Enriching a Grammatical Database with Intelligent Links to Linguistic Resources

Ton van der Wouden
Meertens Institute
Amsterdam, The Netherlands
Ton.van.der.wouden@meertens.knaw.nl

Gosse Bouma
Groningen University
The Netherlands
g.bouma@rug.nl

Matje van de Camp
De Taalmonsters
Tilburg, The Netherlands
matje@taalmonsters.nl

Marjo van Koppen
Utrecht University
The Netherlands
J.M.vanKoppen@uu.nl

Frank Landsbergen
Institute for Dutch Lexicography
Leiden, The Netherlands
Frank.Landsbergen@inl.nl

Jan Odijk
Utrecht University
The Netherlands
j.odijk@uu.nl

Abstract

We describe goals and methods of CLARIN-TPC, a project to enrich the on-line Taalportaal (Language Portal) grammatical database with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database.

1 Introduction

We describe the on-line grammatical database Taalportaal (Language Portal) an particularly how it is being enriched with intelligent links in the form of annotated queries to a variety of interfaces to on-line corpora and an on-line linguistic morphophonological database. This database contributes to the use of the CLARIN research infrastructure in the following ways:

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal.
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists.
- By redirecting the user to the front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

2 Background

Linguistic data is everywhere. The working linguist is confronted with data any moment he/she reads a newspaper, talks to their neighbour, watches television, switches on the computer. To overcome the volatility of many of these data, digitized corpora have been compiled for languages all around the globe since the nineteen sixties. These days, there is therefore no lack of natural language resources. 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 a small country as the Netherlands only. An enormous amount of this linguistic knowledge, however, is stored in paper form: in grammars, dissertations and other publications, both aimed at scholarly and lay audiences. The digitization of linguistic knowledge is only beginning, online grammatical knowledge is relatively scarce in comparison with what is hidden in the bookshelves of libraries and studies.

Of course, there are notable exceptions. One such exception is the Taalportaal (Language Portal) project, that is currently developing an online portal, containing a comprehensive and fully searchable digitized reference grammar, i.e. an electronic reference of Dutch and Frisian phonology, morphology and syntax. With English as its meta-language, the Taalportaal aims at serving the international scientific community by organizing, integrating and completing the grammatical knowledge of both languages. In contrast, the standard reference grammar for Dutch, the *Algemene Nederlandse Spraakkunst* (Haeseryn et al. 1997), is aimed at a broader (and other) audience than the international scientific community only, and is written in Dutch. The digital version[1] is essentially an XML-version of the paper edition.

To enhance the Taalportaal's value, the CLARIN project described here (NL-15-001: TPC) sought to enrich the grammatical information within the Taalportaal with links to linguistic resources. The idea was that the user, while reading a grammatical description or studying a linguistic example, was to be offered the possibility to find both potential examples and counterexamples of the pertinent constructions in a range of annotated corpora, as well as in a lexical database containing a wealth of morphophonological data on Dutch. Although links to raw text (including internet search) are offered as well, we here focus on resources with rich linguistic annotations explicitly, since we want to do more than just string searches: searching for construction types and linguistic annotations themselves is one way to reduce the problem of the massive ambiguity of natural language words and sentences.

In light of the restricted resources in terms both of time and money, this CLARIN project was not aiming at exhaustivity, that is, not all grammatical descriptions and not all examples are adorned with query links. TPC is explicitly to be seen as a pilot project, aiming for a proof of concept by showing the feasibility of efficient coupling of grammatical information with queries in a number of linguistic resources.

