PART II

Infrastructure for Linguistics

CHAPTER 9

Infrastructure for Linguistics: Introduction

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

9.1 Introduction

Given its origins in linguistics and language technology, it should come as no surprise that CLARIN-LC created many infrastructural facilities for linguistics. These will be discussed in this part of the book, with the exception of infrastructural facilities for syntax, to which a separate part of this book is dedicated (Part III).

The chapters in this part only partially cover the work done in CLARIN-LC to support linguistic research. I will first provide a brief overall overview of the relevant data and software that resulted from CLARIN-LC (section 9.2), and then summarise the topics of the chapters of this part (section 9.3).

9.2 Work on Linguistics

I have categorised the various datasets and software applications into a number of categories in order to structure the description. This categorisation is in many respects somewhat arbitrary, but nevertheless groups the resources in natural classes. The categories are drawn from different dimensions, and are thus not mutually exclusive. Many datasets and applications can be used for multiple purposes and therefore belong to multiple categories. For this reason, some resources are mentioned under multiple categories.

The categories we distinguish are on the one hand subdisciplines of linguistics, such as *language documentation*, *language variation*, *language acquisition*, *lexicography*, and *discourse and stylistics*; and on the other hand types of software functionality, such as *enrichment*, *annotation* and *search*.

How to cite this book chapter:
Odijk, J. 2017. Infrastructure for Linguistics: Introduction. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 107–111. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.9.

Language documentation LAISEANG provides an unrivalled collection of multimedia materials and written documents from 48 languages of Insular South East Asia and West New Guinea (see chapter 10). The Typological Database System (TDS, described in chapter 11), provides the user with integrated access to a collection of independently developed typological databases. NEHOL is a digitally accessible and searchable database of the Dutch-lexifier Creole language Negerhollands or Virgin Islands Dutch Creole, VIDC (see chapter 12).

Language Variation DBD/TCULT: The DBD is a rather substantial collection of data (over 1,500 sessions) from a number of projects and research programmes that were directed at investigating multilingualism and comprises data originating from Dutch, Sranan, Sarnami, Papiamentu, Arabic, Berber and Turkish speakers. The basis of the collection is the data from the TCULT project in which intercultural language contacts in the Dutch city of Utrecht were studied. MIMORE integrates three different but related dialectical databases: DynaSAND (the dynamic syntactic atlas of the Dutch dialects), DiDDD (Diversity in Dutch DP Design) and GTRP (Goeman, Taeldeman, van Reenen Project). The associated MIMORE application enables combined searching in and analysis of these three databases. Barbiers et al. (2016) is an example of research that crucially uses this application. Analysis of data on migration flow between Dutch municipalities via MIGMAP (chapter 29) can be used for analysing language variation within the Netherlands, the influence of one dialect on another, etc. Gabmap (Leinonen et al., 2016) enables a researcher to automatically analyse data of language variation, e.g. varying words for the same concepts, varying pronunciations for the same words, or varying frequencies of syntactic constructions in transcribed conversations. D-LUCEA is a database of speech recordings of native and non-native speakers of English. The recorded speakers are students from an international student community where English is used as lingua franca. These students are being recorded longitudinally throughout their 3-year period on campus (see chapter 15).

Some resources are concerned with variation across time: VU-DNC is a unique diachronic corpus of Dutch newspaper articles from five major Dutch newspapers from 1950/1951 and 2002. 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).

Language Acquisition COAVA enables a user to search a combination of historical dialect data and first language acquisition data. (Cornips et al., 2016) describe a case study carried out with this application and the underlying datasets. FESLI enables a user to search in the FESLI data, which have been enriched with part of speech tags. These data are from monolingual and bilingual (Dutch - Turkish) children with and without Specific Language Impairment. LESLLA contains speech of 15 low-educated learners of Dutch as a second language. VALID is an open access multimedia archive of language pathology data collected in the Netherlands, primarily on Dutch, with audio files and transcripts. D-LUCEA was mentioned above under *language variation* but it is evidently also relevant to second language acquisition (see chapter 15).

Lexicography Many lexical data have been curated and/or made accessible through user-friendly web applications. They include the Dictionary of the Brabantic Dialects, Part III, General Vocabulary, (WBD); the Dictionary of the Limburgian Dialects, Part III, General Vocabulary (WLD); the Frisian dictionary of WFT-GTB integrated in the language bank and accessible via the language bank web application (see chapter 13); the multiword expression lexical database of DUELME and its associated DuELME web application (Odijk, 2013a;b); the GrNe web application for the classical GrNe Greek-Dutch dictionary (originally for the letter π only, but for an increasing number of letters); and the lexical data of Cornetto, a lexical resource for the Dutch language which combines two resources with different semantic organisations: the *Dutch Wordnet* with its synset organisation and the *Dutch Reference Lexicon* which includes

definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc. The Cornetto data are easily accessible via the dedicated Cornetto web application developed in CLARIN-LC.

Obviously, many search applications described elsewhere in this book, such as OpenSoNaR, AutoSearch, Nederlab, PaQu, GrETEL, CorpusSearchWeb, SHEBANQ, and MIMORE also support lexicography and lexicological research.

Discourse and Stylistics The diachronic VU-DNC corpus of Dutch newspaper articles from five major Dutch newspapers from 1950/1951 and 2002 is annotated not only with part of speech codes but also with discourse annotations. The DiscAn corpus is a collection of subcorpora of the Dutch language specifically created as a corpus annotated at the level of discourse, in particular for coherence relations and discourse connectives.

