

University of Groningen

Enriching a Scientific Grammar with Links to Linguistic Resources: The Taalportaal

van der Wouden, Ton; Bouma, Gosse; van de Camp, Matje; van Koppen, Marjo;
Landsbergen, Frank ; Odijk, Jan

Published in:
CLARIN in the Low Countries

DOI:
[10.5334/bbi.24](https://doi.org/10.5334/bbi.24)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

van der Wouden, T., Bouma, G., van de Camp, M., van Koppen, M., Landsbergen, F., & Odijk, J. (2017). Enriching a Scientific Grammar with Links to Linguistic Resources: The Taalportaal. In J. Odijk, & A. van Hessen (Eds.), CLARIN in the Low Countries (pp. 299-310). London: UBIQUITY PRESS LTD. DOI: 10.5334/bbi.24

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Enriching a Scientific Grammar with Links to Linguistic Resources: The Taalportaal

Ton van der Wouden^{a,f}, Gosse Bouma^b, Matje van de Camp^c, Marjo van Koppen^d, Frank Landsbergen^e and Jan Odijk^d

^aMeertens Instituut Amsterdam, ^bGroningen University, ^cTaalmonsters, ^dUtrecht University, ^eInstitute for Dutch Lexicology INL.

^fCorresponding Author: ton.van.der.wouden@meertens.knaw.nl

ABSTRACT

Scientific research within the humanities is different from what it was a few decades ago. For instance, new sources of information, such as digital grammars, lexical databases and large corpora of real-language data offer new opportunities for linguistics. The Taalportaal grammatical database, with its links to other linguistic resources via the CLARIN infrastructure, is a prime example of a new type of tool for linguistic research.

24.1 Introduction

This chapter focuses on the ways that the digital Taalportaal grammar is enriched with links to language corpora and other digital linguistic resources. We first give an introduction to the goals and architecture of the Taalportaal, a new type of online scientific grammar that covers the syntax, the morphology, as well as the phonology, of Dutch and Frisian, the two official languages of the Netherlands. In the second part, we elaborate on why and how the Taalportaal's grammatical information is enriched with links to corpora and other linguistic resources.

24.2 The Taalportaal

Language is everywhere. The working linguist is confronted with linguistic data any moment they read a newspaper, talk to their neighbour, watch television, or switch on the computer. To overcome the volatility of many of these data, digitised corpora have been compiled for languages all around the globe since the 1960s. These days, there is therefore no lack of natural language resources, at

How to cite this book chapter:

van der Wouden, T, Bouma, G, van de Camp, M, van Koppen, M, Landsbergen, F and Odijk, J. 2017. Enriching a Scientific Grammar with Links to Linguistic Resources: The Taalportaal. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 299–310. London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbi.24>. License: CC-BY 4.0

least not for commonly studied languages like English and Dutch. Large corpora and databases of linguistic data are amply available, both in raw form and enriched with various types of annotation, and often free of charge or for a very modest fee.

There is no lack of linguistic descriptions either: linguistics is a very lively science area, producing tens of dissertations and thousands of scholarly articles in such a small country as the Netherlands alone. An enormous body of linguistic knowledge, however, is stored in paper form only: in grammars, dissertations and other publications, be they aimed at scholarly or lay audiences. The digitisation of linguistic knowledge is only beginning, and online grammatical knowledge is relatively scarce in comparison with all the treasures that are hidden in the bookshelves of libraries and studies.

Of course, there are notable exceptions. One such exception is the Taalportaal (Language Portal) project, an online portal containing a comprehensive and fully searchable digitised reference grammar, i.e. an electronic reference of Dutch and Frisian phonology, morphology and syntax. Information about the Afrikaans language is currently being added as well. With English as its meta-language, the Taalportaal aims at serving the international scientific community by organising, integrating and completing the grammatical knowledge of the Dutch and Frisian languages, as well as of Afrikaans.

The Taalportaal (www.taalportaal.org) is a collaboration project of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO). The project is aimed at the development of a comprehensive and authoritative scientific grammar for Dutch and Frisian in the form of a virtual language institute (cf. Landsbergen et al., 2014).