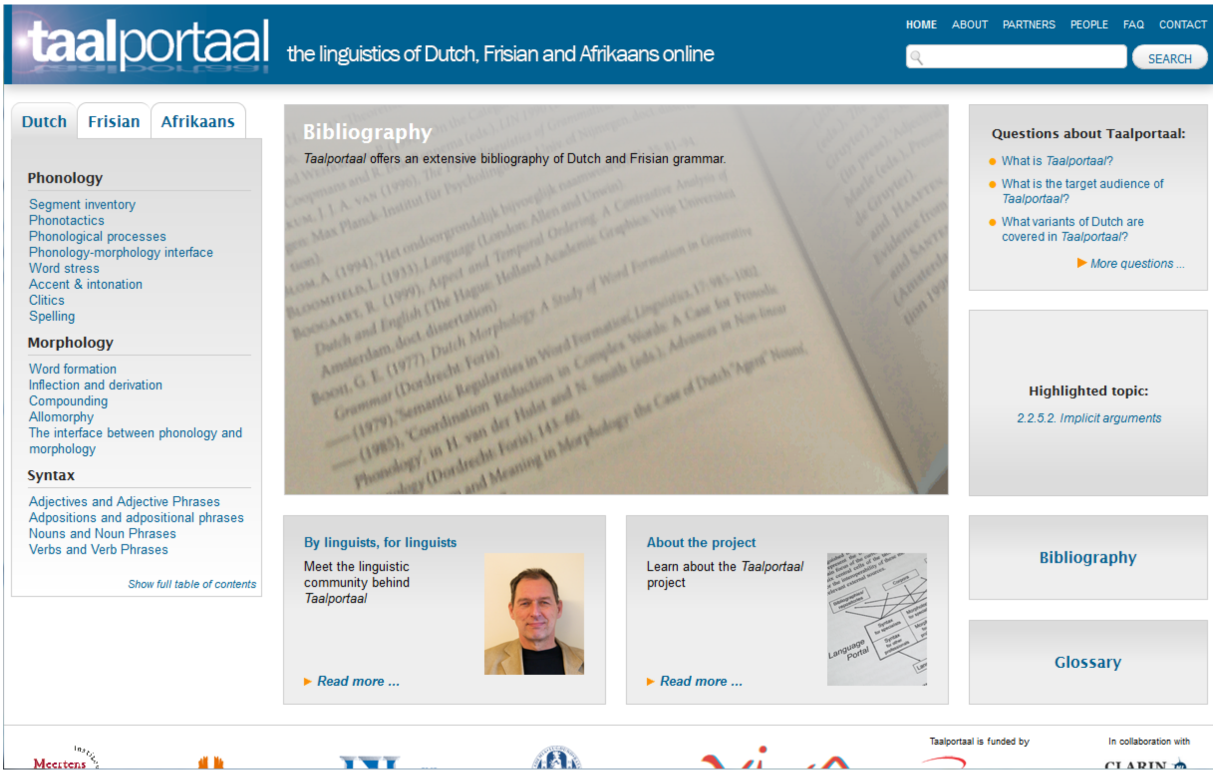
3 The Taalportaal

The Taalportaal project is a collaboration 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 provides 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 printed handbooks. For example, the three sub-disciplines syntax, morphology and phonology are often 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.

As of January 2016, the first release of the Taalportaal is online[2]. Figure 1 shows the portal's opening screen.

[1] http://ans.ruhosting.nl/e-ans/index.html.

[2] http://www.taalportaal.org.



taalportaal the linguistics of Dutch, Frisian and Afrikaans online

HOME ABOUT PARTNERS PEOPLE FAQ CONTACT

SEARCH

Dutch Frisian Afrikaans

Phonology

- Segment inventory
- Phonotactics
- Phonological processes
- Phonology-morphology interface
- Word stress
- Accent & intonation
- Clitics
- Spelling

Morphology

- Word formation
- Inflection and derivation
- Compounding
- Allomorphy
- The interface between phonology and morphology

Syntax

- Adjectives and Adjective Phrases
- Adpositions and adpositional phrases
- Nouns and Noun Phrases
- Verbs and Verb Phrases

Show full table of contents

Bibliography

Taalportaal offers an extensive bibliography of Dutch and Frisian grammar.

Questions about Taalportaal:

- What is *Taalportaal*?
- What is the target audience of *Taalportaal*?
- What variants of Dutch are covered in *Taalportaal*?

More questions ...

Highlighted topic:

2.2.5.2. Implicit arguments

By linguists, for linguists

Meet the linguistic community behind *Taalportaal*

Read more ...

About the project

Learn about the *Taalportaal* project

Read more ...

