The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars

Erhard Hinrichs[1], Steven Krauwer[2]
[1]University of Tübingen, [2]CLARIN ERIC
[1]Seminar für Sprachwissenschaft, Computerlinguistik, Wilhelmstr. 19, 72074 Tübingen, Germany;
[2]Trans 10, 3512 JK Utrecht, The Netherlands
E-mail: erhard.hinrichs@uni-tuebingen.de, s.krauwer@uu.nl

Abstract

CLARIN is the short name for the *Common Language Resources and Technology Infrastructure*, which aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. CLARIN is in the process of building a networked federation of European data repositories, service centers and centers of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centers will be interoperable so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. Interoperability of language resources and tools in the federation of CLARIN Centers is ensured by adherence to TEI and ISO standards for text encoding, by the use of persistent identifiers, and by the observance of common protocols. The purpose of the present paper is to give an overview of language resources, tools, and services that CLARIN presently offers.

Keywords: digital humanities, research infrastructures, interoperability of language resources and tools

1. Introduction

The digital age has opened up entirely new ways for researchers in all fields of science to gain easy access to large amounts of electronically available data, tools and services to analyse and visualize the contents of such datasets, and even more importantly to collaborate on joint research projects, relying on virtual research environments and geographically distributed research infrastructures as a backbone for such collaborations. The term *e-Science* has been coined to refer to this new paradigm for conducting research. This recent trend has also extended to research in the humanities and social sciences, where this innovative approach is subsumed by the terms *eHumanities* or *digital humanities*.

2. Goals of CLARIN

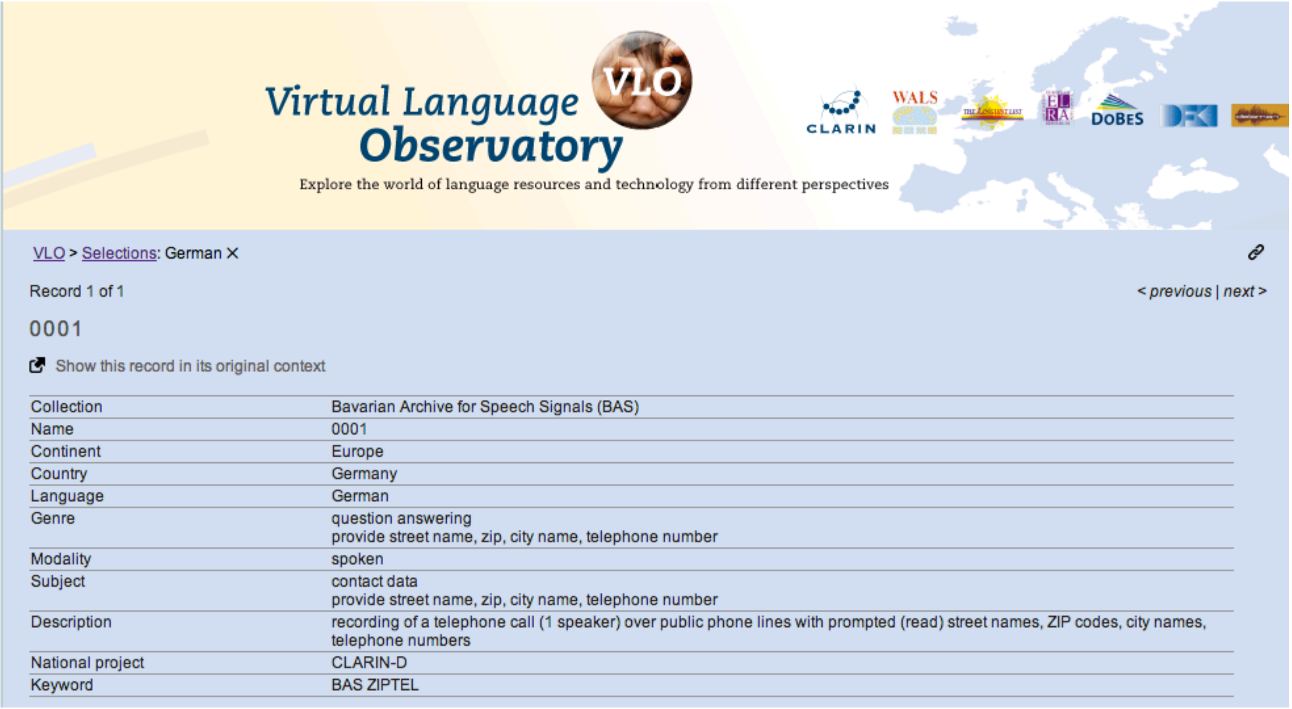
CLARIN is the short name for the *Common Language Resources and Technology Infrastructure*, which aims at providing easy and sustainable access for scholars in the humanities and social sciences (HSS) to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. CLARIN is one of the research infrastructures that were selected for the European Research Infrastructures Roadmap by ESFRI, the European Strategy Forum on Research Infrastructures. The ESFRI Roadmap contains five research infrastructures in the area of social sciences (CESSDA, European Social Survey, and SHARE) and humanities (CLARIN and DARIAH).

