CHAPTER 1

CLARIN in the Low Countries: Introduction

Jan Odijk and Arjan van Hessen
UiL-OTS, Utrecht University
j.odijk@uu.nl, A.J.vanHessen@uu.nl

ABSTRACT

In this chapter we introduce the notion of research infrastructure, CLARIN as a research infrastructure for the Humanities and Social Sciences, the CLARIN projects carried out in the Low Countries (the Netherlands and Flanders), and some closely related projects in the Netherlands. We end with a description of the structure of this book.

1.1 Introduction

This book describes the results of activities undertaken to construct the CLARIN research infrastructure in the Low Countries (CLARIN-LC), i.e., in the Netherlands and in Flanders (the Dutch-speaking part of Belgium).

The activities in the Netherlands were carried out mainly through the CLARIN-NL project, a national project for the design, construction, and exploitation of the Netherlands part of the European-wide CLARIN infrastructure. Through a proposal for joint activities on CLARIN between the Netherlands and Flanders for the (shared) Dutch language, it was possible to obtain some (small) funds for work on CLARIN in Flanders. Results of the close collaboration of the CLARIN communities in the Netherlands and Flanders, as well as results of independent activities of the Flanders CLARIN community, are included in this book.

In this chapter, we will provide some basic information on the background and history (section 1.2) of CLARIN, and its basic characteristics (section 1.3). In section 1.4 we describe the CLARIN projects in the Netherlands and in Flanders, as well as some independent but closely related projects. We sketch the structure of this book in section 1.5.

How to cite this book chapter:

Odijk, J and van Hessen, A. 2017. CLARIN in the Low Countries: Introduction. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 1–9. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.1. License: CC-BY 4.0

1.2 CLARIN: Historical Background

CLARIN Europe CLARIN is an acronym for *Common Language Resources and Technologies Infrastructure*. A proposal for CLARIN was submitted and accepted for inclusion in the 2006 ESFRI Roadmap.[1] A proposal for a CLARIN Preparatory Project coordinated by Utrecht University (CLARIN-PP, 2008–2011) was submitted and received funding from the European Commission.

Since February 2012 CLARIN is coordinated by CLARIN ERIC, hosted by the Netherlands. An *ERIC (European Research Infrastructure Consortium)* is a legal entity at the European level specifically set up for European research infrastructures. An ERIC has countries or intergovernmental organisations as its members. CLARIN ERIC has 19 members (Austria, Bulgaria, the Czech Republic, Denmark, the Dutch Language Union, Estonia, Finland, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Norway, Poland, Portugal, Slovenia, and Sweden), with the UK as an observer, and the number of members is growing.[2] Each ERIC member commits to paying the ERIC yearly fee and to contributing to the CLARIN infrastructure by setting up national projects to this end.

CLARIN in the Netherlands In the Netherlands the national CLARIN project was called *CLARIN-NL*. It ran from 2009 through 2015. Though the CLARIN-NL project finished in 2015, funding was obtained for two projects to continue and extend the work on infrastructures for the Humanities: CLARIAH-SEED (2013–2014), and CLARIAH-CORE (2015–2018).

CLARIN in Flanders In Flanders, the activities for CLARIN were funded from 2010 through 2012, in part thanks to a close collaboration with CLARIN in the Netherlands. Independently funded projects also made several contributions to CLARIN.

1.3 The CLARIN Infrastructure

The CLARIN infrastructure (from now on simply *CLARIN*) is a **research infrastructure** for **Humanities researchers** who work with **digital language resources**. We will explain each of the bold-faced terms.

Infrastructure refers to (usually large-scale) basic physical and organisational resources, structures and services needed for the operation of a society or enterprise.[3] Familiar examples are railway networks (Figure 1.1), road networks, electricity networks, but also (on a smaller scale) Eduroam[4], which provides world-wide wireless internet facilities through higher and further education organisations.