Bibliography

Glossary

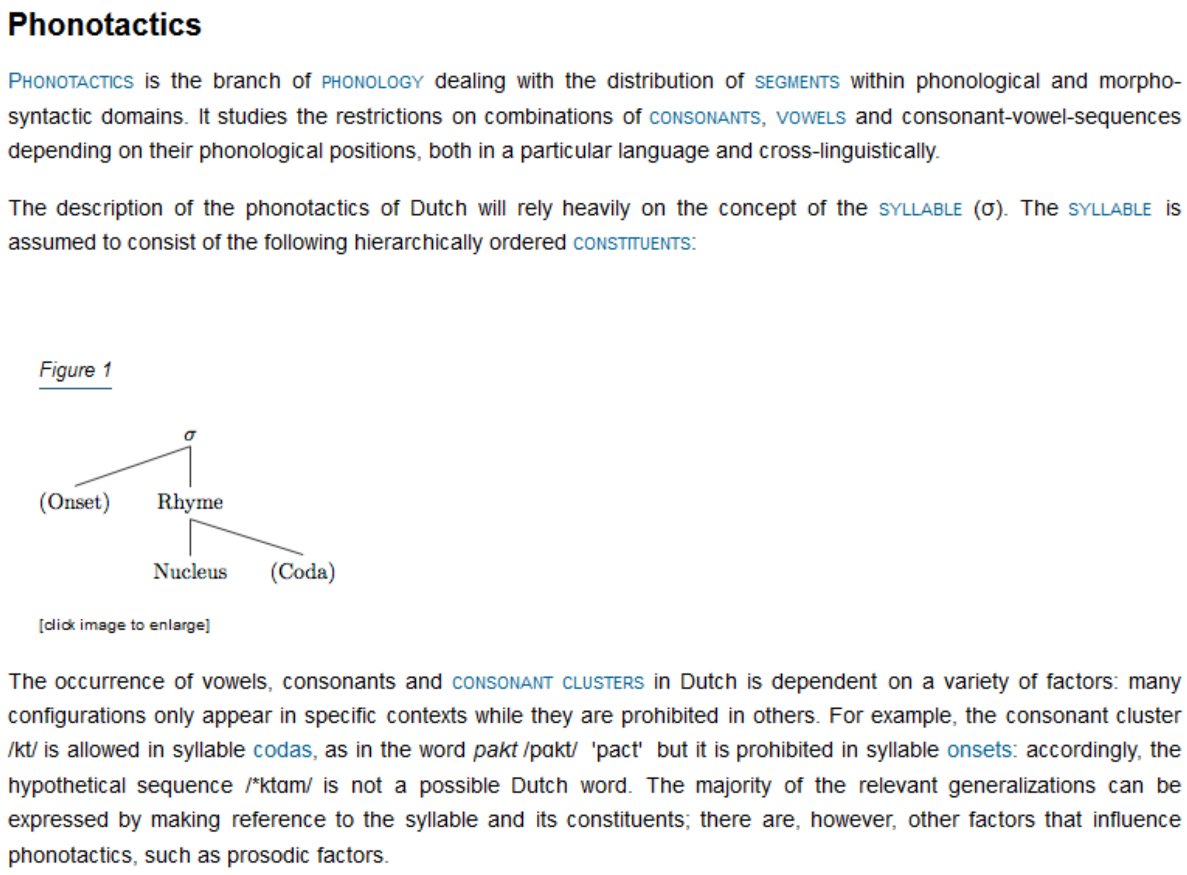
Taalportaal is funded by

In collaboration with

Figure 1: Opening screen of the Taalportaal.

Technically, the Taalportaal is built as an XML-database, organized as DITA-topics.[3] The data is freely accessible via the Internet using any standard internet browser. Organization and structure of much of the linguistic information is reminiscent of, and is to a certain extend 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 authorized experts. Figure 2 shows a small, introductory fragment concerning Dutch phonology.

[3] https://en.wikipedia.org/wiki/Darwin_Information_Typing_Architecture.



Phonotactics

PHONOTACTICS is the branch of PHONOLOGY dealing with the distribution of SEGMENTS within phonological and morpho-syntactic domains. It studies the restrictions on combinations of CONSONANTS, VOWELS and consonant-vowel-sequences depending on their phonological positions, both in a particular language and cross-linguistically.

The description of the phonotactics of Dutch will rely heavily on the concept of the SYLLABLE (σ). The SYLLABLE is assumed to consist of the following hierarchically ordered CONSTITUENTS:

Figure 1

[click image to enlarge]

The occurrence of vowels, consonants and CONSONANT CLUSTERS in Dutch is dependent on a variety of factors: many configurations only appear in specific contexts while they are prohibited in others. For example, the consonant cluster /kt/ is allowed in syllable codas, as in the word *pakt* /pɑkt/ 'pact' but it is prohibited in syllable onsets: accordingly, the hypothetical sequence /*ktɑm/ is not a possible Dutch word. The majority of the relevant generalizations can be expressed by making reference to the syllable and its constituents; there are, however, other factors that influence phonotactics, such as prosodic factors.

Figure 2: Introductory fragment on Dutch phonology.

The Netherlands are not the only country thinking of a virtual language institute like the Taalportaal. Recently, South Africa has started building a virtual language institute called **Viva!**[4] that aims at developing a digital infrastructure for Afrikaans. Among its goals are study and description of the Afrikaans language, and development of comprehensive tools and resources for written and spoken Afrikaans, including digital dictionaries and corpora; language advice is also supplied. Part of the Viva! portal is a comprehensive grammar of Afrikaans, which is based on the Taalportaal architecture, and will be part of the Taalportaal infrastructure.