The Taalportaal is built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian. Its prime intended audience is the international scientific community, which is why English is chosen as the language used to describe the language facts. The Taalportaal aims to provide an exhaustive collection of the currently known data relevant for grammatical research, as well as an overview of the currently established insights about these data. This is an important step forward compared to presenting the same material in the traditional form of (paper) handbooks. For example, the three sub-disciplines of syntax, morphology and phonology are often traditionally studied in isolation, but, by presenting the results of these sub-disciplines on a single digital platform and internally linking these results, the Taalportaal contributes to the integration of the results reached within these disciplines.

This can be illustrated by means of a simple example concerning diminutive formation in Dutch. At first sight, this may look like a strictly morphological phenomenon, but upon closer inspection there are certainly also phonological and syntactic aspects to it. For example, the form of the diminutive morpheme depends on the phonological structure of the preceding noun: *hond-je* 'dog.dim', *kam-metje* 'comb.dim', *konin-kje* 'king.dim', etc. There is also a syntactic effect of diminutive formation in that it changes the gender of the input noun; diminutives are all neuter and thus select the definite singular article *het* 'the' (cf. *de hond* 'the dog' versus *het hondje* 'the dog.dim') and may also trigger different forms of agreement (cf. *een oude hond* 'an old dog' versus *een oud hondje* 'an old dog.dim'). Semantically, many morphological diminutives carry a (positive or negative) emotional load. Thus, the usage possibilities of *hondje* '(cute) doggy' are different from those of *kleine hond* 'small dog'. The Taalportaal makes visible these and less obvious cases of grammatical phenomena that are not restricted to one of the traditional sub-disciplines, to the benefit of each of the three disciplines and thus to the study of grammar in general.

The Netherlands are not the only country considering a linguistic knowledge base like the Taalportaal. Recently, South Africa has started building a virtual language institute called Viva (<http://viva-afrikaans.org/>) that aims at developing a digital infrastructure for the Afrikaans language. Among its goals are the study and description of Afrikaans, as well as the development

of tools and resources for written and spoken Afrikaans, including digital dictionaries and corpora; language advice is also supplied. The cornerstone of the VivA portal is a comprehensive grammar of Afrikaans, which is inspired by and based on the Taalportaal architecture, and is currently being added to the Taalportaal infrastructure.

As of January 2016, the first release of the Taalportaal is online. Figure 24.1 below shows an instance of the portal's opening screen.

Technically, the Taalportaal is built as a number of XML files, organised as DITA-topics.¹ It is freely accessible via the Internet via any standard internet browser. The organisation and structure of much of the linguistic information is reminiscent of, and to a certain extent inspired by, Wikipedia and comparable online information sources. An important difference, however, is that Wikipedia's democratic (anarchistic) model is avoided by restricting the right to edit the Taalportaal information to authorised experts.

Figure 24.2 shows a small, introductory fragment of the portal concerning Dutch phonology:²

Figure 24.1: the opening page of the taalportaal site.

¹ DITA, the Darwin Information Typing Architecture, is an XML data model for authoring and publishing. According to https://en.wikipedia.org/wiki/Darwin_Information_Typing_Architecture (as of 17 June 2016), 'the name derives from the following components:

Darwin: it uses the principles of specialization and inheritance, which is in some ways analogous to the naturalist Charles Darwin's concept of evolutionary adaptation,

Information typing, which means each topic has a defined primary objective (procedure, glossary entry, troubleshooting information) and structure,

Architecture: DITA is an extensible set of structures'.

² A-class vowels are known as 'long vowels' or 'tense vowels' in other frameworks; cf. http://www.taalportaal.org/taalportaal/topic/pid/topic-13998813314542255#a_vowel.

The rounded high front-central vowel /y/

The rounded, high, front-central A-CLASS VOWEL/y/ is found in words such as:

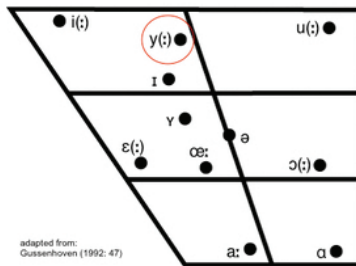
Example 1

- a. nu /ny/ 'now'
- b. humor /hy.mɔr/ ['hymɔr] 'comedy, humour'
- c. bruto /bry.to/ ['bryto] 'gross amount, bruto'
- d. puur /pyr/ 'pure'
- e. kostuum /kos.tym/ [kos'tym] 'suit, dress'

It is spelled with a single letter <u> in open syllables (see (1a)-(1c)); this letter is doubled (<uu>) in closed syllables (see (1d)-(1e)).