The CLARIN governance and coordination body at the European level is CLARIN ERIC. An ERIC is a new type of international legal entity, established by the European Commission in 2009. Its members are governments or intergovernmental organisations. CLARIN ERIC has nine founding members: Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, The Netherlands, Poland, and the Dutch Language Union, an intergovernmental organization representing the Flemish speaking part of Belgium and The Netherlands. Norway has the status of an observer country and is expected to join as a full member shortly, as are several other European countries.

CLARIN is in the process of building a networked federation of European data repositories, service centers and centers of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centers will be interoperable so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. At this moment the CLARIN infrastructure is still under construction, but a number of participating centers are already offering access services to data, tools and expertise. The purpose of the present paper is to give an overview of language resources, tools, and services that CLARIN presently offers.

3. Data Discovery, Acquisition, Data Curation

Every eHumanities project, regardless of whether it is a quantitative or a qualitative study, is grounded in a dataset. In a quantitative study, such a dataset is typically subjected to statistical analysis of some sort, in order to (dis-)confirm a particular hypothesis. In a qualitative study, the initial hypothesis may not be completely clear yet and is often developed in the process of exploring a dataset.


Virtual Language
VLO
Observatory
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DOBES

VLO > Selections: German ×

Record 1 of 1

< previous | next >

0001

Show this record in its original context

Collection	Bavarian Archive for Speech Signals (BAS)
Name	0001
Continent	Europe
Country	Germany
Language	German
Genre	question answering provide street name, zip, city name, telephone number
Modality	spoken
Subject	contact data provide street name, zip, city name, telephone number
Description	recording of a telephone call (1 speaker) over public phone lines with prompted (read) street names, ZIP codes, city names, telephone numbers
National project	CLARIN-D
Keyword	BAS ZIPTEL

Figure 1: Virtual Language Observatory (VLO)