Besides the grammar modules, the Taalportaal contains an ontology of linguistic terms (recently recast in the CLARIN Concept Registry, cf. Schuurman 2015) and an extensive bibliography. Note that the text in the pictures is full of words and phrases that are marked (bluish): these can be clicked, which results in definitions popping up and/or related topics being opened. Moreover, the texts are often amply illustrated, not only in the familiar ways with linguistic examples and tree-like drawings as in the fragment, but also with sound fragments – which is of course unfeasible in old-school paper books of reference.

4 Enriching the Taalportaal with links to linguistic resources

Another manner of enriching the Taalportaal's grammatical information that is also unfeasible in traditional printed grammatical works of reference is to enrich it with links to digital linguistic resources within the CLARIN infrastructure. In a collaborative effort of the Meertens Institute, the Institute of Dutch Lexicology, the Universities of Groningen and Utrecht, and Taalmonsters, a motivated selection of Taalportaal texts has been enriched with links that encompass queries in corpus search interfaces (project CLARIN-NL15-001).

[4] http://viva-afrikaans.org/.

The Taalportaal contains, among other things, an online edition 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 descriptive, the emphasis in the selection and presentation of the phenomena discussed is clearly guided by discussions in the theoretical, more specifically the generative, literature (cf. Bouma et al. 2015).

In his largely positive review of the first SoD volumes on NP syntax, Hoeksema (2013) points out that "There is a growing body of work in empirical studies of judgment variation [...] that future extensions of this grammar could benefit from, especially when coupled to studies of actual usage patterns in corpus material" and that "This particular reader would also have welcomed to see some more lists in the book". By enriching the on-line version of SoD with queries over syntactically annotated corpora, the current project tries to accommodate the needs of researchers like Hoeksema.

Queries are linked to the following:

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

The queries are embedded in the Taalportaal texts as standard hyperlinks to other resources within the CLARIN network, where CLARIN supplies guidelines for things like a common vocabulary, common annotation standards, common interfaces, and single log-in. 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. In general, some kind of caching appears to be an option worth investigating.

Two tools are available for queries that are primarily syntactic in nature:

- The PaQU web application[5] (cf. Odijk 2015)
- The GrETEL web application[6] (cf. Augustinus et al. 2013).

Both tools can be used to search largely the same syntactically annotated corpora, viz. the Dutch spoken corpus CGN (van der Wouden et al. 2003) and the LASSY corpus of written text (van Noord et al. 2006), but they offer a slightly different functionality. Both applications offer dedicated user-friendly query interfaces (word pair relation search in PaQu and an example-based querying interface in GrETEL) as well as XPATH as a query language,[7] so that switching between these tools is trivial. Moreover, it is to be foreseen that future corpora of Dutch (and hopefully for Frisian as well) will be embedded in the very same CLARIN infrastructure, using the same architecture (cf. Landsbergen et al. 2014), the same type of interface and the same kind of linguistic annotation; the latter is the annotation schema for the Dutch Spoken Corpus CGN (cf. Schuurman et al. 2003, van der Wouden et al. 2003) which has become a de facto standard for Dutch corpus annotation, thus allowing for the re-use of the queries on these new data.

Translation of a linguistic example, a linguistic term, or a name or description of a construction is not a deterministic task that can be implemented in an algorithm. Rather, the queries are formulated by student assistants. After proper training, they get selections of the Taalportaal texts to read, interpret and enrich with queries where appropriate. The queries are 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. can be added. The student assistant's work is supervised by senior linguists.

Next to the annotated corpora mentioned above, access to two more linguistic resources have been investigated in TPC. On the one hand, there is the huge SONAR corpus (cf. Oostdijk et al. 2013). The size of this corpus (> 500 M tokens) makes it potentially useful to search for language phenomena that are relatively rare. In this corpus, however, (morpho-)syntactic annotations (pos-tags, inflectional

[5] http://portal.clarin.nl/node/4182.

[6] http://portal.clarin.nl/node/1967.

[7] https://en.wikipedia.org/wiki/XPath.

properties, lemma) are restricted to tokens (i.e., occurrences of inflected word forms). It comes with its own interface,[8] which allows queries in (a subset of) the Corpus Query Processing Language[9] and via a range of interfaces of increasing complexity. The original interface was not directly suited for linking queries as proposed here. For that reason, an update of this interface has been made to make the relevant queries possible.[10]

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 texts 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) that 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 will have to have yet another format, but again, the Taalportaal user will not be bothered with the gory details.

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

Since the links with the queries always go via the corpus search applications' *front-ends*, the Taalportaal user will, when a link has been clicked, be redirected not only to actual search results but also to a corpus search interface. The user can, if desired, adapt the query to better suit his/her needs, change the corpus being searched, search for constructions or sentences that diverge in one or more aspects (features) from the original query, or enter a completely new one. Most applications used (viz. PaQu, GrETEL, and OpenSONAR) have multiple interfaces differing in pre-supposed background knowledge of the user, and we believe that such options will actually be used. In this way, the enrichment of the Taalportaal as described here not only provides linguist users with actual corpus examples of linguistic phenomena, but may also have an educational effect of making the user acquainted with the existing corpus search interfaces.

