EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

Frisian dictionaries, digitized from A to Z

Drenth E.[1], Sijens H.[1], Van de Velde H.[1,2]

[1] *Fryske Akademy, Leeuwarden (NL)*
[2] *Universiteit Utrecht, Utrecht*
edrenth@fryske-akademy.nl; hsijens@fryske-akademy.nl; hvandevelde@fryske-akademy.nl

Abstract
This paper approaches dictionaries as lexical resources with functions for target audiences, which benefit from a strictly defined data format, which means less work and improved interchangeability. Code generation in a reliable automated build process provides validation and documentation. Stable services provide functions that can be realized within the data format. The software can run straight away with a complete docker setup. In this way, creating a dictionary becomes primarily a matter of editing or converting data, for instance with an XML editor that supports editors by means of generated validation and documentation.

Keywords: Frisian; TEI; Universal Dependencies; eXist-db; dictionaries

1 Introduction

1.1 Background

Compiling a Frisian academic dictionary is one of the raisons d'être of the Fryske Akademy (FA). Nowadays, over eighty years since the institute's formation, the creation of Frisian user dictionaries is still a core activity. For the longest time, the process typically spanned several years and resulted in a paper publication. In the early days of digitization (around 1980), the dictionaries that were edited were entered into a database as ASCII text. For coding (for typesetting and other purposes), an at sign was added (e.g., @C = italic). The text was then stored in a full-text database. From this database, the text was generated and converted for printing. Later, the need arose for dictionaries that were digitally accessible. This need was initially met by using scripting languages to query the documents or convert them to queryable formats such as XML or a database. The next step was the introduction of the Fryske Akademy's first 'born digital' dictionary, the Online Dutch Frisian Dictionary (*Online Nederlands-Fries Woordenboek*, ONFW) (Drenth 2017). The ONFW is digital in design. However, the target format is not generic, and this interferes with standardized querying (and editing).

1.2 The next step: Dictionaries as datasets

The Fryske Akademy is now taking the next step in the development of its lexical resources and tools: defining a standardized format and developing generic applications for that format. With this step, dictionaries become lexical resources that can be continually edited and queried, rather than the static end product of an extended lexicographic process. The focus is on developing applications that meet the needs of user groups, and the data format plays a crucial role in this process.

2 Basic principles and preconditions

The major changes in the lexicographic process require a clear formulation of the basic principles and preconditions of the new structure. These are the basic principles and preconditions drawn up by the Fryske Akademy for the lexical infrastructure for the Frisian language.

2.1 Current editing environment

Maintaining the work processes linked to the current editing environment is a precondition. ONFW's editing environment is maintained in collaboration with the *Instituut voor de Nederlandse Taal* (Dutch Language Institute), which secures the lexicographic infrastructure for the Dutch language.

2.2 Serving a variety of end-users

The format and the solutions will be geared to different user groups, such as linguists, professional users (writers, translators, journalists, teachers, civil servants, lawyers), native speakers, and language learners (both L1 and L2 learners). This means that the format should be able to hold information in a variety of formats, handle varying degrees of detail, and make available both comprehensive and simple search options and search results.

2.3 More efficient editing

Lexicographers need to be able to edit lexical information intuitively. In doing so, they require drop-down lists, additional documentation, and validation of their input. A community of language users should be able to contribute in a simple way (citizen science). These contributions must then be identified, edited, and validated by professional lexicographers.

2.4 Quick and easy queries

A data format is a means for querying lexical resources. This requires functions that are stable and clearly defined in terms of input, output, and error handling. The work is greatly simplified if the format's structure and content are strictly defined, so that it is clear where specific information can be found. In addition to online queries, the data should be accessible offline using proprietary tools.

2.5 Ease of conversion

There are many initiatives in the field of lexicography, both open source and proprietary, such as TEI Lex-0 (Salgado 2019), Freedict, grammarly, and wordnet (see Section 4). The option to link up with these initiatives is an important asset and usually entails converting the format, dynamically or otherwise, to that of the target initiative. Such a conversion can be facilitated by (i) the use of a well-defined format, (ii) that is semantically clear and consistent, and (iii) content that does not contain hidden functions (e.g., a # in text with special meaning) to avoid having to carry out an error-prone analysis of optional content.