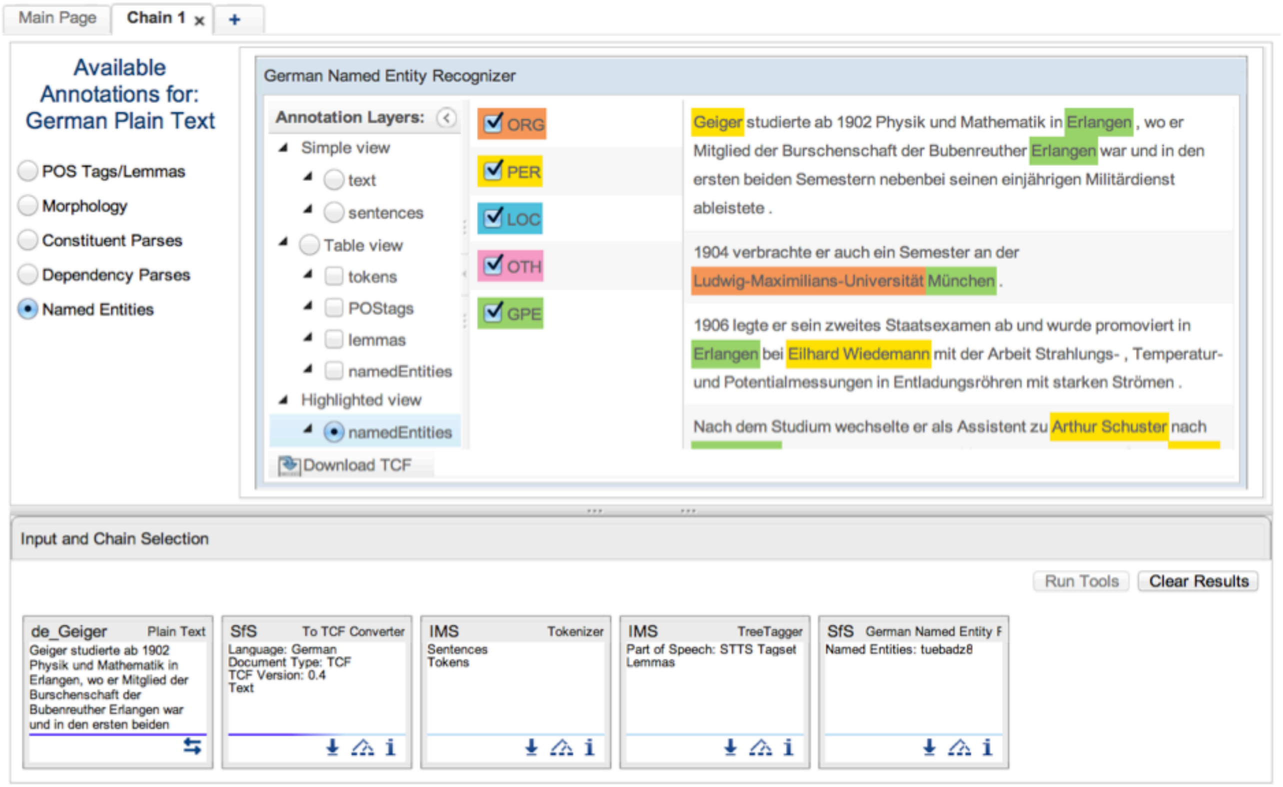
3.1 Reference Datasets

Often investigators know at the outset of their project, which datasets need to be consulted. Such datasets are typically reference sets that are well-known in the scientific community and that are often consulted to ensure that the results obtained in a particular study can be compared to previous studies performed on the same dataset. The federation of CLARIN centers offers a high number of such reference datasets for the languages represented by the CLARIN member countries. The CLARIN Center at the Austrian Academy of Sciences in Vienna offers the Austrian Academy Corpus (AAC; Biber & Breiteneder 2004), a very large collection of German texts and German literature covering the period of 1848 to 1989. The German reference corpus DeReKo (Kupietz et al. 2010), the largest linguistically motivated collection of contemporary German texts with more than 4.0 billion word tokens (as of August 2010), is hosted by the CLARIN Center at the IDS in Mannheim. The CLARIN Center at the Berlin-Brandenburg Academy (BBAW) provides access to the German Text Archive (DTA; Geyken et a. 2010), a digital collection of German-language printed works from around 1650 to 1900 as full text and as digital facsimile. The CLARIN Center at the Polish Academy of Sciences in Warsaw hosts the National Corpus of Polish (Przepiórkowski et al. 2011), a reference corpus of the Polish language with more than fifteen hundred millions of words.