Figure 1 (cf. Gussenhoven 1992) depicts the (Dutch) vowel's position within the vowel chart.

Figure 1



Articulation

/y/ is a rounded, high, front-central, A-class vowel. The tongue body is fronted, the tongue tip is down. Articulation is like that of /i/ except with rounded lips: the front cavity is enlarged because of pursing of the lips (Collins and Mees 2003; Eijkman 1937).

Figure 24.2: Dutch phonology example.

Among other things, the information about the vowels contains data about their distribution, phonetic details, and links to sound files exemplifying the realisation of the sound in several positions within the word. This is illustrated in figure 24.3:

The Taalportaal grammars were not built from scratch. One of their main components is an online version of the *Syntax of Dutch* (SoD; Broekhuis et al., 2012–2016), a descriptive grammar that goes well beyond the level of detail provided by other sources, including reference grammars. Although the SoD grammar is descriptive in nature, the emphasis in the selection and presentation of the phenomena discussed is clearly guided by discussions in the (generative) theoretical literature (Broekhuis, 2013; Hoeksema, 2013; Bouma et al., 2015); by implication, the same holds for the Taalportaal's treatment of Dutch syntax. For Dutch morphology, the Taalportaal has been built, among many other sources, upon the first volume of Haeseryn et al. (1997), as well as on morphological handbooks such as those of De Haas and Trommelen (1993), Booij (2002) and Smessaert (2013). The parts on Dutch phonology are indebted to Booij (1995) and Kooij and van Oostendorp (2003). For Frisian there was no lack of studies that could be profited from either, for instance Visser (1997), Hoekstra (1998), Tiersma (1999), Popkema (2006), and De Haan et al. (2010).

Table 3: Soundfiles, waveforms and spectrograms of the above sound files, with indications of the relevant acoustic parameters of Northern Standard Dutch /y/


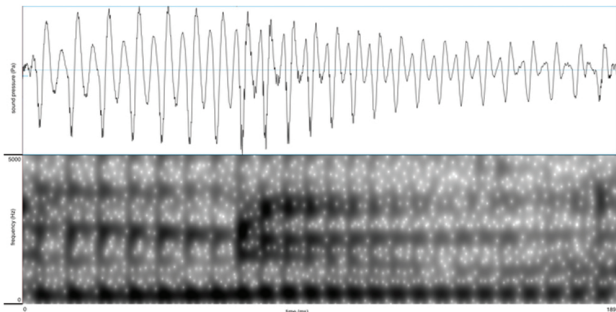

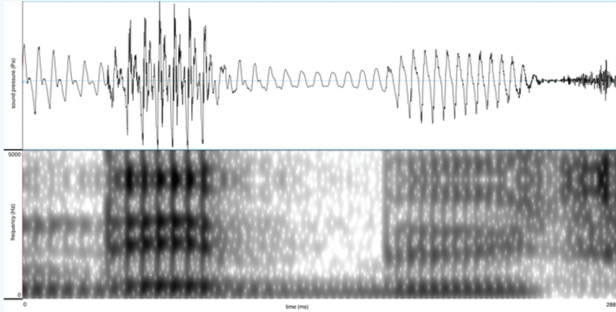
wordgroup	phonological context	soundfile	waveform/spectrogram
(...) <i>en nu gauw naar binnen</i> '(...) and now soon inside'	word-final		 [click image to enlarge]
<i>Van Hooijdonk heeft zijn debuut gemaakt voor Feyenoord</i> 'Van Hooijdonk made his debut for Feyenoord'	pre-obstruent		 [click image to enlarge]

Figure 24.3: Dutch phonetics example.

As an internet grammar portal, the Taalportaal is somewhat comparable to the grammis portal of the German Language Institute (IDS; <http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht>). grammis, however, covers only one language (German) and is not aimed at the international scientific community but primarily at a German audience. These differences explain both the choice of the metalanguage (English for Taalportaal, German for grammis) and the differences in depth of analysis. grammis moreover is far less connected to other data sources than Taalportaal.

Besides the grammar modules, the Taalportaal contains an extensive ontology of linguistic terms (recently recast in the CLARIN Concept Registry; cf. Schuurman, 2015) and a large bibliography. Many of the words and phrases in the texts are marked: they can be clicked on, which results in sounds being played, definitions popping up and/or related topics being opened, a feature that will be elaborated upon in the following sections.