A **research infrastructure** is an infrastructure intended for carrying out research, i.e., facilities, resources and related services used by the scientific community to conduct top-level research. Famous examples are the European Extremely Large Telescope (E-ELT) in Chile (Figure 1.2) and the CERN Large Hadron Collider.

[1] ESFRI is an acronym for European Strategy Forum on Research Infrastructures (`http://www.esfri.eu/`).

[2] This is the situation on 15 November 2016. See https://www.clarin.eu/content/participating-consortia for an up-to-date overview.

[3] This description is an adaptation of the description from English-language Wikipedia (`http://en.wikipedia.org/wiki/Infrastructure`).

[4] We will provide hyperlinks in the text but usually not show the URL. People reading this chapter electronically can directly click on such links. People who read this chapter on paper do not want to copy the URLs by hand anyway, so they will turn to the electronic version if they want to follow a link.

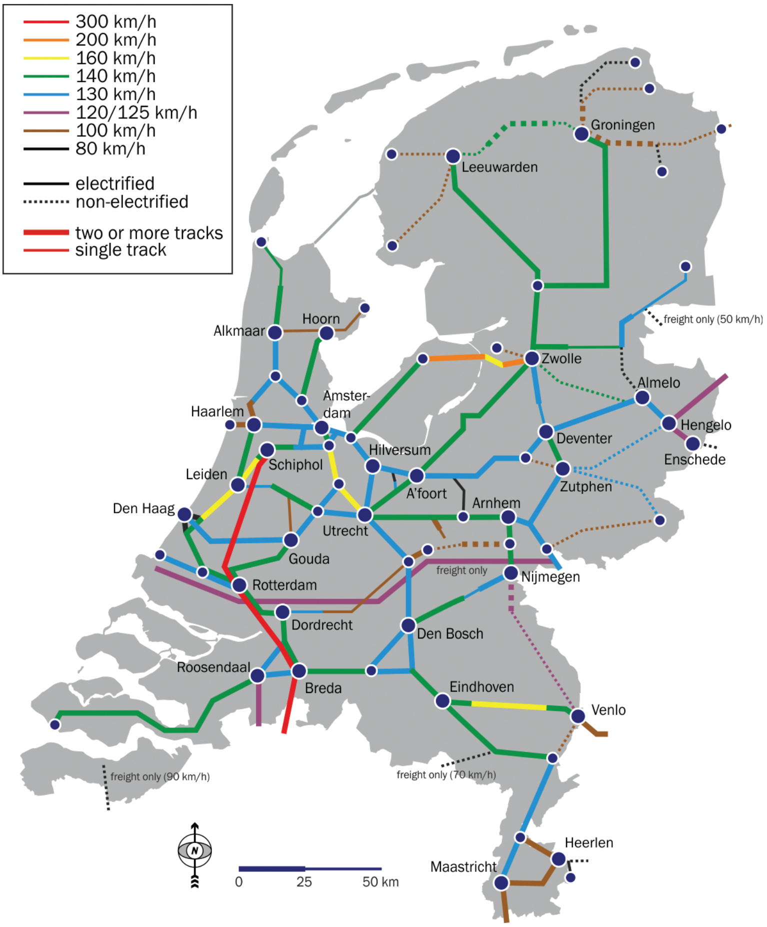

Figure 1.1: Dutch Railway Network (picture from Wikipedia).

Humanities researchers include linguists, historians (including art historians), literary scholars, philosophers, religion scholars, and others, as well as political science researchers, who are usually considered part of the Social Sciences.[5]

Digital language resources includes both data and software. They include a wide spectrum of digital data types:

- Data in natural language (texts, lexicons, grammars, etc.)
- Databases about natural language (typological databases, dialect databases, lexical databases, etc.)
- Audio-visual data containing (written, spoken, signed) language (e.g. pictures of manuscripts, audiovisual data for language description, descriptions of sign language, interviews, radio and TV programmes, etc.)