2.6 Sustainability

The solutions to be set up or developed should be maintainable with the least possible effort: taking no rare technical expertise, little time, and relatively little money. The work is to be carried out with proven methods and technology from the ICT industry. The format co-determines the options: with a leaner and simpler format, it will be easier to maintain the software solutions.

2.7 Adaptability

In the future, the current format may not be suited for new information. In that case, the format may have to be adapted. As a result, the data formats and their functions will start to diverge. A clear procedure must be put in place to avoid problems due to changes and to facilitate switching to later versions.

2.8 Open standards

The significance of open standards is not disputed, so the connection to and use of open standards is a given. The solution itself will also be made available to all, and wider use will be advocated.

2.9 FAIR Principles

The Fryske Akademy is an academic institution and as such adheres to the FAIR principles[1]. The compliance with the FAIR principles and open standards will pay off in terms of interoperability, acceptance, available tools, and the collaboration between parties.

3 Current standards

Based on the principles set out above, we have considered the following standards and infrastructures that could be part of the solution.

3.1 TEI

TEI[2] stands for Text Encoding Initiative; this is a well-established, widely used open guideline for encoding text. It is used primarily for the encoding and online publication of historical manuscripts, and it contains a module for dictionaries and support for linguistic annotations. TEI is comprehensive and designed to support a wide range of scenarios. As a result, there are often many options for encoding and few forced options. To meet the drawbacks of its wide standard, TEI has a powerful mechanism for customization: One Document Does It All (ODD). With ODD, the guidelines can be geared to specific applications by defining which components are used and how they are applied.

3.2 TEI Lex-0

TEI Lex-0 (Bański 2017) is an initiative intended to establish an open dictionary standard based on TEI that is better suited to digital processing. It is an ODD that is the result of a community process; it focuses on limiting the opportunities available in TEI. TEI Lex-0 is primarily intended as a format for existing dictionaries in order to improve interoperability.

3.3 Universal Dependencies

Universal Dependencies[3] (UD) is an open framework for the consistent linguistic encoding of text. UD contains guidelines for word type, morphosyntactic description, treebanks for many languages, and tools such as a part-of-speech tagger. As

[1]https://www.go-fair.org
[2]https://tei-c.org
[3]https://www.universaldependencies.org

such, UD provides a solid foundation for natural language processing.

3.4 Freedict

Freedict[4] is a technical open-source project containing many translation dictionaries in TEI format. Most dictionaries are flat and mainly focus on word-to-word translation. There are multiple applications for Freedict, especially for Android and Linux.

3.5 Ontolex

Ontolex (McCrae 2017) is an open semantic web model used to classify information (language/words) lexicographically. Semantic web technologies allow computers to analyse information and apply a form of artificial intelligence across various information domains. When lexicographic information is converted to, and made available in, the semantic web, it also becomes available for the applications that use artificial intelligence.

3.6 ELEXIS

ELEXIS[5] is not a standard but rather an infrastructure. Internationally speaking, the European Lexicographic Infrastructure (ELEXIS) (elex.is) is of great importance. It is an initiative intended to make linguistic data openly accessible and to make language resources available. There are several ways to connect to ELEXIS, for instance by converting existing material to TEI Lex-0 or Ontolex. Tools available from ELEXIS include Sketch Engine, a large corpus system; Lexonomy, an online editing and publishing environment for dictionaries; and Elexifier, for converting existing dictionaries to TEI Lex-0 or Ontolex.

4 Specifics

In this section, we outline the choices and approaches by which we have arrived at our solutions.

4.1 Choices

We have opted for a solution based on TEI and Universal Dependencies. These two international standards have been used by the Institute and its partners for many years. Both standards are actively maintained, and several matching tools are available, such as the TEI stylesheets, oxygen, udpipe, and teipublisher. The target audience and approach of the Freedict project is clearly different from creating professional dictionaries for more experienced language users and scholars. Therefore, the Fryske Akademy has not selected this project as a foundation for its dictionaries. The availability of Frisian in Freedict dictionaries is relevant, however, as Frisian is a low-resource language that is of little interest to commercial companies. Therefore, we will be supplying data to the Freedict project.
TEI Lex-0 is marketed primarily as a format that facilitates interoperability between lexicographic resources. It can also be used as a basic format for compiling dictionaries. TEI Lex-0 restricts the space allowed by TEI, but still offers a lot of freedom. This makes it less suitable for software development because it is not certain where information can be found within the data structure, nor how information can be recognized. We have decided to set up our own format, very similar to TEI Lex-0, but with more restrictions. This provides a better foundation for software development. It also makes it easier to convert the format from the editing environment. We have developed a conversion to TEI Lex-0 for interoperability with ELEXIS and others.
As a semantic web format, Ontolex is not a suitable format for editing or building a dictionary service. In time, lexicographic data will be published as Linked Open Data.