CLARIN centers offer extensive collections of spoken language. The CLARIN Center at the Meertens Institute in Amsterdam is home to thousands of hours of audio material for Dutch, including more than 1000 hours of dialect recordings (Barbiers et al. 2006). The CLARIN Center at the Bavarian Speech Archive in Munich specializes in digital corpora for contemporary German (Schiel & Draxler & Tilman 1997). The CLARIN Center in Sofia offers the Bulgarian Political and Journalistic Speech Corpus (Osenova & Simov 2012).

The language resources offered by CLARIN are not restricted to the languages spoken in CLARIN member countries. The TLA Language Archive at the MPI for Psycholinguistics in Nijmegen offers easy access to the DOBES Archive (Drude & Trisbeek & Broeder 2012), which documents endangered languages around the world, and to the CHILDES (CHIld Language Data Exchange System) database of child language acquisition data (MacWhinney 2000).

Another key language resource are high-quality lexica. The CLARIN Center at the University of Tartu provides on-line access to a variety of monolingual and bilingual lexica for Estonian (Meister &Vilo 2008). The CLARIN Center at BBAW in Berlin is home to the Digitale Wörterbuch der deutschen Sprache (DWDS; Klein & Geyken 2010). The DWDS lexicon uses extensive digital corpus collections to document the actual usage of German words and offers on-line access to all materials. Apart from traditional lexica, CLARIN also offers access to lexical resources that model word meanings in terms of a network of lexical and conceptual relations. The CLARIN center federation currently hosts such word nets for Czech (Pala & Smrž 2004), Danish (Pedersen et al. 2006), Dutch (Vossen et al. 2013), Estonian (Orav & Vider 2000), Finnish (Lindén & Carlson 2010), German



Geiger studierte ab 1902 Physik und Mathematik in Erlangen , wo er Mitglied der Burschenschaft der Bubenreuther Erlangen war und in den ersten beiden Semestern nebenbei seinen einjährigen Militärdienst ableistete .

1904 verbrachte er auch ein Semester an der Ludwig-Maximilians-Universität München .

1906 legte er sein zweites Staatsexamen ab und wurde promoviert in Erlangen bei Eilhard Wiedemann mit der Arbeit Strahlungs- , Temperatur- und Potentialmessungen in Entladungsröhren mit starken Strömen .

Nach dem Studium wechselte er als Assistent zu Arthur Schuster nach

de_Geiger — Plain Text: Geiger studierte ab 1902 Physik und Mathematik in Erlangen, wo er Mitglied der Burschenschaft der Bubenreuther Erlangen war und in den ersten beiden

SfS — To TCF Converter: Language: German; Document Type: TCF; TCF Version: 0.4; Text

IMS — Tokenizer: Sentences; Tokens

IMS — TreeTagger: Part of Speech: STTS Tagset; Lemmas

SfS — German Named Entity F: Named Entities: tuebadz8

Figure 2: WebLicht Graphical User Interface

(Hamp & Feldweg 1997; Henrich & Hinrichs 2010), and Norwegian (Fjeld & Nygaard, L. 2010).

In addition to reference datasets, CLARIN provides access to an extensive set of metadata records. The Virtual Language Observatory (VLO; van Uytvanck et al. 2010) currently contains more than 500.000 metadata records to language resources and tools. Facetted search and a visual map provide easy-to-use interfaces for HSS scholars to locate language resources and tools that match the needs in a particular research project. The screen shot in Figure 1 shows the VLO web application and displays the metadata for a resource hosted by the CLARIN Center in Munich as the result of a facetted VLO search for speech recordings for the German language.

