Taalportaal: an online grammar of Dutch and Frisian

Frank Landsbergen, Carole Tiberius, Roderik Dernison
Institute for Dutch Lexicology
P.O. Box 9515, 2300 RA, Leiden, The Netherlands
frank.landsbergen@inl.nl, carole.tiberius@inl.nl, roderik.dernison@inl.nl

Abstract

In this paper, we present the *Taalportaal* project. *Taalportaal* will create an online portal containing an exhaustive and fully searchable electronic reference of Dutch and Frisian phonology, morphology and syntax. Its content will be in English. The main aim of the project is to serve the scientific community by organizing, integrating and completing the grammatical knowledge of both languages, and to make this data accessible in an innovative way. The project is carried out by a consortium of four universities and research institutions. Content is generated in two ways: (1) by a group of authors who, starting from existing grammatical resources, write text directly in XML, and (2) by integrating the full *Syntax of Dutch* into the portal, after an automatic conversion from Word to XML. We discuss the project's workflow, content creation and management, the actual web application, and the way in which we plan to enrich the portal's content, such as by crosslinking between topics and linking to external resources.

Keywords: grammar portal, Dutch, Frisian

1. Introduction

The *Taalportaal* project will create an online language portal containing an exhaustive and fully searchable electronic grammar of Dutch and Frisian[1]. This grammar will consist of a large set of texts, called topics, on the three domains phonology, morphology and syntax for both languages. Besides the grammar module, the portal will contain an ontology of linguistic terms and an extensive bibliography. In order to serve the international scientific community, all content will be in English.

In this paper, we give an overview of the *Taalportaal* project. In section 2, we discuss the project's background and compare it to other language portals. In section 3, we present the structure and workflow of the project. Section 4 deals with the infrastructure for the authors who are contributing to the project. Section 5 discusses the implementation of the *Syntax of Dutch* into the portal, and section 6 deals with the actual web application.

2. Project background

2.1 Organization

The *Taalportaal* project started in 2011 and runs until the end of 2015. It is carried out by a team of authors, advisors and IT-personnel from four universities and research institutions: the Meertens Institute, the Fryske Akademy, Leiden University and the Institute for Dutch Lexicology. The latter is responsible for the technical infrastructure.

The main motivation for the portal is the fact that currently, no comprehensive scientific online grammar of Dutch and Frisian exists. Furthermore, the digital design of the portal enables interoperability between the linguistic categories of phonology, morphology and syntax on the one hand, and between the two languages on the other. It is expected that the portal's rich crosslinking will benefit these domains of research, which are now often studied in isolation.

Content for the portal is provided by a team of authors, who describe the grammars of the two languages in XML (see section 4). In addition, the complete works of the *Syntax of Dutch* (Broekhuis & Keizer, 2012; Broekhuis & Den Dikken, 2012; Broekhuis, 2013) are automatically converted from Word into XML, and added to the portal integrally (see section 5)[2]

The final version of the portal will be delivered at the end of 2015. As of February 2014, a beta version of the application is available online at www.taalportaal.org.

2.2 *Taalportaal* in relation to other portals

The focus of *Taalportaal* is on a thorough, linguistic description of the grammatical properties of Dutch and Frisian, meant for linguists in particular and scientists in general. Several sites exist on Dutch grammar, but they differ from *Taalportaal* in terms of content or target audience. *NEON* (http://neon.niederlandistik.fu-berlin.de/) offers comprehensive information on Dutch grammar, but focuses on language learners rather than linguists. A well known grammar for Dutch is the *Algemene Nederlandse Spraakkunst* (Haeseryn et al., 1997), which was originally printed as a book and has been retrodigitized (http://ans.ruhosting.nl). It is only available in Dutch and it does not contain information on Dutch phonology.

Albeit for German, the *Grammis* portal seems to be the most closely related to the goals of *Taalportaal*. *Grammis* (http://hypermedia.ids-mannheim.de/) has been developed by the Institut für Deutsche Sprache in Mannheim and contains an elaborate information system on German grammar, a grammatical dictionary, a list of

[1] Frisian and Dutch are the two official languages of the Netherlands. Frisian is spoken mostly in the north of the Netherlands in the province of Friesland (*Fryslân*) and has around 470.000 speakers.

[2] The completed volumes of the Syntax of Dutch are also available online at http://www.oapen.org/search?keyword=Broekhuis+syntax.