As for software, digital language resources include software dedicated to browsing and searching in digital language data (e.g. software to search in a linguistically annotated text corpus), as well

[5] CLARIN at the European scale is intended for the Humanities *and* the Social Sciences, but the Netherlands has focused on the Humanities.

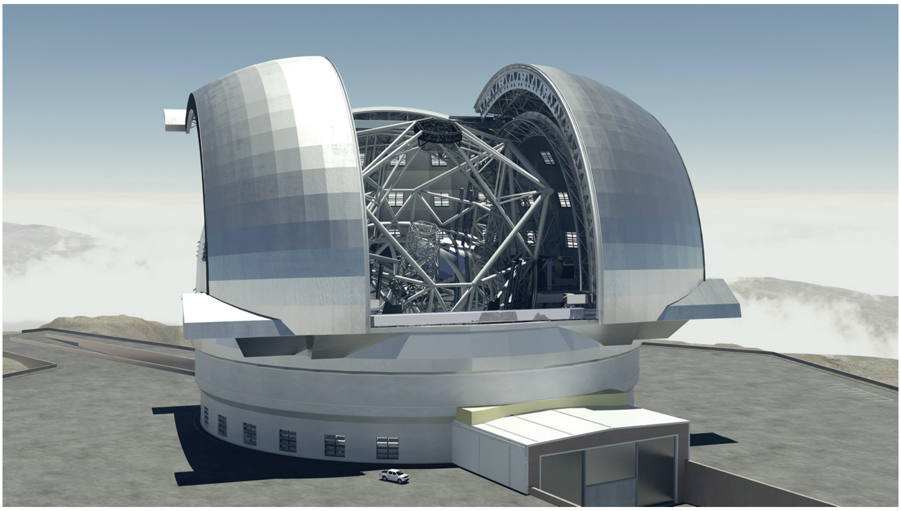

Figure 1.2: European Extremely Large Telescope in Chile (source: Wikipedia).

as software to analyse, enrich, process, and visualise digital language data, (e.g., a parser, which enriches each sentence in a text corpus with a syntactic structure). We will often use the term *resource* as shorthand for *digital language resource*.

CLARIN is intended for **language** in various functions, including:

- As an object of inquiry
- As a carrier of cultural content
- As a means of communication
- As a component of identity

Though the creation of data for research certainly is part of creating a research infrastructure, CLARIN-LC has **not** created any new data. It has mainly adapted existing data and software to make them compliant with CLARIN-requirements and interoperable, and it has created new user-friendly software for searching, analysing and visualising data.

CLARIN is not one big physical installation on a single location such as the CERN Large Hadron Collider or the Chile Large Telescope. On the contrary, CLARIN is

- a **distributed** infrastructure, which has been implemented as a network of **CLARIN centres**. A CLARIN centre is a centre that is certified as such and provides CLARIN services. The Netherlands has several such centres. These will be discussed in more detail in chapters 3 and 4.
- a **virtual** infrastructure, which provides services electronically (via the internet). Every user can use CLARIN from any location where (s)he has access to internet.[6]

Many applications have been developed that enable searching, enriching, analysing or visualising huge amounts of data. These applications are web applications, so that no software needs to be downloaded or installed on the local computer of a researcher. The data that the applications apply to are stored on servers at CLARIN centres, so that no data need to be downloaded and stored locally. This is important, because the huge size of the relevant data makes storing them locally

[6] Though CLARIN also makes available software that operates locally on a single computer. This is necessary in some cases where internet access is absent or limited.

increasingly more difficult. Many applications have been developed in CLARIN with multiple user-interfaces, ranging from very user-friendly interfaces intended for novice users to expert interfaces which offer full functionality but require expert knowledge, and many intermediate forms. In many search applications, queries formulated through a simple interface can also be seen in the more complex interfaces, so that more complex queries can be built up incrementally, starting with a relatively simple query in a simple interface, and extended and refined in one of the more advanced interfaces. In this way CLARIN makes many more data accessible to researchers than ever before, and the researchers can actually use the data thanks to the dedicated user interfaces of the web applications that apply to these data. We will see many examples of such applications in the book parts II through IV.