24.3 Enriching the Taalportaal

It is becoming more and more common for 21st-century linguists to want to check whether and to what extent the linguistic facts as they are presented in the linguistics literature correspond to the linguistic reality. In this context, '[c]reating a link between a descriptive grammar and a

syntactically annotated corpus can be valuable for various reasons. Illustrating a given construction with corpus examples may help to get a better understanding of the variation of the construction and the frequency of these variants. Corpus data may also convince a reader that a given variant actually occurs in (well-formed) text, or in some cases may illustrate that examples judged ungrammatical by the authors of the descriptive grammar do occur with some frequency in actual text' (Bouma et al., 2015).

Searching for realistic language data becomes easier by the day, thanks to joint efforts such as CLARIN (www.clarin.eu) that seek to enhance the scientific research infrastructure by, among many other things, linking and making available large existing corpora and other linguistic resources under a single user licence.

The (syntactically annotated part of the) Spoken Dutch Corpus (manually verified, speech from various situations, 1M words; Oostdijk, 2000; van der Wouden et al., 2003), the Lassy Small treebank (manually verified, written material from various genres, 1M words, 65,200 sentences) and the Lassy Large treebank (automatically created, written material from various genres, 700M words, 8.6M sentences; van Noord et al., 2013) are all suitable corpora for this kind of application. The first two resources provide high-quality data for a limited amount of text, while the last resource provides wide-coverage, but noisy, data. All treebanks follow (with minor modifications) the same annotation standard (Van Eynde, 2003 for lemmatisation and POS tagging; Schuurman et al., 2003 for syntax), which has become a *de facto* standard for Dutch corpus annotation, allowing for the re-use of the queries on these new data.

Taalportaal has been enriched with a range of queries that search for relevant constructions in these corpora. Queries are linked to:

- Linguistic examples;
- Linguistic terms; or
- Names or descriptions of constructions.

The queries are embedded in the Taalportaal texts as standard hyperlinks. Clicking these links brings the user to a corpus query interface where the specified query is executed – or, if it can be foreseen that the execution of a query takes a lot of time, the link may also connect to an internet page containing the stored result of the query.

Syntactic annotations are a complex type of data, usually formally encoded in accordance with a well-defined schema in XML at present. Sometimes these syntactically annotated corpora come with a search interface, but to search these complex data efficiently and optimally, one needs a command of an XML search language such as XPath.³ Many researchers in linguistics lack these skills. Although the basics of the XPath language are not difficult, interesting queries often become very complex. Moreover, one has to know every particular detail of the encoding of constructions in a particular treebank.

24.3.1 Automatic Links

Many of the links have been generated automatically: all examples in the Taalportaal can be clicked on, which will open a 'pop-up' window like the one in figure 24.4:

By clicking the links, the example sentence *Jan is niet boos (over die opmerking)* can be searched in a number of resources, as is illustrated in the screen dump above: in this case the choices are Google, DBNL, GrE TEL, CHN, OpenSoNar, and TaalPortaal. Suppose we choose the third option, the GrE TEL web application (<http://portal.clarin.nl/node/1967>; cf. Augustinus et al., 2013 and chapter 22 in this volume); we can then search for linguistic structures in the most user-friendly

³ Cf. <https://en.wikipedia.org/wiki/XPath>.

Example 2

a. Jan is niet boos (over die opmerking).
Jan is not angry about that remark

b. Jan is niet tevreden (over zijn beloning).

JAN IS NIET BOOS (OVER DIE OPMERKING). [EDIT]

Find this example:

- with [Google](#)
- in the [DBNL](#) (linguistic literature subpart / the entire site)
- with [GrE TEL](#) in LASSY Small, the Corpus Gesproken Nederlands (CGN) or SoNar
- in the [Corpus Hedendaags Nederlands \(CHN\)](#) (CLARIN ACCESS REQUIRED)
- in the SoNar Corpus with [Open SoNar](#)
- in [TaalPortaal](#)

a. Jan is (er) boos (over) dat Peter niet gekomen is.
Jan is there angry about that Peter not come is

Figure 24.4: Taalportaal pop-up example.

