Available online at www.sciencedirect.com

ScienceDirect

Lingua 178 (2016) 1–4

Lingua

www.elsevier.com/locate/lingua

Introduction

Linguistic research using CLARIN

Keywords: CLARIN; Research infrastructure; Linguistically annotated corpora

This Lingua Special Issue aims to show how linguistic research can benefit from the CLARIN research infrastructure.

CLARIN[1] is a research infrastructure for carrying out humanities research that deals with language resources. Its design and development started in 2008, is coordinated by CLARIN ERIC[2] since 2012, and is carried out in part by CLARIN ERIC but mostly by research organisations and data centres in the CLARIN ERIC member countries in national projects set up to make contributions to the CLARIN infrastructure. The number of CLARIN ERIC members is currently 17,[3] and growing. The CLARIN infrastructure is *distributed*, i.e., implemented in a network of CLARIN Centres,[4] and *virtual*, i.e., it provides services via the Internet.

CLARIN offers a number of basic services to researchers, in particular

1. Services to find digital data and get access to digital data
2. Services to find software for processing digital data
3. Services to apply the software to the digital data in a user friendly manner
4. Services to create and describe new data and software
5. Services to store new data and software in CLARIN for long term preservation and for making them accessible to other researchers

I will briefly describe some aspects of each of these services below.

The CLARIN research infrastructure enables researchers to conduct better linguistic research, in various respects: First, it enables researchers to give their research a much more solid empirical basis for all linguistic research than was possible until now. Second, the availability of data and services in the CLARIN-infrastructure, and their interoperability, makes it possible to address existing research questions in an easier and/or more efficient way. Third, the CLARIN infrastructure will enable addressing research questions that could not be dealt with until recently, e.g., because the relevant data, amount of data, or combination of data was not easily accessible and searchable before, Finally, it will generate new research questions that could not have been posed before CLARIN came into existence.

The articles in this Lingua special issue demonstrate these features of the CLARIN infrastructure clearly, thus showing how linguistic research can benefit from it.

Though CLARIN aims to provide services for all humanities researchers who work with language data, which includes not only linguists but also literary scholars, historians, philosophers, etc., the focus here will be on what CLARIN has to offer to linguists.

[1] CLARIN is an acronym for Common Language Resources and Technology Infrastructure. See http://www.clarin.eu.

[2] An ERIC (European Research Infrastructure Consortium) is a legal entity at the European level with countries or intergovernmental organisations as members. See http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric.

[3] Austria, Bulgaria, Czech Republic, Denmark, Dutch Language Union, Estonia, Finland, Germany, Greece, Italy, Lithuania, the Netherlands, Norway, Poland, Portugal, Slovenia, and Sweden, with the UK as observer.

[4] See https://clarin.eu/clarin-eric-datatables/centres for an overview of the CLARIN centres.

http://dx.doi.org/10.1016/j.lingua.2016.04.003

CLARIN is not the only research infrastructure for humanities researchers. In Europe, DARIAH[5] is another example, and in the U.S. there was the Bamboo project.[6] However, these do not have a specific focus on language and are therefore less relevant to linguists. Within linguistics, TalkBank[7] is an important predecessor, which is still highly in use, for a variety of subdisciplines (e.g., the CHILDES corpora and tools for children's language, see MacWhinney, 2000). CLARIN, however, has a broader scope, and is a federation of repositories and service providers, of which TalkBank actually is one.[8] CLARIN's federative set-up offers better opportunities for stability, and avoids problems with storing data at a single organisation and problems related to storing data and services at a foreign organisation.

We now turn to a brief description of the basic services that CLARIN offers.

Finding data Digital data that are relevant for linguistic research include plain text corpora and text corpora with rich linguistic annotations, lexicons and lexical databases, audio recordings (possibly with annotations), audio-visual data for language documentation, micro-comparative databases, typological databases and many others. The service of finding such digital data is offered by CLARIN mainly via the *Virtual Language Observatory* (VLO),[9] which offers a faceted search interface to descriptions ('metadata') of digital data. It contains not only descriptions of data produced in the CLARIN member countries but also descriptions for data from different sources. It is constantly fed with new metadata by systematically harvesting the metadata made available by the CLARIN centres and independent sources, e.g., the OLAC[10] community, which includes, inter alia, the Linguistic Data Consortium.[11] The VLO currently enables search in over 700k metadata records. The descriptions of the data also contain links to the actual data, so that through the VLO also access to the data itself is provided.