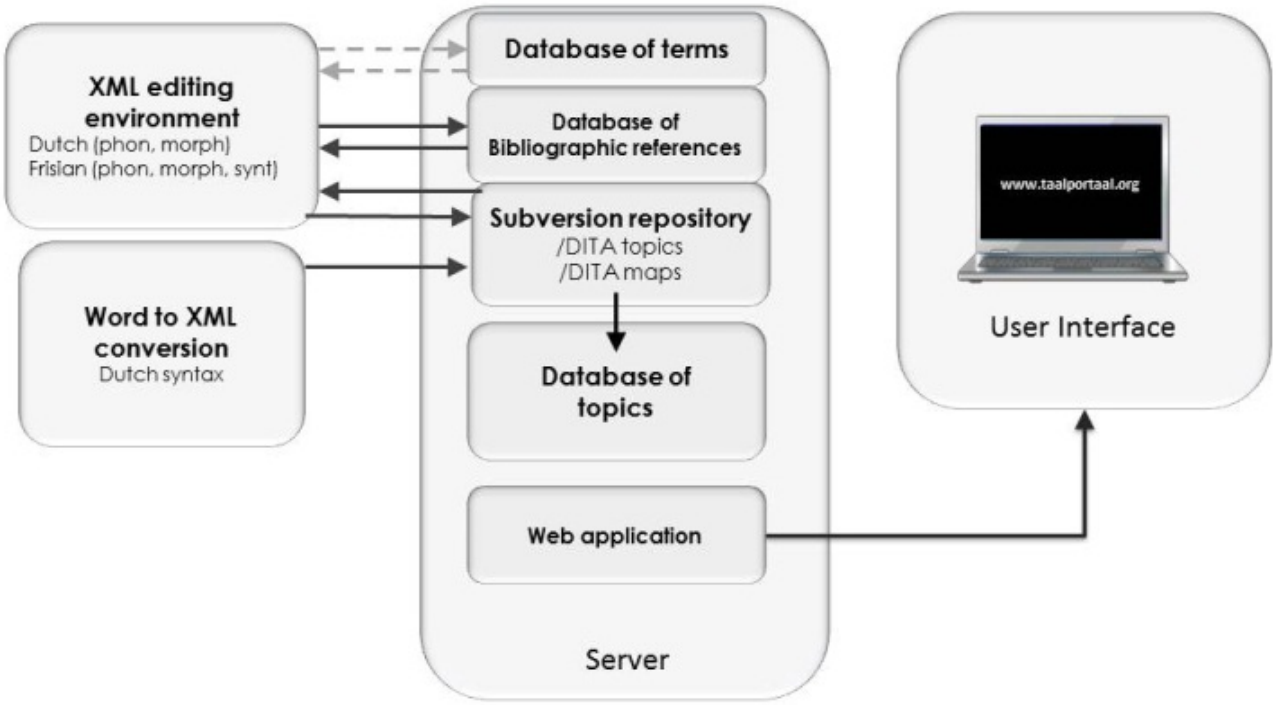

Figure 1: Workflow of the *Taalportaal* project

terms, a bibliography and a page with questions and answers on grammatical topics. Its content is in German.

Another aspect that sets *Taalportaal* apart from the aforementioned portal sites is that it contains parallel, crosslinked grammatical information on *two* languages, instead of one[3].

3. Project structure

Figure 1 shows the workflow of the *Taalportaal* project. The input for *Taalportaal* is formed by grammatical texts in XML. Most *Taalportaal* authors write directly in XML using the *Taalportaal* editing environment. The materials of the *Syntax of Dutch* follow a different route though, due to its genesis, and consist of Word documents which are automatically converted to XML.

The *Taalportaal* authors who are based at different institutes and universities throughout the country, store their grammatical texts in a central Subversion repository. This data is copied onto a local file system and forms, together with the database of terms and the database of bibliographic references, the core of the *Taalportaal* web application. The process from retrieving the data from the Subversion repository to displaying the data online is completely automated such that data updates can be realized at regular intervals.

4. Author infrastructure

The editing environment in which *Taalportaal* authors create their topics consists of the XML-editor Oxygen, which has been customized to the specific needs of the project. This editor contains a link to an extensive database of bibliographical references, that authors can use to insert entries in their texts, and to add or edit them. This database has been filled by merging the references from several textbooks on Dutch phonology, morphology and syntax.

