Enriching a grammatical database with intelligent links to linguistic resources

Ton van der Wouden Meertens Institute The Netherlands Ton.van.der.wouden@meertens.knaw.nl	Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, and Jan Odijk

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 how the on-line Taalportaal (Language Portal) grammatical database is 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 contributes to the use of the CLARIN research infrastructure, 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 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 since the 1960 for languages all around the globe. These days, there is 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, 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.

To enhance the Taalportaal's value, the CLARIN project described here (NL-15-001: TPC) seeks to enrich the grammatical information within the Taalportaal with links to linguistic resources. The idea is that the user, while reading a grammatical description or studying a linguistic example, is 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. We explicitly focus on resources with rich linguistic annotations, 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.

In light of the restricted resources, in terms both of time and money, this CLARIN project is not aiming at exhaustivity, that is, not all grammatical descriptions and not all examples will be provided with query links. TPC is thus 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 corpora.

3 The Taalportaal

The Taalportaal project (www.taalportaal.org) 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. The Taalportaal is built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian. The Taalportaal's prime intended audience is the international scientific community, which is why the language used to describe the language facts is English. The Language Portal will 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 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 Language Portal contributes to the integration of the results reached within these disciplines.

Technically, the Taalportaal is an XML-database that is accessible via any internet browser. Organization and structure of the linguistic information will be reminiscent of, and is is to a large extend inspired by, Wikipedia and comparable online information sources. An important difference, however, is that Wikipedia's democratic (anarchic) model is avoided by restricting the right to edit the Taalportaal information to authorized experts.

4 Enriching the Taalportaal with links to linguistic resources

CLARIN-NL15-001 is a collaborative effort of the Meertens Institute, the Institute of Dutch Lexicology, the Universities of Groningen and Utrecht, and Taalmonsters. In this pilot project, a motivated selection of Taalportaal texts will be enriched with links that encompass queries in corpus search interfaces. Queries will be linked to

- Linguistic examples
- Linguistic terms
- 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. 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
- The GrETEL web application (cf. Augustinus et al. 2013)

Both can be used to search largely the same syntactically annotated corpora (viz. the Dutch spoken corpus CGN (cf. van der Wouden et al. 2003) and the (written) LASSY corpus (cf. 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 (cf. https://en.wikipedia.org/wiki/XPath), 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 type of interface, allowing for the reuse 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. The page http://www.clarin.nl/node/2080 already contains a small example of a Taalportaal fragment adorned with a few query links using PaQu.

Next to the annotated corpora mentioned above, access to two more linguistic resources will be 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 properties, lemma) are restricted to tokens (i.e., occurrences of inflected word forms). It comes with its own interface (OpenSONAR), which allows queries in (a subset of) the Corpus Query Processing Language and via a range of interfaces of increasing complexity. The current interface is 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. This updated version is available (as a beta version) on http://zilla.taalmonsters.nl:8080/whitelab/search/simple.

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.

5 Concrete Examples

The final paper will contain several concrete examples of Taalportaal descriptions and the links to their associated queries; we hope to demonstrate the system at the conference.

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. Since most applications used (viz. PaQu, GrETEL, and OpenSONAR) have multiple interfaces differing in pre-supposed background knowledge of the user, 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.

6 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 will contribute to the research infrastructure for linguistics and related scientific disciplines, 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 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.

References

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and 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. pp. 423-428.

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

Frank Landsbergen, Carole Tiberius, and Roderik Dernison: 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, European Language Resources Association (ELRA), 2014, 26-31.

Gertjan van Noord, Ineke Schuurman, Vincent Vandeghinste. Syntactic Annotation of Large Corpora in STEVIN. In: LREC 2006 (http://www.lrec-conf.org/proceedings/lrec2006/) .

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

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, 129-141. Amsterdam/New York: Rodopi.