Finding software The VLO also provides access to descriptions of software (applications and services; operating locally or via the web) but it has been mainly designed for finding data. For finding software the Language *Resource Inventory*[12] is better suited, as well as the faceted search for services provided by the CLARIN project in the Netherlands (CLARIN-NL) in the CLARIN-NL Portal.[13] The software includes applications for browsing, searching and analysing data, for enriching data with all kinds of linguistic annotations, and for analyzing and visualising search results, among others.

Applying software Services to apply the software to the digital data in a user-friendly manner is a core functionality offered by CLARIN. The contributions in this issue will provide several concrete examples and indicate how they were used to carry out linguistic research. Most of the data and many of the software tools offered via CLARIN already existed before CLARIN. However, in most cases they require downloading and installing software packages, often only on a specific computer platform and with dependencies on other software packages. They require downloading huge amounts of data, and often require ad-hoc adaptations because the software and the data do not fit together. CLARIN makes the data available via the Internet (no downloading needed, though that is still possible), through web applications that provide interfaces tuned to the intended users, and the data and the software have been made interoperable. In several cases, dedicated new services and interfaces were created that did not exist before. This lowers the barrier for using the data and software significantly, so that it can benefit linguistic research on a much larger scale than ever before.

Several web applications developed in CLARIN offer multiple interfaces to the same data: each interface has its own target in terms of the complexity of the query, and/or the expected technical proficiency of the user. For example, GrETEL[14] offers interfaces for example-based querying (in two varieties) and a full XPATH interface. The query interface of PML-TQ[15] is supported by a graphical visualisation of the query, as well as by options for composing a query by clicking on nodes of an example parse from the treebank. Several other applications have similar features.[16] Dedicated interfaces usually restrict the usage, but by combining these different interfaces in one application an environment is created in which this can be avoided. It is in general difficult to create a correct query from scratch, but making small adaptations to an existing well-formed query is much easier. These interfaces enable users to create an initial query by means of an example (e.g., in GrETEL) or supported by a graphical interface (e.g., in PML-TQ), after which it can be refined by making (often minor) modifications in the query generated. There is an additional positive educational side effect, since the technical proficiency of a user will gradually and naturally increase in such an environment.

---

[5] DARIAH is an acronym for Digital Research Infrastructure for the Arts and the Humanities. See http://www.dariah.eu.
[6] http://www.projectbamboo.org/.
[7] http://talkbank.org/.
[8] CMU-TalkBank is indeed a certified CLARIN Type B centre.
[9] http://www.clarin.eu/vlo.
[10] Open Language Archives Community, http://www.language-archives.org/.
[11] https://www.ldc.upenn.edu/.
[12] http://lindat.mff.cuni.cz/repository/xmlui/discover.
[13] https://portal.clarin.nl/.
[14] https://portal.clarin.nl/node/1967 (used by Van Eynde *et al.* in this issue).
[15] http://lindat.mff.cuni.cz/services/pmltq/.
[16] For example PaQu (https://portal.clarin.nl/node/4182), which offers a word relations interface and facilities for full XPATH queries, and OpenSONAR (https://portal.clarin.nl/node/4195), which offers 4 interfaces, differing in complexity and suitability for specific queries.