5 An example

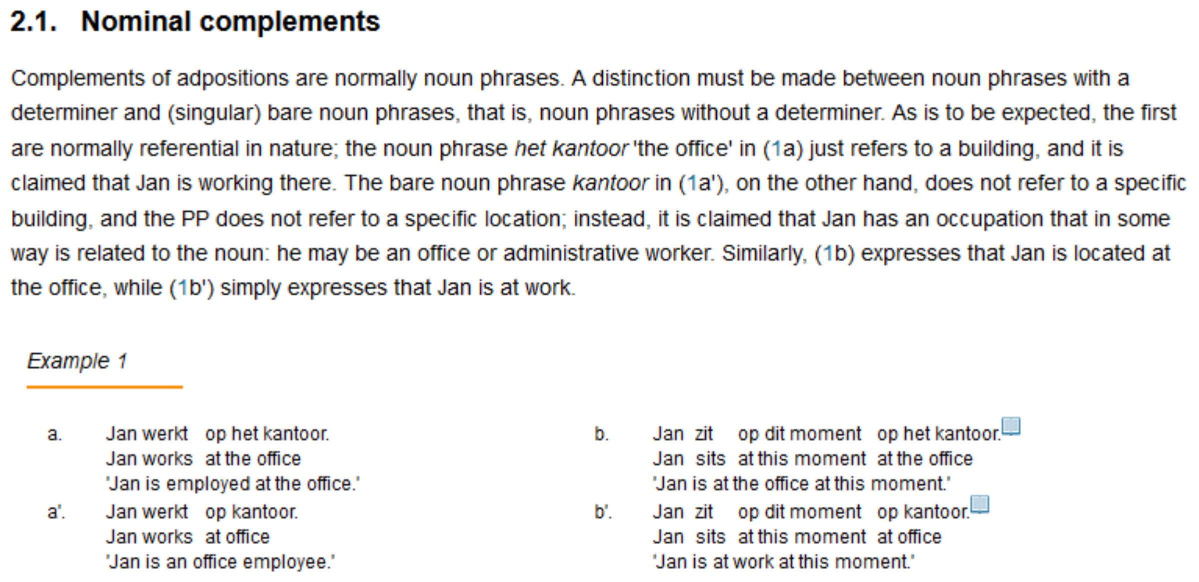
To get the gist of our approach, we will discuss a little example here. The beginning of the Taalportaal's chapter on nominal complements of adpositions[11] discusses the fact that both full noun phrases and bare nouns are possible as complements of prepositions. This is explained in terms of referentiality: in the variant with the full noun phrase *Jan werkt op het kantoor* (Jan works at the office) 'Jan is employed at the office' the noun phrase *het kantoor* 'the office' just refers to a building, and it is claimed that Jan is working there, whereas in the variant with a bare noun *Jan werkt op kantoor* (John works at office) 'Jan is an office employee', the prepositional phrase *op kantoor* 'at office' does not refer to a specific location. Figure 3 shows the relevant fragment.

[8] OpenSONAR via http://portal.clarin.nl/node/4195.

[9] Cf. http://cwb.sourceforge.net/files/CQP_Tutorial.

[10] https://portal.clarin.inl.nl/opensonar_whitelab/search/.

[11] http://taalportaal.org/taalportaal/topic/link/syntax__Dutch__adp__adp2__p2_compl.2.1.xml.



2.1. Nominal complements

Complements of adpositions are normally noun phrases. A distinction must be made between noun phrases with a determiner and (singular) bare noun phrases, that is, noun phrases without a determiner. As is to be expected, the first are normally referential in nature; the noun phrase *het kantoor* 'the office' in (1a) just refers to a building, and it is claimed that Jan is working there. The bare noun phrase *kantoor* in (1a'), on the other hand, does not refer to a specific building, and the PP does not refer to a specific location; instead, it is claimed that Jan has an occupation that in some way is related to the noun: he may be an office or administrative worker. Similarly, (1b) expresses that Jan is located at the office, while (1b') simply expresses that Jan is at work.

Example 1

a. Jan werkt op het kantoor.
Jan works at the office
'Jan is employed at the office.'