Stylene is a system for stylometry and readability research on the basis of existing techniques for automatic text analysis and machine learning, and offers a web service that allows researchers in the Humanities and Social Sciences to analyse texts with this system. It is described in more detail in chapter 16.

Enrichment There are many applications and services for enriching data. These include a web application and service for orthographic normalisation (TICClops), which is also embedded in a workflow for converting digital images into textual resources in TEI[1] format (@PhilosTEI, described in more detail in chapter 32)

The TTNWW application, described in chapter 7, provides a wide range of workflows for enriching text corpora with linguistic annotations, among them workflows for tokenisation, lemmatisation, named entity recognition, coreference marking, and marking of semantic roles, as well as workflows for enriching an audio file with an automatically generated orthographic transcription. NameScape enables a researcher to have a text corpus enriched with annotations for named entities.

OpenConvert consists of a set of web services for format conversions between a variety of formats for textual resources, thus enabling a wide variety of formats to be processed by applications such as TTNWW.

Manual annotation A number of applications focus on annotating resources, i.e. manually (or semi-automatically) enriching them with new information. This was described in chapter 2, but is repeated here. Prominent in CLARIN-LC are the ELAN, and ANNEX applications for the creation of complex annotations on video and audio resources. These applications existed before CLARIN-LC but were significantly improved and enhanced in CLARIN-LC. These enhancements include a web service (AAM-LR) for annotating where in an audio file there is speech (instead of other sounds), and identifying who is speaking in the parts containing speech (diarisation). In the SignLinC project it was made possible to link lexical databases and annotated corpora of signed language in these applications. The ColTime project extended ELAN and ANNEX with a referencing and note exchanging system. The EXILSEA project enhanced these applications for users of different languages with multilingual features based on ISOcat. The MultiCon project enhanced ELAN and ANNEX with multilayer visualisation of multilayer collocates. TQE is a web application for evaluating the quality of phonetic transcriptions of speech files.

The FLAT application described in chapter 6 is an application for manual verification and correction of annotations on text corpora encoded in the FoLiA format.

Several of the tools described under *Enrichment* can also be used for annotation purposes. They can bootstrap the annotation by automatically enriching a resource with annotations, followed by manual verification and correction, e.g. through FLAT, ELAN or ANNEX.

[1] TEI (Text Encoding Initiative) is a widely used standard for encoding textual resources supported by CLARIN.

Search The search applications OpenSoNaR, AutoSearch, Nederlab, PaQu, GrETEL, CorpusSearchWeb, SHEBANQ, and MIMORE will be described in more detail in part III on infrastructure for syntax, but they can obviously also be used for linguistic research other than syntax, e.g. for lexicography, morphology and semantics, and some even for phonology and phonetics. NameScape enables searching for names and analysing their use in literary works (see chapter 30). 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 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. Besides syntax, the links cover morphology and phonology (see chapter 24).

9.3 Contents of Part II: Infrastructure for Linguistics

Part II of this book covers only a small sample of the resources described in section 9.2. For the reader's convenience, we describe the contents of each chapter here:

Chapter 10 describes the LAISEANG collection of multimedia materials and written documents from 48 languages in Insular South East Asia and West New Guinea. The language resources for this collection were gathered by 20 linguists at or in collaboration with Dutch universities over the last 40 years, and were compiled and archived in collaboration with The Language Archive (TLA) at the Max Planck Institute in Nijmegen in accordance with CLARIN standards.

Chapter 11 describes the curation of the Typological Database System (TDS), which provides the user with integrated access to a collection of independently developed typological databases. Curating this independently developed database system was urgently needed to save this valuable resource in a durable, archival environment and convert access to it into a true web service architecture, thus safeguarding future access to the data.

Chapter 12 investigates variation in supra-locative prepositional phrases in two varieties of VIDC, crucially using the NEHOL database curated in CLARIN-LC.

Chapter 13 discusses the use of the WFT dictionary after its integration into the GTB language bank. The authors demonstrate a case of usage for research, and suggests possible improvements and expansions of the online version of the WFT.

Chapter 14 reports on research aiming to determine which language measures are diagnostic indicators of SLI on the basis of narrative data. To that end, morphosyntactic and lexical accuracy and complexity were investigated, crucially using the VALID open access multimedia archive of language pathology data. The authors argues that their results reveal the urgency to have identical, precise protocols in handling and analysing complex data, which is exactly one of the goals of the VALID archive.

Chapter 15 describes the UCU Accent project, and the curation of the resulting D-LUCEA database, which made the recorded speech data and their concomitant metadata widely available to the research community at large. The authors describe some of the research that has been made possible via this project, as well as current plans for applying a similar method for data curation to a new speech accent corpus, *Sprekend Nederland* (*The Netherlands Speaking*).

Chapter 16 describes Stylene, which consists of an educational demonstration interface and tools for stylometry (authorship attribution and profiling) and readability research for Dutch. Stylene is again a typical CLARIN result in that it makes advanced computational methods available in a user-friendly manner to researchers from the Humanities and Social Sciences.

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.

Cornips, Leonie, Jos Swanenberg, Wilbert Heeringa, and Folkert de Vriend (2016), The relationship between first language acquisition and dialect variation: Linking resources from distinct disciplines in a CLARIN-NL project, *Lingua* **178**, pp. 32–45. Linguistic Research in the CLARIN Infrastructure. http://www.sciencedirect.com/science/article/pii/S0024384115002375.

Leinonen, Therese, Çağrı Çöltekin, and John Nerbonne (2016), Using Gabmap, *Lingua* **178**, pp. 71–83. Linguistic Research in the CLARIN Infrastructure. http://www.sciencedirect.com/science/article/pii/S0024384115000315.

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.