The aim of the linguistic department of the FA is to expand the databases and to create, add and link more databases. We see it as in important task of the FA to optimize the databases and their use. Finally, we summarize this task in the following roadmap, divided into three important items: (a) expanding the databases, (b) including spoken language material, and (c) linking data.

- Expanding
    - The Modern Frisian corpus will be expanded to include texts from domains and genres that are currently underrepresented, supplemented by texts from social media, weblogs. Distribution over time is also out of balance: nineteenth-century Frisian, in particular, is underrepresented, and the period after 1990 also requires attention.
    - The WFT can be linked to the Modern Frisian corpus to make more corpus evidence available.
    - There is a long-held wish to create a WordNet-type lexical semantic database of Frisian.
    - On the basis of the Old and Middle Frisian corpora, lexicographical resources should be made for each of these two language phases.
- Speech
    - Spoken corpora must have a place in the landscape described here, and possibly in the API as well.
    - The preferred vocabulary will be supplemented in future by pronunciation information in the International Phonetic Alphabet (IPA) and by morphological information about the keywords.
- Linked data
    - Geographical and diachronic information are already present, but are not yet clear and unequivocal. In the new system, solutions will be sought to give both data types a good home, in data sources and the API.
    - Digitized dialect-geographical material could be linked to the various databases.
    - The corpora, in particular, contain information that can be made available as Linked Open Data (LOD). This could involve metadata information, such as author, location, year of publication and publisher, as well as information in the text, such as geonames and named entities. Through LOD, links can be made to other data sources, such as HISGIS (Historisch Geografisch Informatiesysteem / Historical Geographical Information System), which also works with LOD.

8. References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Breuker, P. (2001). The Development of Standard West Frisian. In H.H. Munske (ed.) Handbuch des Friesischen / Handbook of Frisian Studies. Tübingen: Max Niemeyer Verlag, pp. 711-721.

Depuydt, K., De Does, J., Duijff, P. & Sijens, H. (2017). Making the Dictionary of the Frisian Language available in the Dutch historical dictionary portal. In J. Odijk & A. van Hessen (eds.) CLARIN in the Low Countries. London: Ubiquity Press, chapter 13.

Duijff, P. (2008). Towards Standard Frisian in the Friesch Woordenboek. In M. Mooijaart & M. van der Wal (eds.) Yesterday's Words: Contemporary, Current and Future Lexicography. Newcastle: Cambridge Scholars Publishing, pp. 53-66.

Duijff, P. (2016). Towards Modern Standard Frisian. In I. Tieken-Boon van Ostade & C. Percy (eds.). Prescription and Tradition in Language. Establishing Standards across Time and Space. Bristol / Blue Ridge Summit: Multilingual Matters, pp. 532-549.

Duijff, P. & Van der Kuip, F. (2017). Lexicography in a minority language: a polyfunctional online Dutch-Frisian dictionary (in progress).

Dykstra, A. & Reitsma, J. 1993, De struktuer en de ynhâld fan 'e Taaldatabank fan it Frysk. It Beaken, 55, pp. 55-82.

Schoonheim, T. & Tempelaars, R. (2010). Dutch Lexicography in Progress: the Algemeen Nederlands Woordenboek (ANW). In A. Dykstra and T. Schoonheim (eds.), Proceedings of the XIV Euralex International Congress. Ljouwert: Fryske Akademy / Afûk, pp. 718-725.

Sijens, H. & Dykstra, A. (2013). Language Web for Frisian. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 93-105.

Versloot, A.P. (2008). Mechanisms of Language Change, Vowel Reduction in 15th Century West Frisian. Utrecht: LOT.

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

Dictionaries and websites:

OCDSE: *Oxford Collocations Dictionary for Students of English.* (2009). 2nd edition. Oxford: Oxford University Press.

Anw.inl.nl. Accessed at: http://anw.inl.nl. (9 May 2017).

FHW: Frysk Hânwurdboek. (2008). Ljouwert: Fryske Akademy / Afûk.

Gtb.inl.nl. Accessed at: http://gtb.inl.nl/ (9 May 2017).

HISGIS. Accessed at: www.hisgis.nl/ (26 May 2017).
Hofmann, D. & Popkema, A.T. (2008). Altfriesisches Handwörterbuch. Heidelberg: Universitätsverlag Winter.
Linguistic-customization. Accessed at: https://bitbucket.org/teibestpractices/linguistic-customization (12 April 2017)
ONFW: Online Nederlands-Fries woordenboek. In progress.
Taalweb.frl. Accessed at: www.taalweb.frl https://taalweb.frl/. (28 April 2017)
Tei-c.org. Accessed at: http://www.tei-c.org/index.xml/. (19 July 2016)
Universaldependencies.org. Accessed at: http://universaldependencies.org/. (15 May 2017)
Visser, W. (1985). Frysk Wurdboek. Nederlânsk-Frysk. Leeuwarden: A.J. Osinga Uitgeverij.
WFT: Wurdboek fan de Fryske taal / Woordenboek der Friese taal. (1984-2011): Ljouwert / Leeuwarden: Fryske Akademy.