Step 2: Input Parse

The structure of the **tagged** ^[7] and **parsed** ^[7] sentence: *Jan is niet boos (over die opmerking)*

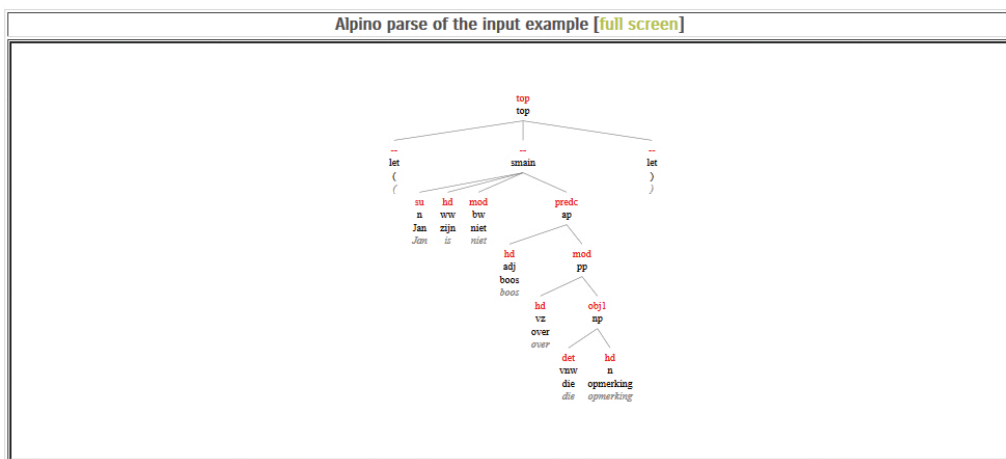


Figure 24.5: Taalportaal syntactic analysis example.

way – that is, without having to learn a corpus query language – in a number of large annotated corpora of Dutch (cf. Augustinus et al., 2012). The sentence is parsed using the Alpino parser (cf. van Noord, 2006). The resulting parse is shown in figure 24.5.

The query can be edited via a menu, for example by replacing specific lexical items by syntactic categories, as illustrated in figure 24.6:

If the user has made their choices, the structure can be searched for. Part of the result is given in figure 24.7:

The example sentences show copula sentences with an adjective that has a prepositional complement: *Peking is niet tevreden met zijn groeiende economische macht* ‘Beijing is not satisfied with its growing economic power’ and *Rotterdam is ook bezig met zo’n plan* ‘Rotterdam is busy with such a plan as well’.

Step 3: Select relevant parts

Indicate the relevant^[?] parts of the sentence, i.e. the parts you are interested in. [\[view input parse\]](#)

sentence	Jan	is	niet	boos	(over	die	opmerking)
word	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
lemma	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
word class	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
optional in search	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

OPTIONS

- Respect word order
- Ignore properties of the dominating node ^[?]

GUIDELINES

- **word:** The exact word form. This is a case sensitive feature.
- **lemma:** Word form that generalizes over inflected forms. For example: *zin* is the lemma of *zin*, *zinnen*, and *zinnetje*; *gaan* is the lemma of *ga*, *gaat*, *gaan*, *ging*, *gingen*, and *gegaan*. Lemma is case insensitive (except for proper names).
- **word class:** Short Dutch part-of-speech tag. The different tags are: *n* (noun), *vw* (verb), *adj* (adjective), *1id* (article), *vnw* (pronoun), *vg* (conjunction), *bw* (adverb), *tw* (numeral), *vz* (preposition), *tsw* (interjection), *spec* (special token), and *let* (punctuation).
- **optional in search:** The word will be ignored in the search instruction. It may be included in the results, but it is not necessary.

Figure 24.6: GrE TEL input example.

dpc-ind-001645-nl-sen.p.35.s.1	Peking is niet tevreden met zijn groeiende economische macht , maar wil ook politieke en diplomatieke invloed verwerven in Azië .
WS-U-E-A-0000000216.p.11.s.10	Rotterdam is ook bezig met zo'n plan .

Figure 24.7: GrE TEL output example.

2.1. Prepositional complements

The examples in (2) show that complements of adjectives are normally PPs, which are often optional.

Figure 24.8: Taalportaal PP example.

24.3.2 Manually Prepared Queries

Whereas it is relatively easy to automatically translate example words or sentences into corpus queries, this usually does not hold for grammatical descriptions meant for human readers. Still, these readers might be interested to check the grammarian’s claims in corpus data.