3.2 Creation of New Resources

If an eHumanities study requires the construction of new digital datasets, this is usually due to the fact that no existing resource can provide the answer to a particular research question. In fact, it may be the main goal of the project to fill a gap in the existing set of resources for a particular research question or for a particular language. Resource creation is a costly effort, especially if significant manual effort is required. Special care must therefore be taken that such data creation efforts (i) adhere to best practises or standards for text encoding whenever possible and (ii) are guided by a data management plan that covers the entirely data lifecycle from data collection, to data sharing and data archiving. HSS scholars often lack the necessary experience or access to data repositories to meet these expectations. The CLARIN-D User Guide (Herold & Lemnitzer 2012) provides practical information on the use of standards for language resources and on following good practises in data creation.

4. Data Mining and Data Analysis

4.1 Query Tools and Federated Content Search

Since datasets available in electronic form are typically very large, CLARIN centers support HSS scholars by providing powerful and easy-to-use query tools for many of the resources described above. Access is greatly facilitated if such query tools are realized as web applications and thus available in any web browser. Two good examples of this kind are the web application for querying the German Text Archive (Jurish & Thomas & Wiegand 2014).

In addition to query interfaces for individual resources, CLARIN offers a Federated Content Search (FCS) functionality that enables HSS scholars to construct a virtual corpus collection hosted by different CLARIN centers and to query this virtual corpus via a common search interface. Currently, nine CLARIN centers in Germany and in the Netherlands make more than 20 resources available to the linguistic researches via the common interface of the CLARIN-D Federated Content Search, and this number is steadily growing. The CLARIN Center at the University of Oslo also provides

FCS functionality via the GLOSSA corpus query tool.

4.2 Work Flows for Data Annotation

Language data that are annotated with linguistic information can be searched with high accuracy for specific data patterns. The CLARIN Centers in Oslo, Prague, Tübingen and at the Dutch Language Union offer linguistically annotated corpora, so-called *treebanks*, for Czech (Prague Dependency Treebank; Hajic et al. 2000), Dutch (ALPINO; van der Beek et al. 2002), German (TüBa-D/Z; Telljohann & Hinrichs & Kübler 2004), and Norwegian (INESS; Meurer et al. 2013) with accompanying query tools.

If a collection of language resources does not contain sufficient linguistic information, for example if the word forms in a corpus have not been lemmatized, it is impossible to obtain meaningful word frequency distributions. Likewise, if an HSS scholar wants to search for all person names in a very large newspaper corpus in order obtain an overview of who is currently in the news, then the person names in such a corpus needs to be marked up. CLARIN offers support for HSS scholars who need to add annotations of this kind. The web application WebLicht (Hinrichs & Hinrichs & Zastrow 2010), hosted by the CLARIN Center in Tübingen, is a tool-suite for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers can be combined into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format. WebLichts's graphical user interface in Figure 2 shows the annotation of a German text corpus with five different subtypes of named entities (location, person, organization, geo-political entity, and other) in table format along with the processing chain of five different web services that produced these layers of linguistic annotation.

Recently the WebLicht tool suite has been extended to spoken language. This can be achieved with the integration of the webservice WebMaus (Kisler & Schiel & Sloetjes 2012), which is provided by the CLARIN Center in Munich. WebMaus takes as input an audio file and its transcription and automatically aligns the speech signal with its transcriptions. The WebLicht tool can then further annotate the transcriptions so that via the automatic alignment, a user can find the relevant portions of the speech signal for particular data patterns.

5. Data Visualization

A key aspect to a successful eHumanities project involves visualization tools that render the data analysis results in an easy-to-grasp fashion. This is particularly important if the datasets involved are very large. Accordingly, data visualization has become an active research field in its own right within digital humanities research. While CLARIN cannot and does not want to claim that it can provide a comprehensive suite of eHumanities visualization tools, it can already support HSS scholars with a number of helpful applications. The CLARIN Center at the University of Copenhagen has developed a visualization tool for parallel inspection of word nets.