The CLARIN infrastructure is still under construction, is highly incomplete, and is fragile in some respects. The development of the infrastructure also differs dramatically from country to country. Some countries started their national projects rather early (e.g. Germany, the Czech Republic and the Netherlands), but others only recently joined CLARIN ERIC and are still to start their national project. Budgets also differ significantly from country to country, which also determines the amount of work that can be done.

Many parts of the CLARIN infrastructure can already be used. In fact, it is already used for carrying out research, and yielded scientific articles. Concrete examples are the articles in the Lingua Special Issue on CLARIN (Odijk, 2016b), the PhD thesis (Augustinus, 2015), which crucially uses GrETEL (see chapter 22), and the PhD thesis (Hansen-Morath, 2016), which crucially uses Gabmap (Leinonen et al., 2016).

1.4 Projects in CLARIN-LC

In the Netherlands, the CLARIN-NL project ran from 2009 through 2015. It will be described in more detail in section 1.4.1. The project in Flanders is described in section 1.4.2. There are also several independent but related projects in the Netherlands, which will be described in section 1.4.3.

1.4.1 The CLARIN-NL project

The CLARIN-NL project received a funding of approximately 9 million euros for the period from 2009 through 2015 from the Netherlands Organisation for Scientific Research (NWO) roadmap for large-scale research infrastructures.

The CLARIN-NL project had a mixed set-up. On the one hand, a top-down approach was taken to implement essential functionality for the Netherlands part of the CLARIN infrastructure, for setting up the network of CLARIN centres, and for contributions to the central CLARIN infrastructure. Projects for these activities were defined and were assigned to relevant experts in the field selected by the CLARIN-NL Board. Originally two big projects were defined for this purpose, but in later stages multiple additional (usually relatively small) projects turned out to be necessary due to new developments, lacking functionality, increased use of certain services which required coordination, support for newly developed software, etc. More than 21% of the total budget was spent on these top-down activities.

On the other hand, a more bottom-up approach was taken for populating the infrastructure with data and software services. Here a small consortium consisting of one or more Humanities researchers and a CLARIN centre could make a proposal for the curation of existing data (i.e., for making them CLARIN-compatible) and/or for creating or updating a software application for browsing, searching, enriching, annotating, analysing and/or visualising data. The submitted proposals were evaluated by independent national and international experts, and the best-scoring project proposals were awarded funding. Four calls for such projects were launched. This approach has been very successful in that it offered much opportunity to react to emerging problems, bring in more partners, increase the coverage of Humanities disciplines in CLARIN, and to react to ideas and proposals coming from our prospective users. Almost 46.9 % of the total budget was spent on these activities.[7]

Though CLARIN was initiated by the linguistics and language technology communities, it was always the intention to make it an infrastructure for the Humanities more generally, and even to include the Social Sciences: it is intended for all Humanities researchers that work with language. The CLARIN-NL project was quite successful in involving other disciplines from the Humanities.[8] There were projects on history, linguistics, literary studies, religion studies, media studies, archaeology, political studies, and philosophy, covering quite a broad spectrum of the Humanities.

Of these, linguistics was most dominant, and covered linguistic subfields such as dialect studies, discourse studies, historical linguistics, first-language acquisition, language attrition, language documentation, language typology, lexicography and lexicology, morphology, phonetics, second-language acquisition, semantics, sign language and syntax. History was also prominently present with projects on subfields such as the history of the Second World War, mediaeval studies, naval history, oral history and parliamentary history. In the domain of literary studies there were projects on Arthurian novels, emblem studies, literary reception, mediaeval studies, and songs. For many more details, we refer to Odijk (2016a).