4.2 Approach

For the TEI customization, an ODD was developed in which the primary objective was to achieve a simple structure that could be properly validated. In the development stage, five different dictionaries were simultaneously converted to the target format. These five dictionaries are the ONFW, the Frysk Hânwurdboek (Duijff 2008), the Dutch-Frisian dictionary (Visser 1985), the Frisian-Dutch dictionary (Zantema 1984) and the legal dictionary (Duijff 2000). In addition, a REST service was developed for querying the target format. By running these development paths in parallel, we were able to test the target format and service in practice. During the development process, we carefully considered which components were generic. These components and their development processes were separately made available as open source, see section 5.

4.3 Components

Figure 1 visualizes current components used in lexicography at the Fryske Akademy.

Figure 1: schematic of components for lexicography. Generic parts are marked in green

[4]https://freedict.org
[5]https://elex.is

4.3.1 ODD with library

An ODD is an XML file in TEI format that can be used to capture a TEI customization. The resulting scheme roughly accommodates word forms with grammatical designations, homonyms, meanings, paradigms, synonyms, variants, examples, collocations, and proverbs. Translations can be included as well. In the meta-information, editors can indicate which of the following functions are supported.

- Formtranslation: translation of words
- Textsearch: search for words in text
- Synonyms: search for synonyms of words
- Variants: search for variants of words
- Compounds: search for compounds of words
- Pronunciation: search for the pronunciation of word forms
- Hyphenation: search for the hyphenation of word forms
- Usage: search for usage information regarding word forms
- Stress: search for emphasis in word forms
- Definition: search for definition of word forms
- Grammar: search for grammar including position regarding words
- Paradigm: search for the paradigm in headwords
- Examples: search for examples of words
- Collocations: search for collocations of words
- Proverbs: search for proverbs containing specific words

These functions are available in the REST service, see section 4.3.3.
Using the TEI stylesheets and some freely available tools, the ODD is translated into a validation file containing documentation. This translation is part of a tightly defined, repeatable Maven[6] build process. Once a version of the format has been approved, it is published in a globally available repository: Maven Central. The published version can be used in dictionary projects, for instance by editors who are aided in their editing environment by drop-down lists, documentation of items to be added, and the option to validate their work, or by ICT staff who write conversions and want to validate the result.

4.3.2 Exist-db

Dictionaries based on our method use XML, which has prompted us to store them in an XML database. We have opted for exist-db[7] [b]ecause exist-db has a long track record, is open source, standards-based, has an active community, and because the FA and its partners are familiar with exist-db.

4.3.3 REST service

In exist-db, a REST service was developed to query dictionaries. This service provides powerful, lucene-based search functions that can be used to search for translations of words, grammatical properties, pronunciation, examples, etc. (See the list under 'ODD with library'.) The results are presented in a simple standardized manner in Json for further processing, see table 1.

/translate?form=dag&lang=nl	/paradigm?form=wurkje&lang=fry	/synonyms?form=fyts~
	0: "pronoun.clitic"	

Table 1: example queries and results

4.3.4 Application

In addition to this service, a web application was developed in three languages, English, Dutch, and Frisian, suitable for computers and smartphones. Since user needs and wants may vary, this application is secondary to the REST service as far as we are concerned.

5 Dissemination

5.1 REST services

In accordance with the setup described above, four dictionaries are now available as REST services:

- the Frysk Hânwurdboek
- the Frysk Wurdboek Nederlânsk-Frysk
- the Frysk Wurdboek Frysk-Nederlânsk
- the Juridysk Wurdboek

[6]https://maven.apache.org
[7]https://exist-db.org

These services can be accessed at https://frisian.eu/dictionaries.