The Microcomparative Morphosyntactic Research (MIMORE) tool offered by the CLARIN Center at the Meertens Institute in Amsterdam enables researchers to investigate morphosyntactic variation in Dutch dialects. MIMORE accesses three related databases with a common on-line search engine. Figure 3 shows the MIMORE search results for the 2nd person singular form *bist* of the copula verb *zein* (English: *be*). This verb form is restricted particular dialects of Dutch and differs from *bent*, the corresponding verb form in Standard Dutch. Since the language data in the three databases searchable by MIMORE contain geographical references, the search results can be automatically plotted on a geographic map. Visualizations of this kind provide a natural way to identify dialectal variation, as in the case of the word *bist*: The results all cluster in the northern part of the Netherlands and along the border with Germany. This is not surprising since German uses the same verb form.

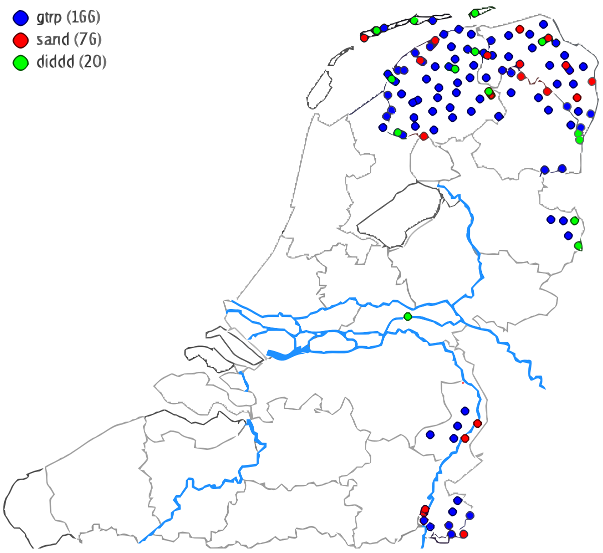


Figure 3: MIMORE

The CLARIN Center at the University of Tübingen provides access to the web application CiNaViz (short for: *City Name Visualisation*). CiNaViz is a search interface for more than a million geographical locations all over Europe and automatically displays the search results on a geographical map. Figure 4 shows the distribution of city names ending in *bach* (red), *beck* (blue) and *bek* (green) in Central Europe. Interestingly, the dividing line between *bach* and *beck* follows the famous Benrath Line that separates northern from southern German dialects. As with the MIMORE tool, CiNaViz offers a valuable tool for HSS scholars interested in language variation.

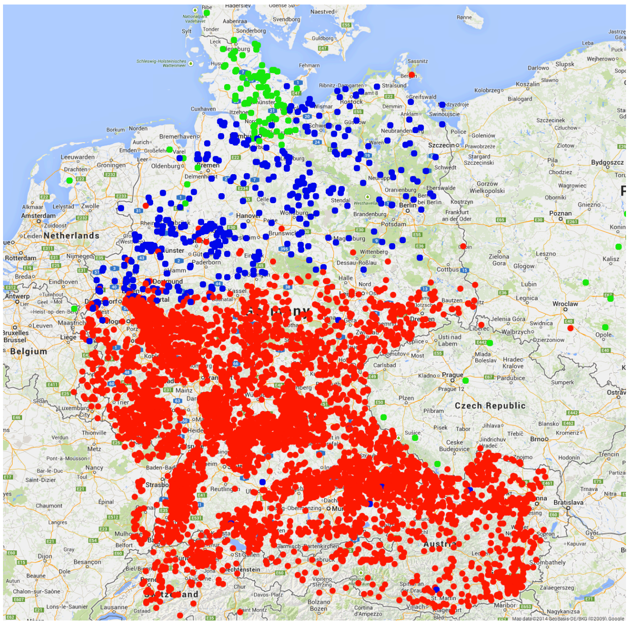


Figure 4: CiNaViz

6. Data Sharing and Data Archiving

The last stage in an eHumanities project typically involves the sharing, publishing and archiving of the datasets that were created, annotated or analysed during the project. CLARIN provides support for this phase of an eHumanities project by providing SimpleStore and OwnCloud solutions for data sharing and for collaborative work on the same dataset. Many CLARIN data repositories offer archiving services for external resources and for finished datasets. For quality assurance, all CLARIN Centers are assessed by the CLARIN Assessment Committee, according to strictly defined technical requirements and have to obtain the Data Seal of Approval (Sesink & van Horik & Harmsen 2010) for their services.