CLARIN-NL made it possible to enrich Taalportaal fragments, most of them dealing with Dutch syntax, with more sophisticated queries in annotated corpora (cf. Bouma et al., 2015).

The translation of a linguistic example, a linguistic term, or a name or description of a construction is not a task that is easy or that even has deterministic results that could be implemented in an algorithm (cf. Bouma et al., 2015). Therefore, the queries were formulated by experts, who got selections of the Taalportaal texts to read, interpret, and enrich with queries where appropriate. The queries were amply annotated with explanations concerning the choices made in translating the grammatical term or description or linguistic example into the corpus query. When necessary, warnings about possible false hits, etc., were added as well. The results were checked by senior linguists. Consider the small section from Dutch syntax depicted in figure 24.8:

1. Met cd-rom voor Windows en Mac , 15.000 Nederlandse trefwoorden **afkomstig** uit andere talen , 684 blz. , EUR 55 ; ISBN 90-6648-0270 . ✚
2. Zo is het woord japon bijvoorbeeld **afkomstig** uit Japan , werd het woord sauna geleend van de Finnen en is bazaar een woord dat uit het Perzisch komt . ✚
3. De cd-rom maakt gebruik van standaardbrowsertechniek en is daardoor **geschikt** voor zowel Windows als Mac . ✚

Figure 24.9: first Lassy output example.

1. Maar hoe dat precies komt is **niet zo duidelijk** . ✚
2. In het Frans zijn buitenlandse woorden officieel **niet welkom** en op IJsland wordt voor elk nieuw begrip een IJslands woord bedacht . ✚
3. Hij studeerde medicijnen en filosofie aan de Universiteit van Leiden (1685 - 1690) en emigreerde vervolgens naar Engeland , waar hij **als arts werkzaam** was en een aantal politieke en geschriften in het Engels vervaardigde . ✚

Figure 24.10: second Lassy output example.

The result of the annotation process is as follows:

QUERY/QUERIES: ? ✕

DESCRIPTION:
Adjectives with a prepositional complement (**@rel="pc***) as sister.

XPATH:
//node[@rel="hd" and @pt="adj" and ../node[@rel="pc" and @cat="pp"]]

[Show results of this query in lassysmall with PaQu](#)

DESCRIPTION:
Modified adjectives without a PP-complement.

XPATH:
//node[@rel="predc" and @cat="ap" and node[@rel="mod"] and node[@rel="hd" and @pt="adj"] and not(node[@cat="pp"])]

[Show results of this query in lassysmall with PaQu](#)

Figure 24.11: result of the annotation process.

The sentence about adjectives has been translated into two, radically different queries: the first one searches for adjectives with a prepositional complement as sister, the second one for predicatively used modified adjectives without a PP-complement or any kind of modifier. Clicking the (blue) link ‘Show results of this query in lassysmall with PaQu’ will open a new browser window to the PaQu interface (<http://portal.clarin.nl/node/4182>; cf. Odijk, 2015). The first query results in a number of hits from the Lassy Small corpus; the first ones are given in figure 24.9:

We twice see the adjective *afkomstig* ‘originating’ followed by a prepositional phrase headed by *uit* ‘from’, and once the adjective *geschikt* suitable with a prepositional phrase with *voor* ‘for’.

The second query results in the sentences, among many others, given in figure 24.10:

Here we see the three adjectives *duidelijk* ‘clear’, *welkom* ‘welcome’, and *werkzaam* ‘active’, modified with *niet zo* ‘not so’, *niet* ‘not’, and *als arts* ‘as a doctor’, respectively.

If the user clicks the small plus sign following the result sentence, a parse tree is shown. (PaQu offers corpus statistics as well, but that is beyond the scope of this chapter.)

As the corpora dealt with so far offer little or no morphological or phonological annotation, they cannot be used for the formulation of queries to accompany the Taalportaal on morphology and phonology. There is, however, a linguistic resource that is in principle extremely

useful for precisely these types of queries, namely the CELEX lexical database (cf. Baayen et al., 1995), which offers morphological and phonological analyses for more than 100,000 Dutch lexical items. This database is currently being transferred from the Nijmegen Max Planck Institute for Psycholinguistics (MPI) to the Leiden Institute for Dutch Lexicology (INL). It has its own query language, which implies that Taalportaal queries that address CELEX have to have another format – but again, the Taalportaal user will not be bothered with the small details.