Authors use a *Taalportaal*-XML schema, which is based on the DITA format (http://dita.xml.org) to enforce a topic-based approach to writing. The content of the portal is, where possible, taken from existing descriptions from Dutch and Frisian, and updated where needed. Since the original sources generally have a linear structure, they have to be rewritten to make them more suitable for the internet. This is done by splitting them into smaller, more or less independently accessible parts called 'topics', and by using a three-layered structure: (1) quick info, (2) read more, (3) extra. 'Quick info' contains a short introductory description of what the topic is about. 'Read more' contains the actual content. The layer 'extra' is used for additional information such as theoretical debates.

This three-layered structure also allows us to ensure a certain amount of content homogeneity which is further enforced by a number of structural restrictions to the XML-schema, such as a compulsory use of sections. On the other hand, the XML-schema is flexible enough to account for the different sorts of tagging wishes from the domains of phonology, morphology and syntax. An example is the structure of example sentences, which contains separate layers of analysis (table 1).

[3] There are currently plans for a virtual institute of Afrikaans, using the *Taalportaal* infrastructure and data.

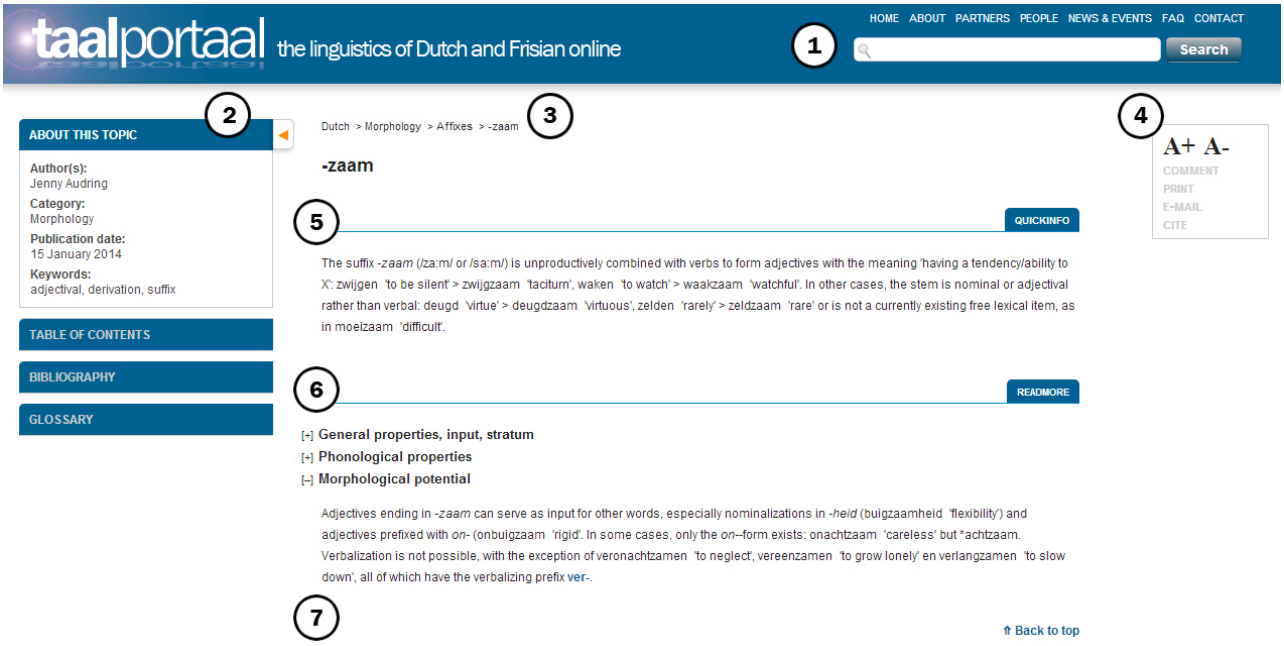

Figure 2: Screenshot of a typical Taalportaal page. The numbers indicate the following areas: (1) general search box; (2) left panel with metadata information, table of contents and links to the bibliography and glossary; (3) breadcrumbs showing the position of the current page in the table of contents hierarchy; (4) right panel with the options zoom, comment, print, email, cite (not available in the current beta version); (5) introduction of the main text; (6) main text; (7) reference list and links to related topics (not available in the current beta version).