7. Technical Aspects

Interoperability of language resources and tools in the federation of CLARIN Centers is ensured by adherence to TEI and ISO standards for text encoding, by the use of persistent identifiers as long-lasting references to digital language data as well as by the observance of common protocols: Shibboleth for user authentication and authorization, SRU/CQL for Federated Contents Search, and OAI-PMH for metadata harvesting.

8. Conclusion and Outlook

Here we could describe only a subset of the resources and tools available in CLARIN. Section 10 (References) of this paper provides URLs for many of the ones mentioned in the present paper. For comprehensive and up-to-date information about all CLARIN resources and tools, we refer interested readers to the CLARIN homepage: http://www.clarin.eu/.

9. Acknowledgements

We would like to thank all CLARIN centers for their profound contributions to the sharing of resources, tools, and knowledge as well as all CLARIN ERIC members for their financial support in the preparation and construction of the CLARIN infrastructure. All CLARIN centers owe a special thank you to Peter Wittenburg of the Max-Planck Institute for Psycholinguistics in Nijmegen, whose vision was essential in the conception of CLARIN and in the blue print for its architecture. The research and development of the first author are supported by grants from the Bundesministerium für Bildung und Forschung (BMBF), from the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, and from the Eberhard Karls Universität Tübingen. The second author expresses his gratitude to Utrecht University for their long-standing support as the host institution of CLARIN ERIC. Both authors are grateful to their colleagues on the Board of Directors of CLARIN ERIC for their support and for their comments on earlier drafts of this paper.

10. References

Barbiers, S. et al. (2006). *Dynamic Syntactic Atlas of the Dutch dialects* (DynaSAND). Amsterdam, Meertens Institute. URL: http://www.meertens.knaw.nl/sand/ .

Biber, H., Breiteneder, E. (2004). The AAC [Austrian Academy Corpus] - An Enterprise to Develop Large Electronic Text Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004,* May 26--28, 2004, Lisbon, Portugal. URL: http://www.aac.ac.at/.

Drude, S., Trilsbeek, and Broeder, D. (2012). Language Documentation and Digital Humanities: The (DoBeS) Language Archive. In J. C. Meister, (Ed.), *Digital Humanities 2012 Conference Abstracts.* University of Hamburg, Germany, July 16--22, pp. 169-173. URL: http://dobes.mpi.nl/.

Fjeld, R. V., Nygaard, L. (2010). NorNet - a monolingual wordnet of modern Norwegian. In B. S., Pedersen, A. Braasch, S. Nimb & R. V. Fjeld (Eds.), *Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. NEALT Proceedings Series, Vol. 7, pp. 13--16.

Geyken, A.; Haaf, S.; Jurish, B.; Schulz, M.; Steinmann, J.; Thomas, C. and Wiegand, F. (2010). Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin & C. Puschmann (Eds), *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, 20./21. September 2010. Beiträge der Tagung. Hrsg. von. 2., ergänzte Fassung. hbz, 2011, pp. 157--161. URL: www.deutschestextarchiv.de.

Hajič, J.; Böhmová, A.; Hajičová, E.; Vidová, B. and Hladká, V. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora*, Amsterdam: Kluwer, 2000, pp. 103--127. URL: https://ufal.mff.cuni.cz/pdt2.0/.

Hamp, B., Feldweg, H. (1997): GermaNet -- a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid, 1997, pp. 9--15. URL: http://www.sfs.uni-tuebingen.de/GermaNet/.