As was mentioned above, the Frisian language – the other official language of the Netherlands, with Dutch – is described in the Taalportaal as well, in parallel to Dutch. Although there is no lack of digital linguistic resources for Frisian, internet accessibility of these resources is lagging behind. This makes it difficult at this point to enrich the Frisian parts of the Taalportaal with queries. It is to be hoped that this CLARIN project will stimulate further efforts to integrate Frisian language data in the research infrastructure.

24.4 Concluding Remarks

In the first part of this chapter, we have introduced the Taalportaal grammar portal, a digital scientific grammar of Frisian and Dutch, covering the syntax, morphology and phonology of the two official languages of the Netherlands. In the second part of the chapter, we focused on the dynamic links from the grammatical descriptions to other linguistic resources of various sorts – something that is of course impossible in traditional paper grammars. By this extension, the Taalportaal functions as a hub within the scientific infrastructure supplied by CLARIN. This is relevant for the Taalportaal users in at least two ways:

- it increases the value of the Taalportaal as a research tool; and
- it lowers the threshold to use the linguistic resources involved.

The Taalportaal's open architecture allows for extension with new languages (Afrikaans is well under way), but also with new language varieties (dialectal data, historical data, etc.). Moreover, the CLARIN network allows for extension with links to new and so far largely unexplored linguistic resources, such as the huge digital dictionaries of the INL and semantically organised lexical databases such as Open Dutch WordNet (<http://wordpress.let.vuwr.nl/odwn/>; cf. Postma et al., 2016), which may make linguists' practical work even easier and, at the same time, even more exciting.

It is to be foreseen that future corpora of Dutch (and hopefully of Frisian as well) will be embedded in the very same CLARIN infrastructure, using the same architecture, the same type of interface, and the same kind of linguistic annotation.

Acknowledgements

The Taalportaal project was a joint effort of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology, and Leiden University. It was made possible by a grant from the Netherlands Organisation for Scientific Research (NWO Grant 175.010.2009.003).

Parts of the Taalportaal were enriched with queries to corpora in a separate project, CLARIN TPC, which was a collaboration of the Meertens Institute, the Institute of Dutch Lexicology, the Universities of Groningen and Utrecht, and De Taalmonsters. CLARIN TPC was made possible by a grant from CLARIN-NL (CLARIN-NL-15-001).

Previous Publications

Earlier publications that cover parts of the Taalportaal project or linguistic resources mentioned in this chapter are the following: Landsbergen et al. (2014), Bouma et al. (2015), and Van der Wouden et al. (2015, 2016).

References

- Augustinus, Liesbeth, Vincent Vandeghinste, & Frank Van Eynde (2012). Example-Based Treebank Querying. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 3161–3167.
- Augustinus, Liesbeth, Vincent Vandeghinste, Ineke Schuurman, & Frank Van Eynde (2013). Example-Based Treebank Querying with GrETEL – now also for Spoken Dutch. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. NEALT Proceedings Series 16. Oslo, Norway. 423–428.
- Baayen, R. Harald, Richard Piepenbrock, & L. Gulikers (1995). *The CELEX Lexical Database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Booij, Geert (1995). *The Phonology of Dutch*. Oxford: Oxford University Press.
- Booij, Geert (2002). *The Morphology of Dutch*. Oxford: Oxford University Press.
- Bouma, Gosse, Marjo van Koppen, Frank Landsbergen, Jan Odijk, Ton van der Wouden, & Matje van de Camp (2015). Enriching a Descriptive Grammar with Treebank Queries. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk and Adam Przepiórkowski (eds.), *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*. Warsaw: Institute of Computer Science, Polish Academy of Sciences, 13–25.
- Broekhuis, Hans (2013). De Syntax of Dutch: nieuw gereedschap voor de internationale neerlandistiek. *Internationale Neerlandistiek*, 51, 3, 243–260.
- Broekhuis, Hans, Norbert Corver, Marcel den Dikken, Evelien Keizer, & Riet Vos (2012–16). *Syntax of Dutch*. Amsterdam: Amsterdam University Press (7 volumes).
- Van Eynde, Frank (2003). Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands. Centrum voor Computerlinguïstiek K.U.Leuven.
- de Haan, Germen, Jarich Hoekstra, Willem Visser, & Goffe Jensma (red.) (2010). *Studies in West Frisian Grammar: Selected Papers by Germen J. de Haan*. Amsterdam: John Benjamins Publishing Company.
- de Haas, Wim, & Mieke Trommelen (1993). *Morfologisch Handboek van het Nederlands*. Den Haag: SDU uitgeverij.
- Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij, & Maarten C. van den Toorn (1997). *Algemene Nederlandse Spraakkunst*, 2e, geheel herz. dr. Groningen en Deurne: Martinus Nijhoff and Wolters Plantijn.
- Hoeksema, Jack (2013). Review of: Syntax of Dutch. Noun and Noun Phrases vols. 1 and 2. *Lingua*, 133, 385–390.
- Hoekstra, Jarich (1998). *Fryske Wurdfoarming*. Ljouwert: Fryske Akademy.
- Kooij, Jan, & Marc van Oostendorp (2003). *Fonologie, uitnodiging tot de klankleer van het Nederlands*. Amsterdam: University Press.
- Landsbergen, Frank, Carole Tiberius, & Roderik Dernison (2014). Taalportaal: an online grammar of Dutch and Frisian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2206–2210.
- van Noord, Gertjan (2006). At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors: *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42.
- van Noord, Gertjan, Gosse Bouma, Frank van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, & Vincent Vandeghinste (2013). Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk (red.), *Essential Speech and Language Technology for Dutch: the STEVIN Programme*, Springer, 147–164.
- Odijk, Jan (2015). Linguistic Research with PaQU. *Computational Linguistics in The Netherlands journal* 5, 3–14.