We summarise here the major achievements of the CLARIN-NL project, and indicate where more details on these achievements can be found in this book:

- CLARIN-NL created the Netherlands part of the CLARIN infrastructure with five centres, four of which are certified CLARIN centres (chapter 4);
- CLARIN-NL has incorporated a wide range of data and dedicated software applications into the CLARIN infrastructure, enabling their use by a much larger community than before CLARIN-NL (parts II, III and IV);
- CLARIN-NL has raised wide awareness of the existence and importance of the CLARIN infrastructure within the Humanities researcher community in the Netherlands;
- The CLARIN infrastructure and the data and software applications contained in it are actually used in research, and its use is increasing (see section 1.3 for some examples);
- CLARIN-NL has a clear focus on language but covers a large spectrum within the Humanities (see above, and part IV);
- Big steps have been taken in improving interoperability of data and software, both on the syntactic and the semantic level (chapters 5, 6 and 7);
- Through CLARIN-NL, the Netherlands have played a leading role in CLARIN at the European level and promoted international cooperation (this chapter).

Of course, there is still room for significant improvement. We list the major issues:

- There is as yet no business model that makes the CLARIN infrastructure sustainable, i.e. so that it can continue to exist without occasional funding through the National ESFRI Roadmap funds;
- Interoperability of software and data still requires a lot of improvements, not only in the Netherlands but also in the whole CLARIN infrastructure;

[7] This also includes the CLARIN expertise centres on data curation and historical resources.

[8] But it intentionally did not focus on the Social Sciences except for some closely related disciplines, e.g. political sciences, which are, in terms of infrastructural needs, very close to the study of history.

- Visibility of the resources (e.g. via the CLARIN Virtual Language Observatory) must be significantly improved;
- The creation of common CMDI metadata must be made much simpler;
- More sophisticated options for searching through distributed content must be created.

Fortunately, significant amounts of funding were obtained for successor projects (CLARIAH-SEED, CLARIAH-CORE), in which these (and other) issues can be addressed.

1.4.2 CLARIN in Flanders

It was difficult to obtain funding for work on CLARIN in Flanders. Fortunately, by focusing on cooperation between the Netherlands and Flanders, in particular with regard to the shared language (Dutch), funds were obtained for some work in Flanders. The activities in Flanders consisted of two parts:

- Close collaboration with the Netherlands on turning natural language processing tools that were developed earlier (inter alia in the joint Netherlands-Flanders STEVIN project; Spyns and Odijk, 2013) into web services and integrating them in a workflow system. The results of this TTNWW project are described in chapter 7.
- A number of small projects carried out fully in Flanders, in particular on syntactic search (see chapter 22), stylistics (see chapter 16), and tools for extraction of pregnancy and ideological context from speeches.

These projects were carried out successfully, and some results were extended in independently financed projects (e.g., the GrETEL application resulting from the project on syntactic search).

1.4.3 Related Projects

There were other projects in the Netherlands that were independent of CLARIN-NL but played a role in CLARIN-NL: the CKKC-project, Nederlab, and Taalportaal.

The CKCC project (Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic) was an independently financed project. It aimed to build an application to browse and analyse around 20,000 letters that were written by and sent to 17th century scholars who lived in the Dutch Republic, as well as to enable visualisations of geographical, time-based, social network and co-citation inquiries. It was selected in the CLARIN-EU call for Humanities and Social Sciences projects as the project proposal that '[would] best demonstrate the use of LRT and would show the potential of a research infrastructure in the Humanities' (CLARIN Newsletter 6, p. 3).[9] It has received funding from CLARIN-NL to apply language technology in the project (esp. part of speech tagging) and to make their results CLARIN-compatible. The results of the project are described in Part IV, chapter 26.