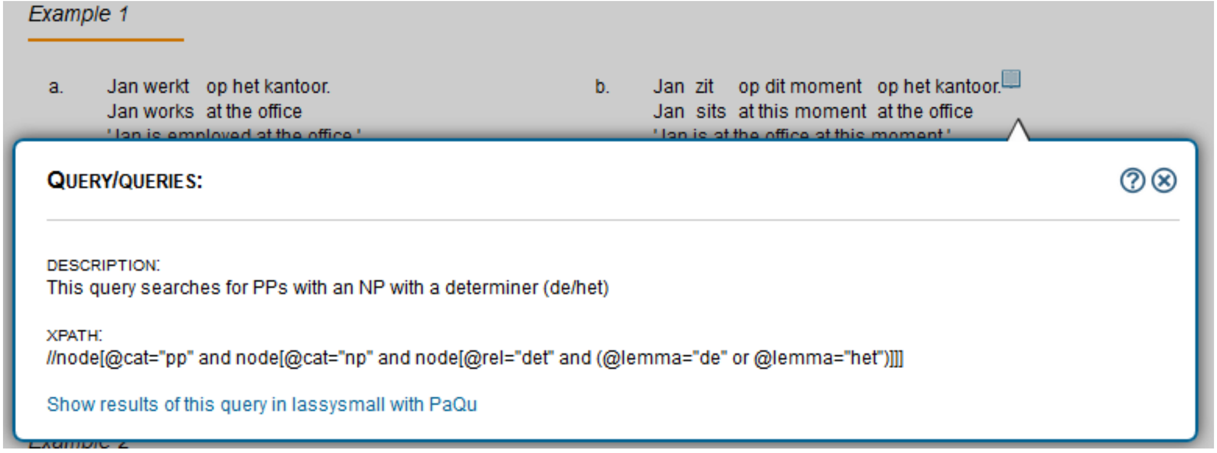
a'. Jan werkt op kantoor.
Jan works at office
'Jan is an office employee.'

b. Jan zit op dit moment op het kantoor.
Jan sits at this moment at the office
'Jan is at the office at this moment.'

b'. Jan zit op dit moment op kantoor.
Jan sits at this moment at office
'Jan is at work at this moment.'

Figure 3: Nominal complements of adpositions.

The b-examples are enriched with corpus queries, as indicated by the little icons. Clicking the first icon opens a pop-up window, illustrated in Figure 4:



Example 1

a. Jan werkt op het kantoor.
Jan works at the office

b. Jan zit op dit moment op het kantoor.
Jan sits at this moment at the office

QUERY/QUERIES:

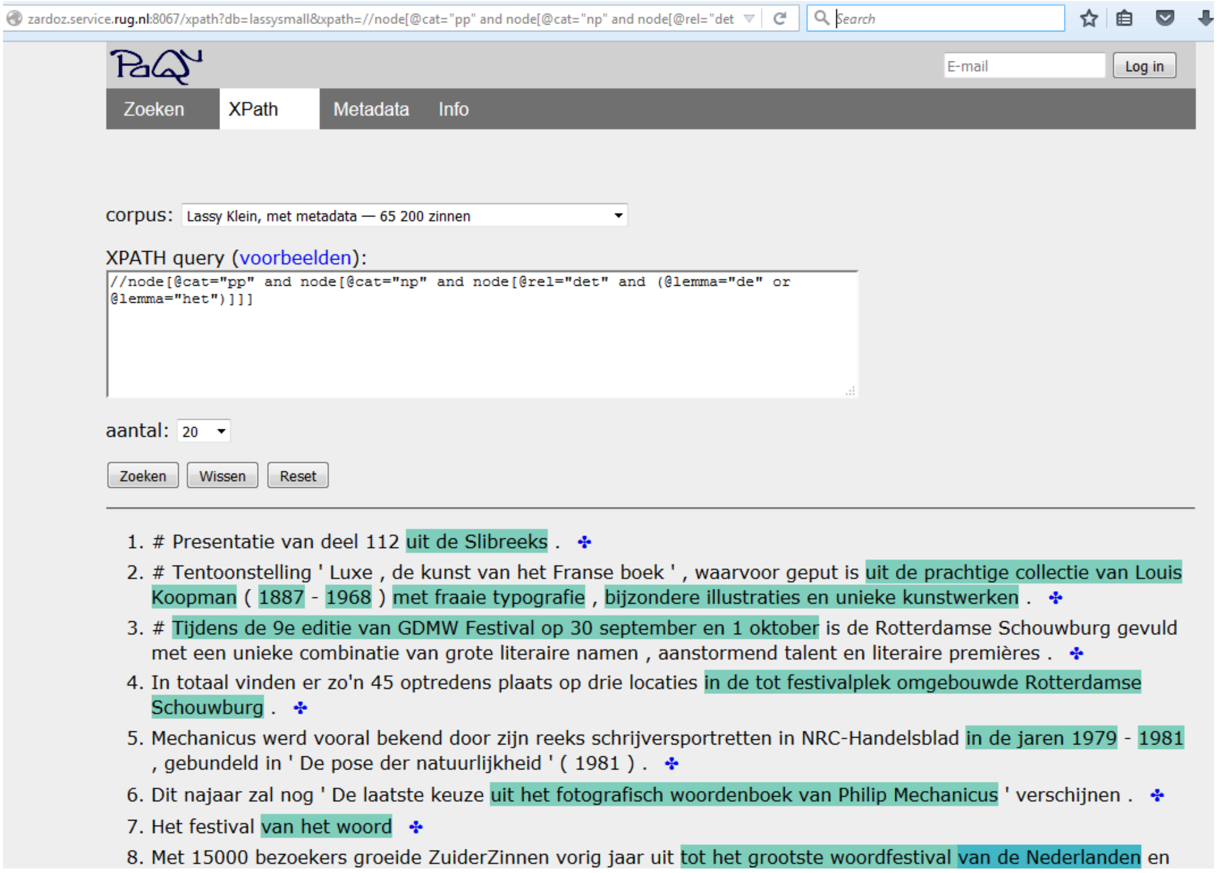
DESCRIPTION:
This query searches for PPs with an NP with a determiner (de/het)