Creating resources and resource descriptions Research projects can yield new data and software, or enrich existing data with new annotation layers. Such new data and software should be made available to the research community, for a variety of reasons. These include transparency, enabling verification and possibly even replication of research results, integrity, and efficiency. Most research results are made with public money, and should therefore be made available to the whole research community. They should be easily accessible also after a research project ended. In this way other researchers can benefit from them in their own research, and the original researchers can benefit from them in a later stage. CLARIN offers services for including newly created data and tools in the CLARIN infrastructure. In order to make this work, CLARIN imposes certain requirements on the description (metadata) of the resources, the formats of the resources (to achieve formal or syntactic interoperability), and provisions for specifying the meaning of elements in the resource (to ensure semantic interoperability). CLARIN provides a range of tools and services to assist researchers in meeting these requirements (see e.g., Odijk, 2014a, section 6.3). CLARIN also requires that the intellectual property rights and any ethical issues have been dealt with properly (and it provides assistance with this[17]), and it requires that the resources are stored on a server of one of the CLARIN centres. These centres will make sure that the data are accessible to other researchers through their metadata, which are harvested by the VLO and other tools for searching data.

Storing resources CLARIN also provides services, through the CLARIN centres, to ensure long term preservation of the data and software.[18] Each CLARIN centre must have a clear procedure in place for ensuring long term preservation, and work according to this procedure. This is one of the ingredients of the *Data Seal of Approval* (DSA),[19] which each CLARIN Centre must be awarded as one of the conditions of becoming a certified CLARIN Centre.[20]

This concludes the brief discussion of the basic services offer by CLARIN. For more details I refer to Odijk (2014a,b) and references there.

Contributions in this issue The current issue contains 3 papers dealing with *syntax* (Van Eynde *et al.*, Hinrichs, Barbiers *et al.*), 5 on *language variation* (Cornips *et al.*, Leinonen *et al.*, Barbiers *et al.*, Hinrichs, Van Sluijs *et al.*), one of which focuses on *historical* language variation (Hinrichs). Other topics covered by the papers include the *study of Creole languages* (Van Sluijs *et al.*), the *lexicon* (Cornips *et al.*), and *phonology* (Leinonen *et al.*). Another recurrent theme, along a different dimension, is the need for combining multiple sets of data and from that the need for their interoperability (Hinrichs, Van Sluijs *et al.*, Barbiers *et al.*, and Cornips *et al.*).

In the research reported on in the paper by *Barbiers et al.*, crucial use is made of the MIMORE web application for browsing, searching, analysing and visualising data in three independently existing databases of microvariation in Dutch. MIMORE is an excellent example of a web application created in CLARIN that lowers the barriers for using the existing microvariation databases. Its facilities (e.g., set operations on search results) to find, analyse and visualise correlations between data enabled the researchers to detect a correlation between pronominal doubling in DP (demonstrative doubling) and CP (subject doubling). They go on to provide an analysis of this correlation, which accounts not only for this correlation but also for the differences with regard to doubling that are found in the dialects that have it.

The paper by *Cornips et al.* shows that the COAVA application developed in CLARIN enables researchers to access two datasets from two different subdisciplines simultaneously, viz. Dutch first child language acquisition data in CHILDES (MacWhinney, 2000) and historical Dutch Dialect Dictionaries. The authors show that this makes it possible to examine the common assumption in historical linguistics that language change from the past is due to the process of non-target transmission of linguistic features, forms and structures between generations, thus between parents or adults and children. The authors illustrate it with one small case study.

The paper by *Van Sluijs et al*. investigates the variation and change in two Caribbean Creole language clusters in terms of *genealogical blends*, i.e., languages in which typological properties from different lineages from different parts of the world (here: Europe and West-Africa) are combined. In their study, they crucially use multiple sets of digital data in CLARIN on 18th and 20th century texts from the Suriname Creoles and Virgin Island Creole Dutch.

The paper by *Van Eynde et al.* investigates number agreement in copular constructions. It crucially uses syntactically annotated corpora (treebanks) and demonstrates how treebanks can be exploited in order to guide the formulation of relevant generalizations. The author crucially relies on tools and data that have recently been developed in the framework of the Dutch-Flemish STEVIN programme (2004–2011, see Spyns and Odijk, 2013) and for which CLARIN has developed applications to lower the barriers to their use, in particular the GrETEL[21] web application (Augustinus et al., 2012).