The Nederlab project is an independently funded project that aims to provide data and tools for the longitudinal study of the Dutch language and culture. It has also been supported financially by CLARIN-NL and has been set up as the second (virtual) CLARIN-NL centre of expertise, more specifically on data and tools for the study of the Dutch language and culture across time. Within the Nederlab project large amounts of historical data are curated, and their metadata created or curated. A dedicated search application has been constructed to search in the data, their metadata and in annotations on these data on multiple tiers. A first version of the Nederlab application was demonstrated at the CLARIN-NL final event (March 2015), and extended versions that incorporate

[9] https://www.clarin.eu/sites/default/files/cnl06_web.pdf.

parts of the multitier annotation search facilities at various workshops in 2015. Development of the full dataset and search application is ongoing, and the project will, when it is finished, bring huge amounts of historical Dutch text corpora and a dedicated search application into the CLARIN infrastructure.

The Taalportaal project is an independently financed project that aims to create a comprehensive and authoritative scientific digital grammar, the *Taalportaal* (Language Portal), which is an interactive knowledge base about the three languages Dutch, Frisian, and Afrikaans. It covers syntax, morphology and phonology. From the Taalportaal links are made to language resources such as annotated text corpora and lexical databases. CLARIN-NL has supported a project to create such links with ready-made queries to illustrate the description of specific constructions with actual examples from richly annotated corpora. The results of this project are described in chapter 24.

1.5 Structure of this book

This book starts with two introductory chapters (of which the current chapter is one). They are followed by multiple chapters grouped into 4 parts, each dealing with a specific topic and addressing a specific user group. Each part starts with an introductory chapter that sketches the background, and relates the individual chapters to each other and to the CLARIN infrastructure as a whole.

Chapter 2 provides a more detailed overview of the CLARIN infrastructure and explains how it can benefit a researcher.

In Part I, the technical infrastructure is described: the technical facilities that are needed to implement the functionality described in chapter 2, as well as their organisation as a network of CLARIN centres.

The remaining parts deal with specific data and software that CLARIN has been populated with in CLARIN-LC.

Part II deals with linguistics: data and applications that may benefit research into linguistics. Given the linguistic roots of the CLARIN infrastructure, data and applications for linguistics are of course prominently represented. In fact, one subdiscipline of linguistics, syntax, was so well represented that a special book part is dedicated to it: Part III deals with data and applications for syntactic research.

Part IV covers data and application from other disciplines than linguistics.

The book structure is summarised here:

Introduction (this chapter)
The CLARIN Infrastructure in the Low Countries (chapter 2)
Part I The Technical Infrastructure (chapters 3 through 8)
Part II Infrastructure for Linguistics (chapters 9 through 16)
Part III Infrastructure for Syntax (chapters 17 through 24)
Part IV Infrastructure for Other Humanities Disciplines (chapters 25 through 32)

Acknowledgements

This work was financed by CLARIN-NL and CLARIAH.

References

Augustinus, Liesbeth (2015), *Complement Raising and Cluster Formation in Dutch: A Treebank-supported Investigation*, Phd thesis, KU Leuven, Leuven.

Hansen-Morath, Sandra (2016), *Regionale und soziolinguistische Variation im alemannischen Dreiländereck – Quantitative Studien zum Dialektwandel*, Phd thesis, Albert-Ludwigs-Universität, Freiburg.

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 (2016a), CLARIN-NL final report, *CLARIN-NL report*, Utrecht University, Utrecht, The Netherlands. `http://www.clarin.nl/sites/default/files/CLARIN%20NL%20Final%20Report%202016-06-08%20FINAL.pdf`.

Odijk, Jan (2016b), Linguistic research using CLARIN, *Lingua* **178**, pp. 1–4. Linguistic Research in the CLARIN Infrastructure, `http://dspace.library.uu.nl/handle/1874/339377`. `http://www.sciencedirect.com/science/article/pii/S0024384116300237`.

Spyns, Peter and Jan Odijk (2013), Essential speech and language technology for Dutch. Results by the STEVIN-programme. (on-line ISBN:) 978-3-642-30910-6. `http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1`.