PART III

Infrastructure for Syntax

CHAPTER 17

Infrastructure for Syntax: Introduction

Jan Odijk
UiL-OTS, Utrecht University, j.odijk@uu.nl

17.1 Introduction

A lot of work has been done by CLARIN-LC to fill CLARIN with data and applications to support syntactic research. For this reason, a separate part of this book is dedicated to this topic.

The chapters in this part only partially cover the work done in CLARIN-LC to support syntactic research. I will first briefly describe the work done that is not covered by the chapters in this part (section 17.2), and then introduce the chapters (section 17.3).

17.2 Work on Syntax

Some data and applications have a broader scope than syntax but are nevertheless highly relevant for syntactic research. The Typological Database System (TDS, described in chapter 11), provides the user with integrated access to a collection of independently developed typological databases which also contain syntactic properties. The TTNWW application, described in chapter 7, provides a wide range of workflows for enriching text corpora, but some of these workflows are highly relevant for syntactic research, such as the workflows for tokenization, lemmatization, part-of-speech tagging, and parsing of modern Dutch texts. Tools for tokenising, lemmatising, part of speech tagging and parsing mediaeval Dutch have been made available through Adelheid and INPOLDER.

Most applications for syntactic research focus on search for syntactic properties. The DuELME (Odijk, 2013a;b) data and associated search application enable a user to search for syntactic properties of Dutch multiword expressions. FESLI has provided a corpus with monolingual and bilingual children (Dutch - Turkish) with and without Specific Language Impairment, enriched with part of speech annotations on tokens, and an application for searching for morpho-syntactic properties in this corpus.

How to cite this book chapter:
Odijk, J. 2017. Infrastructure for Syntax: Introduction. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 213–215. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.17.

AutoSearch enables a user to search in one's own corpora once they have been enriched with part of speech tags using TTNWW (see chapter 7), and Nederlab enables a user to search in all digitised texts relevant for the Dutch national heritage and the history of Dutch language and culture (ca 800 - present).

The MIMORE application enables combined searching in and analysis of three databases on dialect variation (see Barbiers et al. (2016) for an example of research that crucially used this application).

Finally, the LASSY Word Relations Search application (Tjong Kim Sang et al., 2010) enables a user to search for syntactic dependency triples in specific Dutch treebanks.

17.3 Contents of Part III: Infrastructure for Syntax

All chapters focus on applications for search in and analysis of syntactically annotated corpora.

Some applications enable search in corpora with linguistic annotations on tokens, e.g. part of speech codes. Chapter 18 describes the SHEBANQ application, which enables search in the WIVU Hebrew Bible Text Database. It illustrates a typical project in which certain data, originally stored in an idiosyncratic format, have been curated and converted to a CLARIN-supported format based on the Linguistic Annotation Framework (LAF; Ide and Suderman, 2014), and a web application has been built for searching in these data, either by creating queries oneself, or by reusing queries created by others and stored here.

Chapter 19 describes the application interface (front-end) of the OpenSoNaR application, which enables search in the large scale Dutch reference corpora SoNaR and SoNaR New Media (Oostdijk et al., 2013), while chapter 20 describes the search engine (back-end) of this application. The second version of this application also provides access to the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; Oostdijk et al., 2002). The OpenSoNaR web application, with multiple interfaces varying in complexity, opens up the SoNaR corpora for research by humanities scholars, who until recently could access these data only with great difficulty. For the CGN data an exploration application existed already (called COREX; Hellwig and Weijers, 2004), but it was developed more than 12 years ago, is a desktop application, and is not being maintained anymore.

Another set of applications enable search in text corpora in which each sentence has been assigned a syntactic structure (treebanks). One example is CorpusStudioWeb, which not only enables searching in treebanks, but offers additional important functionality for researchers, in particular keeping a number of related searches together in a search 'project' and annotating search results automatically or semi-automatically. This is described in chapter 21.

GrETEL is a web application for search in treebanks in which a user can create queries on the basis of an example sentence (example-based search) and does not have to know all annotation guidelines of the treebank or even a formal query language. It is described in chapter 22.

Chapter 23 describes PaQu and a small case study using PaQu. PaQu is an extension of the LASSY Word Relations Search application (Tjong Kim Sang et al., 2010), and offers searching for grammatical dependency triples in the user's own corpora.

The Taalportaal is a comprehensive and authoritative digital scientific grammar for Dutch, Frisian, and Afrikaans. In the Taalportaal, links to several search applications were made to provide concrete evidence related to specific constructions described in the Taalportaal. The links cover morphology and phonology, but the links to stored queries in treebanks are most prominent. This is the topic of chapter 24.

Acknowledgements

This work was financed by CLARIN-NL and CLARIAH.

References

Barbiers, Sjef, Marjo van Koppen, Hans Bennis, and Norbert Corver (2016), MIcrocomparative MOrphosyntactic REsearch (MIMORE): Mapping partial grammars of Flemish, Brabantish and Dutch, *Lingua* **178**, pp. 5–31. Linguistic Research in the CLARIN Infrastructure. http://www.sciencedirect.com/science/article/pii/S0024384115002211.

Hellwig, Birgit and Erik Weijers (2004), COREX: A tool for exploiting the Corpus Gesproken Nederlands (CGN), *Manual*, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. 2nd version. http://www.mpi.nl/corpus/manuals/manual-corex.pdf.

Ide, Nancy and Keith Suderman (2014), The linguistic annotation framework: A standard for annotation interchange and merging, *Language Resources and Evaluation* **48** (3), pp. 395–418.

Odijk, Jan (2013a), DUELME: Dutch electronic lexicon of multiword expressions, *in* Francopoulo, G., editor, *LMF - Lexical Markup Framework*, ISTE / Wiley, London, UK / Hoboken, US, pp. 133–144.

Odijk, Jan (2013b), Identification and lexical representation of multiword expressions, *in* Spyns, P. and J.E.J.M Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, Springer, Berlin/Heidelberg, pp. 201–217. http://link.springer.com/content/pdf/10.1007%2F978-3-642-30910-6_12.

Oostdijk, N., M. Reynaert, V. Hoste, and I. Schuurman (2013), The construction of a 500 million word reference corpus of contemporary written Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer, Berlin, pp. 219–247. http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1.

Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen (2002), Experiences from the Spoken Dutch Corpus project, *in* González Rodriguez, M. and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, ELRA, Las Palmas, pp. 340–347.

Tjong Kim Sang, Erik, Gosse Bouma, and Gertjan van Noord (2010), LASSY for beginners, Presentation at CLIN 2010, Utrecht, http://ifarm.nl/erikt/talks/clin2010.pdf. http://ifarm.nl/erikt/talks/clin2010.pdf.