XPATH:
//node[@cat="pp" and node[@cat="np" and node[@rel="det" and (@lemma="de" or @lemma="het")]]]

Show results of this query in lassysmall with PaQu

Figure 4: The query pop-up window.

The pop-window offers a brief description, showing the annotator’s interpretation of the text fragment, and the exact query in XPATH. The final line of the window is a direct link to the exact query. Clicking this link opens a new window in the user’s internet browser that shows the result of the query, as illustrated in Figure 5.



zardoz.service.rug.nl:8067/xpath?db=lassysmall&xpath=//node[@cat="pp" and node[@cat="np" and node[@rel="det

corpus: Lassy Klein, met metadata — 65 200 zinnen

XPATH query (voorbeelden):

```
//node[@cat="pp" and node[@cat="np" and node[@rel="det" and (@lemma="de" or
@lemma="het")]]]
```

aantal: 20

1. # Presentatie van deel 112 uit de Slibreeks .
2. # Tentoonstelling ' Luxe , de kunst van het Franse boek ' , waarvoor geput is uit de prachtige collectie van Louis Koopman (1887 - 1968) met fraaie typografie , bijzondere illustraties en unieke kunstwerken .
3. # Tijdens de 9e editie van GDMW Festival op 30 september en 1 oktober is de Rotterdamse Schouwburg gevuld met een unieke combinatie van grote literaire namen , aanstormend talent en literaire premières .
4. In totaal vinden er zo'n 45 optredens plaats op drie locaties in de tot festivalplek omgebouwde Rotterdamse Schouwburg .
5. Mechanicus werd vooral bekend door zijn reeks schrijversportretten in NRC-Handelsblad in de jaren 1979 - 1981 , gebundeld in ' De pose der natuurlijkheid ' (1981) .
6. Dit najaar zal nog ' De laatste keuze uit het fotografisch woordenboek van Philip Mechanicus ' verschijnen .
7. Het festival van het woord
8. Met 15000 bezoekers groeide ZuiderZinnen vorig jaar uit tot het grootste woordfestival van de Nederlanden en

Figure 5: Result of the query in Figure 4.

The researcher can study the examples, download them to his/her own computer, and/or edit the query if the result is not completely satisfactory.

6 Evaluation

After completion of approximately 1200 queries that cover the subchapters of the SoD on complementation and modification of adjectives and adpositions, we have learned that creating suitable queries for a given fragment from the SoD requires creativity and careful experimentation, tuning, and documentation (cf. Bouma et al. (2015) for details and statistics). Construction of queries is far from deterministic, that is, different annotators will have different opinions concerning the most suitable query for a given example or phenomenon. In a surprisingly high number of cases, there are mismatches (in constituent structure, in part-of-speech) between the presentation in the SoD and the treebank annotation. While this makes the development of queries harder, it also underlines the value of the current project: by systematically exploring the way various linguistic examples are annotated in the treebank, we provide a starting point for further corpus exploration for users that have a general linguistic interest but who are not necessarily experts on Dutch treebank annotation.

The manually verified treebanks almost always provide sufficient examples of basic word order patterns for queries that are not restricted to a specific adjective or preposition. For queries that search for a specific lexical head or for less frequent word order patterns, the Lassy Large treebank usually has to be used. In that case, users must be prepared to see also a certain number of false hits. However, there are also examples in the SoD that cannot be found in a 700M word corpus. The conclusion that such word orders are not found in the language would be too strong, but it might be a starting point for further research (i.e. *does this construction occur only in certain registers or discourse settings?*) or for an alternative analysis (i.e. *do these cases really involve adjectives?*).