5.2 Test applications

For a first impression and testing purposes, web applications for these dictionaries are available through the following links:

- the Frysk Hânwurdboek: https://frisian.eu/dictionaries/fhwbapp
- the Frysk Wurdboek Nederlânsk -Frysk: https://frisian.eu/dictionaries/nfwbapp
- the Frysk Wurdboek Frysk-Nederlânsk: https://frisian.eu/dictionaries/fnwbapp
- the Juridysk Wurdboek: https://frisian.eu/dictionaries/jurwbapp

5.3 Software library

A software library for the solution is freely accessible at https://search.maven.org/search?q=a:teidictionaries. The library can be used to validate whether XML documents conform to the format that we have developed. It can also be used to convert XML into programmable objects and vice versa, and to convert XML into TEI Lex-0 with validation.

5.4 Exist-db extension

During the development phase, the need arose for additional features for exist-db: periodic synchronization of articles with exist-db, the ability to properly configure applications, and some search options. These features were donated to the exist-db community: see https://search.maven.org/search?q=a:exist-db-addons and https://github.com/eXist-db/documentation/pull/549.

6 Next steps

Figure 2 visualizes future developments in lexicography at the Fryske Akademy. The sections below describe this in more detail.

Figure 2: schematic of future developments in lexicography

6.1 Editing environment

The current situation is a good starting point for considering future developments. One of the first things we want to realize is a generic editing environment in which both lexicographers and volunteers can edit material, aided by documentation, drop-down lists, validation, and a simple workflow with the phases ‘under consideration’, ‘review’, and ‘approved’.

6.2 Communities

Lexicography is all about language, and language connects people, so, the FA wants to facilitate communities by developing opportunities for feedback.

6.3 Conversions

As mentioned above, a conversion to TEI Lex-0 has been developed, and a conversion to Ontolex will be developed at a later stage. In this way, we aim to connect with ELEXIS and Clarin.

6.4 Integrations

In the near future, this dictionary functionality of this solution will be integrated with other language solutions such as corpora, lexicons, text translation, and spell checking. A GraphQL-based interface is under development for this purpose.

7 In conclusion

Based on analyses of existing solutions and standards, using proven methods and techniques from ICT, we have realized a robust open solution for lexical resources. The solution greatly reduces the amount of work required for publishing lexical data. Moreover, it is a solid foundation for software development, and promotes integration with other initiatives in lexicography. Our solution is available for all languages, and is of particular interest to low-resource languages that cannot afford commercial support. We hope that our initiative will be applied more widely and that a community will emerge to continue the development of the solution.

8 References

Drenth E., Duijff, P., Sijens, H. (2017). Open Access to Frisian Language Material.
Duijff, P. (2000) Juridysk Wurdboek Nederlânsk-Frysk
Duijff, P., Van der Kuip, F., De Haan, R. and Sijens, H. (2008). Frysk Hânwurdboek
McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017). The OntoLex-Lemon Model: Development and

Applications.
Bański, P., Bowers, J., Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms.
Visser, W. (1985). Frysk Wurdboek Nederlânsk-Frysk
Zantema, J.W. (1984). Frysk Wurdboek Frysk-Nederlânsk

8.1 Websites

Apache lucene query syntax. Accessed at: http://www.lucenetutorial.com/lucene-query-syntax.html [2016].
CLARIN - European Research Infrastructure for Language Resources and Technology. Accessed at: https://www.clarin.eu/ [2018].
Docker. Accessed at: https://www.docker.com [2019].
Dictionary software. Accessed at: https://search.maven.org/search?q=a:teidictionaries [2020].
European lexicographic infrastructure. Accessed at: https://elex.is [2019].
Exist-db Xml database. Accessed at: https://exist-db.org [2018].
Fair principles. Accessed at: https://www.go-fair.org [2021].
Free dictionaries. Accessed at: https://freedict.org [2020].
Maven build and dependency management tool. Accessed at: https://maven.apache.org/ [2016].
R package for Tokenization, Tagging, Lemmatization and Dependency Parsing. Accessed at: https://www.rdocumentation.org/packages/udpipe/ [2020].
TEI guidelines. Accessed at: https://tei-c.org [2018].
TEI publisher. Accessed at: https://teipublisher.com [2018].
Universal Dependencies. Accessed at: https://www.universaldependencies.org [2019].

Acknowledgements
We want to thank our colleagues at the FA for their cooperation and patience. We would also like to thank the communities, especially those around TEI, UD and eXist-db, for thinking along, which improved the solution and made it more interoperable. Finally, we would like to thank the province of *Fryslân* and the Dutch government for providing the finances that made these developments possible via project PF-2017/172874 (ONFW) and 01785576 (*Taalweb*).