Layer	**content**
Wordgroup	mooie films
Gloss	beautiful movies
Translation	beautiful movies
Phonetics	mo:jə fɪlms
Phonology	'moo.''ie 'films
Morphology	[mooi-e] [film-s]
Syntax	[(NP) [(ADJ) mooie] [(N) films]]

Table 1: The set of analysis layers for examples within *Taalportaal* topics

Authors are encouraged, but not obliged, to fill in most of the analysis layers per example, since the richer the source data, the more options this gives for future search queries. For example, an option is to enable pattern searches, so a user can search for all examples of the type 'ADJ films' or all example words with the stress pattern 'main stress-secondary stress'.

Another way in which we strive to maintain a high level of content homogeneity is by having the initial writing phase followed by an editorial phase in the fall of 2014, during which all topics will undergo thorough revision where needed. Special attention during this process will have to be given to the insertion of the *Syntax of Dutch* in the portal, since this material has a different background and structure than that of the other modules. This topic is discussed in the next section.

5. The *Syntax of Dutch*

The authors who work in Oxygen create their grammatical topics either from scratch or base them on existing resources. An exception in this is the module of Dutch syntax, for which all data is directly taken from the existing (and almost finished) *Syntax of Dutch*, an up-to-date and comprehensive syntactic description of Standard Dutch. Currently, three parts of this seminal work have been published (Broekhuis & Keizer, 2012; Broekhuis & Den Dikken, 2012; Broekhuis, 2013). Since its data is written in Word and covers thousands of pages, a necessary conversion to XML has to be performed before it can be added to the portal. This conversion is done using a Perl-script that has been especially created for this particular job. Since the converted material can still contain errors, all texts need to be manually checked before they can appear on the portal site. Currently, one of three parts of the *Syntax of Dutch* has been converted and published on the *Taalportaal* website.

6. The web application

The *Taalportaal* web application is a servlet that is written in Java and which runs in Apache Tomcat. It consists of a thin frontend and a fat backend. The frontend contains the website www.taalportaal.org and runs in the web browser of the user. It presents the user interface and allows the user to formulate and run queries as well as to navigate through the texts, an ontology of linguistic terms, a bibliography of relevant linguistic literature and links to

outside corpora. The backend runs on a dedicated server, assists the frontend by executing queries, retrieving materials, etc. and holds the databases.

As for the frontend, special attention is paid to the critical feature of how topics are cross-linked and can be retrieved by users. As for the backend, the portal will be constructed in such a way that after completion, its content can be updated, adapted and extended automatically.

6.1 Website and search options

The website is the actual 'portal' for users. At its core are the set of grammatical topics created by the authors, and the full *Syntax of Dutch*. Furthermore, the portal will contain an ontology of linguistic terms, a bibliography, links to online corpora (see section 6.3), and information about the project, contributors, frequently asked questions and relevant links, news and events. The bibliography page will be linked to the same database that the authors use. Users can search this database and navigate to topics with a pointer to a specific reference. We plan to extend the database with links to electronic versions of references (where available) in the future.

The topic page contains not only the topic, but metadata information about the topic as well (author, language, category, publication and revision dates) (see figure 2). Special care is given to the search and navigation options of the portal, which play a critical role in user satisfaction. We will offer a range of search options, so as to suit a diverse target audience. First, users can search the entire content of the portal with a search box offering simple search, wild card options and 'advanced search' function. Second, users can use the complete table of contents to navigate through the list of available topics. Third, users will also be able to access the register of linguistic terms, which links to the occurrence of a term in actual topics. As for navigation, each topic will contain a clickable breadcrumb trail, providing the possibility to jump to a parent topic. Furthermore, each topic contains an excerpt of the table of contents showing the topic's position in its parent topic, as well as a list of related topics.

6.2 Ontology of linguistic terms

An ontology of linguistic terms is currently being constructed. This ontology serves the functionality of the portal for three reasons. First, it will provide users with a list of terms and their definitions. This list will be available, and searchable, on a separate page of the portal. Also, as mentioned above, the definitions will be linked to occurrences of the terms in the actual topics. Thus, users will be able to get instant information on a linguistic term (by means of a popup, see figure 3), and can also be pointed to other relevant topics containing this term. Second, the list will serve as a register, providing links from each term to a series of topics. Third, the ontology will enable query expansion of user searches, thus enriching retrieval options. For example, a search for the term *article* can be extended to occurrences of the word *determiner* or even the actual Dutch determiners *de* and *het* (table 2).