- Oostdijk, Nelleke (2000). The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of LREC 2000*, 887–894.
- Popkema, Jan (2006). *Grammatica Fries: de regels van het Fries*. Utrecht: Prisma.
- Postma, Marten, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, & Piek Vossen (2016): Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.
- Schuurman, Ineke, Machteld Schouppe, Heleen Hoekstra, & Ton van der Wouden (2003). CGN, an annotated corpus of spoken Dutch. In Anne Abeillé, Silvia Hansen-Schirra, and Hans Uszkoreit (eds.): *Proceedings of 4th International Workshop on Language Resources and Evaluation*, Budapest: European Language Resources Association (ELRA), 340–347.
- Schuurman, Ineke (2015). Concept revival: from ISocat to CLARIN Concept Registry. *CLARIN News* 7 January 2015 <https://www.clarin.eu/news/concept-revival-isocat-clarin-concept-registry>.
- Smessaert, Hans (2013). *Basisbegrippen morfologie*. Leuven/Den Haag: ACCO.
- Tiersma, Piter Meijes (1999). *Frisian Reference Grammar*, 2e ed. Ljouwert: Fryske Akademy.
- Visser, Willem (1997). *The Syllable in Frisian*. Dissertation Vrije Universiteit Amsterdam (Holland Academic Graphics).
- van der Wouden, Ton, Ineke Schuurman, Machteld Schouppe, & Heleen Hoekstra (2003). Harvesting Dutch trees: Syntactic properties of spoken Dutch. In Tanja Gaustad (ed.): *Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting*. Amsterdam/New York: Rodopi, 129–141.
- van der Wouden, Ton, Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, & Jan Odijk (2015). Enriching a grammatical database with intelligent links to linguistic resources. In Koenraad De Smedt (ed.): *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*. Linköping University Electronic Press, Linköpings Universitet.
- van der Wouden, Ton, Jenny Audring, Hans Bennis, Frits Beukema, Geert Booij, Hans Broekhuis, Norbert Corver, Crit Cremers, Roderik Dernison, Marcel den Dikken, Siebren Dyk, Carlos Gussenhoven, Ger de Haan, Vincent van Heuven, Eric Hoekstra, Jarich Hoekstra, Bart Hoogeveen, Gerbrich de Jong, Evelien Keizer, Anna Kirstein, Björn Köhnlein, Frank Landsbergen, Kathrin Linke, Marc van Oostendorp, Nina Ouddeken, Koen Sebregts, Carole Tiberius, Arjen Versloot, Willem Visser, Riet Vos, Truus de Vries, & Joke Weening (2016). Het Taalportaal: Een nieuwe wetenschappelijke grammatica voor het Nederlands en het Fries (en het Afrikaans). *Nederlandse Taalkunde* 21, 1 2016, 157–168.