The paper by *Hinrichs* investigates the historical development of auxiliary fronting in German subordinate clauses. It makes crucial use of two contemporary syntactically annotated text corpora, and a corpus collection for diachronic data covering the period from 1610 to 1900. These corpora are all part of CLARIN. The study demonstrates not only the added

[17] See https://www.clarin.eu/content/licenses-agreements-legal-terms for details.
[18] See http://www.clarin.eu/content/depositing-services for details.
[19] http://www.datasealofapproval.org/en/community/.
[20] https://www.clarin.eu/clarin-eric-datatables/centres/1.
[21] http://nederbooms.ccl.kuleuven.be/eng/gretel.

value that annotated corpora can provide for in-depth studies in historical syntax, but also, and especially, the added value of interoperable language resources for linguistic investigations that require access to and analysis of multiple linguistic resources. Here again, a web application developed in CLARIN (Tündra[22]) plays a crucial role in facilitating search and analysis of the relevant data.

The paper by *Leinonen et al.* describes *Gabmap*, a web application that analyses 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. It is shown how *Gabmap* can be used to create distributional and clustering maps, for measuring linguistic distances, for characterising dialect continua using multidimensional scaling, for identifying dialect groups by clustering techniques, and for finding typical features or ''shibboleths''. *Gabmap* is another clear example where CLARIN has lowered the barriers for using an existing tool (L04[23]), in this case by turning (parts of the) tool into a web application with a clear interface targeted at the intended users, viz. linguists.

The papers in this issue clearly demonstrate the added value of the facilities that CLARIN offers for linguistics in general, including theoretical linguistics, which Lingua focuses on. This is shown by the fact that data are easily found and accessed through CLARIN (all contributions), by the user-friendly options to search in large linguistically annotated corpora made available through CLARIN (as used by Barbiers *et al.*, Van Eynde *et al.*, Cornips *et al.*, Hinrichs), by the options that such tools offer for analysis of the search results and finding correlations between data (Barbiers *et al.*, Leinonen *et al.*), and by the systematic efforts towards interoperability so that multiple linguistic resources can be searched together (Barbiers *et al.*, Cornips *et al.*, Hinrichs). But the papers in this special issue show only a tiny sample of linguistic research that may benefit from the services CLARIN has to offer. CLARIN has already a lot to offer, even though it is still under construction. The development pace differs from country to country (and depends on financing by national governments). For this reason the coverage of various languages is still somewhat unbalanced. Nevertheless, I would like to invite all linguistic researchers to explore how CLARIN can serve them. For questions about this contact a relevant CLARIN centre in your neighbourhood,[24] or contact one of the national CLARIN helpdesks.[25]

References

Augustinus, L., Vandeghinste, V., Van Eynde, F., 2012. Example-based treebank querying. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), pp. 3161–3167. http://www.lrec-conf.org/proceedings/lrec2012/pdf/756_Paper.pdf

MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk, 3rd ed., vol. 818. Lawrence Erlbaum Associates, Mahwah, NJ.

Odijk, J., 2014a. The CLARIN Infrastructure in the Netherlands: What is it and How Can You Use It? Working Paper. Utrecht University. http://dspace.library.uu.nl/handle/1874/307045

Odijk, J., 2014b. The CLARIN Infrastructure in the Netherlands: Design and Construction. Working Paper. Utrecht University. http://dspace.library.uu.nl/handle/1874/303787

Spyns, P., Odijk, J., 2013. Essential Speech and Language Technology for Dutch. Results by the STEVIN-Programme. Springer, Berlin/HeidelbergISBN: 978-3-642-30910-6. http://dx.doi.org/10.1007/978-3-642-30910-6 413 p.

Jan Odijk*
Utrecht University, Netherlands

**Tel.: +31 302536006*
E-mail address: j.odijk@uu.nl

Available online 29 April 2016

[22] https://weblicht.sfs.uni-tuebingen.de/Tundra/.
[23] http://www.let.rug.nl/~kleiweg/L04/.
[24] See https://clarin.eu/clarin-eric-datatables/centres for an overview of the CLARIN centres.
[25] See http://www.clarin.eu/content/support for an overview.