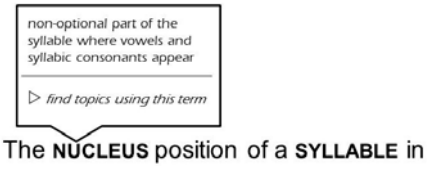

Figure 3: Example of a popup appearing when a user hovers over a linguistic term in a topic text.

query	**expanded query**
article	article, determiner, *de*, *het*, *een*, *it*
agreement	agreement, concord
lexical morpheme	lexical morpheme, free morpheme

Table 2: Examples of query expansion through the use of the ontology of linguistic terms

6.3 Interoperability

The critical feature in which the digital portal distinguishes itself from traditional paper resources is the way in which its content will be crosslinked. This is done in several ways in *Taalportaal*: by the authors themselves, by an editor, and automatically.

First, when writing, authors can link to other elements, topics or outside sources in their XML-editor. However, it is impossible for authors to have a profound knowledge of the content of all existing topics, let alone topics that have not been written yet. Therefore, we plan to add extra links in the editorial phase, which is due to start when most of the topics have been written.

Second, an automatically generated list of 'related topics' will be added to each topic. This list will be created on the basis of several aspects such as language (Dutch vs. Frisian), category (phonology, morphology, syntax), keywords, backlinks, examples and the occurrence of marked terms.

Third, topics can be automatically linked on the basis of shared terms, examples or references. Like the popup for linguistic terms (figure 3), we foresee popups when users hover over an example or reference, showing links to other relevant topics. For the linking of similar or related terms, the ontology of linguistic terms will be used. As for the linking of examples, apart from links to other relevant topics, we plan to improve interoperability by linking to outside sources such as the LASSY-corpus (a syntactically annotated corpus of 700 million words, Van Noord et al. 2013), and the *Algemeen Nederlands Woordenboek* (Schoonheim & Tempelaars, 2010, online available at http://anw.inl.nl).

7. Conclusion

In this paper, we presented the *Taalportaal* project, a project that will deliver an online, fully searchable grammar of Dutch and Frisian for the scientific

community. We discussed how content is created and how the workflow of the project is structured. At the time of writing (March 2014), a beta version of the portal is available at www.taalportaal.org. The final version of the portal will be available at the end of 2015.

There are several ways in which we expect *Taalportaal* to benefit the scientific community. First, *Taalportaal* will be the first comprehensive, online reference of Dutch and Frisian phonology, morphology and syntax. Second, the portal's language is English, which will enable global use of its data. Third, its online nature will give us the opportunity to provide rich crosslinking between topics, the two languages and linguistic domains, as well as linking between examples and outside sources such as corpora.

8. Acknowledgements

The *Taalportaal* project is funded by NWO, the Netherlands Organisation for Scientific Research, under grant 175.010.2009.003.

9. References

Broekhuis, H., Den Dikken, M. (2012). *Syntax of Dutch: Nouns and noun phrases – volume 2*. Amsterdam: Amsterdam University Press. *Comprehensive Grammar Resources*. Series editors: Henk van Riemsdijk & István Kenesei.

Broekhuis, H., Keizer, E. (2012). *Syntax of Dutch: Nouns and noun phrases – volume 1*. Amsterdam: Amsterdam University Press. *Comprehensive Grammar Resources*. Series editors: Henk van Riemsdijk & István Kenesei.

Broekhuis, H. (2013). *Syntax of Dutch: Adjectives and adjective phrases*. Amsterdam: Amsterdam University Press. *Comprehensive Grammar Resources*. Series editors: Henk van Riemsdijk & István Kenesei.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997). *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk. Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn.

Schoonheim, T., Tempelaars R. (2010). Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW). In: Anne Dykstra and Tanneke Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress. Ljouwert, Fryske Akademy/ Afûk*, p. 179.

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., Tjong Kim Sang, E., Vandeghinste, V. (2013). Large Scale Syntactic Annotation of Written Dutch: Lassy. In: P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*. Springer, pp. 147--164.