During the process of formulating corpus queries, the student assistants also reported to have run into serious problems:

- There is a certain "mismatch" between the phenomena described by linguists and the phenomena found most often in the wild.
- There is a mismatch between grammar formalisms used by grammarians and the grammar formalisms used by corpus linguists (generative style in the case of Broekhuis, dependency style in the Dutch corpora used).
- The grammars use more semantics than can be handled by/is encoded in the corpora.

These problems are not without scientific interest. Annotated corpora deal with types of annotation that are encoded relatively easily without too many errors. This implies, among other things, that a lot of the semantic subtleties discussed in grammars are not addressed in current corpora. Moreover, the language described by grammar (albeit called "descriptive") turns out to be not exactly the same as language as covered by corpora. This holds for the corpora used in the experimental project described in this project – although they deal both with spoken and written language varieties – and it will probably hold for all corpora: "Grammars describe the things grammarians are used to describe", and for good reasons, albeit often biased through discussions (and fashions) in the theoretical literature. Parasitic gaps are a prime example of an extremely rare phenomenon (in the wild) with very serious theoretical consequences (cf. Engdahl 1983, Phillips 2006).

7 Concluding remarks

We have described goals and methods of CLARIN-NL15-001, a co-operation project to enrich the on-line Taalportaal (Language Portal) grammatical database with intelligent links that take the form of annotated queries in a number of on-line language corpora and an on-line linguistic morphophonological database. The project contributes to the research infrastructure for linguistics and related scientific disciplines in various ways, since

- It provides users with actual corpus examples for linguistic phenomena described in Taalportaal;
- It points out the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL and OpenSONAR to linguists;
- By redirecting the user to these front-ends, it stimulates the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries. Together with the multiple interfaces of most of these applications, this may also have a significant educational role.

8 Acknowledgements

The Taalportaal project is a collaboration 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 Groot 175.010.2009.003).

The project Enriching a grammatical database with intelligent links to linguistic resources (TPC) is a collaboration of Groningen University, Utrecht University, the Institute for Dutch Lexicology (INL), Taalmonsters and the Meertens Institute. It is financed by a grant of CLARIN-NL (NL-15-001: TPC).

References

Liesbeth Augustinus, 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, pp. 423–428.

R. Harald Baayen, Richard Piepenbrock, & L. Gulikers. 1995. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Gosse Bouma, Marjo van Koppen, Frank Landsbergen, Jan Odijk, Ton van der Wouden, & Matje van de Camp. 2015. Enriching a Descriptive Grammar with Treebank Queries. In: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 13–25.

Hans Broekhuis, Norbert Corver, Marcel den Dikken, Evelien Keizer, & Riet Vos. 2012. *Syntax of Dutch*. Amsterdam: Amsterdam University Press, 2012–16 (7 volumes).

Elisabet Engdahl. 1983. Parasitic gaps. *Linguistics and Philosophy* 6, pp. 5–34.

Walter Haeseryn, Kirsten Romijn, Guido Geerts, Jaap de Rooij, & Maarten C. van den Toorn (eds.). 1997. *Algemene Nederlandse Spraakkunst*. Groningen and Deurne: Martinus Nijhoff and Wolters Plantijn. 2nd rev. ed. (2 vols.).

Jack Hoeksema. 2013. Review of: Syntax of Dutch. Noun and Noun Phrases vols. 1 and 2. *Lingua*, 133, pp. 385–390.

Frank Landsbergen, Carole Tiberius, & Roderik Dernison. 2014. Taalportaal: an online grammar of Dutch and Frisian. In Nicoletta Calzolari et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland, pp. 26–31. ELRA.

Gertjan van Noord, Ineke Schuurman, & Vincent Vandeghinste. 2006. Syntactic Annotation of Large Corpora in STEVIN. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* pp. 1811–1814. ELRA.

Jan Odijk. 2015. Linguistic Research with PaQU. *Computational Linguistics in The Netherlands Journal* 5, pp. 3–14.

Nelleke Oostdijk, Martin Reynaert, Veronique Hoste, Ineke Schuurman. 2013. The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In P. Spyns and J. Odijk (eds.): *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer Verlag, pp. 219–247.

Colin, Phillips. 2006. The real-time status of island phenomena. *Language* 82, pp. 795–823.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and 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, pp. 340–347.

Ineke Schuurman. 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.

Ton van der Wouden, Ineke Schuurman, Machteld Schouppe, and Heleen Hoekstra. 2003. Harvesting Dutch trees: Syntactic properties of spoken Dutch. In *Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting*, ed. by Tanja Gaustad, pp. 129–141. Amsterdam/New York: Rodopi.