Henrich, V., Hinrichs, E. (2010) Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010),* Beijing, China, August 2010, pp. 456--464.

Hinrichs, E., Hinrichs M., and Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010).* Uppsala, Schweden, pp. 25--29. URL: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

Jurish, B., Thomas, C., and Wiegand, F. (2014). *Querying the Deutsches Textarchiv.* In: U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Eds.), *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities* (co-located with iConference 2014, Berlin, 4. März, 2014), pp. 25--30.

Klein, W., Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In U. Heid; S. Schierholz; W. Schweickard; H. E. Wiegand; R. Gouws & W. Wolski, (Eds.), *Lexikographica*. Berlin/New York, pp. 79--93.

Kisler, T.; Schiel, F. and Sloetjes, H. (2012). Signal processing via web services: the use case WebMAUS. In J. C. Meister, (Ed.), *Digital Humanities 2012 Conference Abstracts.* University of Hamburg, Germany, July 16--22, pp. 30--34. URL: https://clarin.phonetik.uni-muenchen.de/BASWebServices/index.html

Kupietz, M.; Belica, C.; Keibel, H. and Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari et al. (Eds.), *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010).* Valletta, Malta: European Language Resources Association (ELRA), pp. 1848--1854, URL: http://www1.ids-mannheim.de/kl/projekte/korpora/.

Lindén, K, Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, Vol. 17, pp. 119--140. URL: http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Mahwah,* NJ: Lawrence Erlbaum Associates.

Meister, E., Vilo, J. (2008). Strengthening the Estonian Language Technology. In Calzolari; K. Choukri; B. Maegaard; J. Mariani; J. Odijk; S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th conference on International Language Resources and Evaluation (LREC 2008),* Marrakesh, Marocco, pp. 3101--3104. URL: http://www.keeleveeb.ee/

Meurer, P.; Dyvik, H.; Rosén, V.; De Smedt, K.; Lyse, G. I.; Losnegaard, G. S. and Thunes. M. (2013). The INESS treebanking infrastructure. In S. Oepen, K. Hagen & J. B. Johannessen (Eds.), *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013),* May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16, number 85 in Linköping Electronic Conference Proceedings. Linköping University Electronic Press, pp. 453—458.

Nygaard, L.; Priestley, J.; Nøklestad, A. and Bondi Johannessen, J. (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In N. Calzolari; K. Choukri; B. Maegaard; J. Mariani; J. Odijk; S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th conference on International Language Resources and Evaluation (LREC 2008),* Marrakesh, Marocco, pp. 617--622. URL: http://hf-tekstlab.uio.no/glossa2/front

Orav, H., Vider, K. (2000). Estonian WordNet. *Kogumikus Congressus Nonus Internationalis Fenno-Ugristarum*.. Pars V. Dissertationes sectionum: Linguistica II. Tartu, pp. 490--497.

Osenova, P., Simov, K. (2012). The Political Speech Corpus of Bulgarian. In N. Calzolari; K. Choukri; T. Declerck; M. U. Dogan; B. Maegaard; J. Mariani; J. Odijk & S. Piperidis (Eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012),* Istanbul, pp. 1744--1747.

Pala, K., Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology.* Vol. 7, pp. 79--88.

Piasecki, M., Szpakowicz, S., Broda, B. (2009). *A Wordnet from the Ground Up.* Wroclaw: Oficyna Wydawnicza Politechniki Wrocławskie.

Pedersen, B.S.; Nimb, S.; Asmussen, N., J.; Sørensen, N.; Trap-Jensen, L. and Lorentzen H. (2006). DanNet - A WordNet for Danish. In *Proceedings from Third International Conference on Global Wordnets,* Jeju, South Korea, pp. 329--331.

Przepiórkowski, A.; Bańko, M.; Górski, R. L.; Lewandowska-Tomaszczyk, B.; Łaziński, M. and Pęzik. P. (2011). National Corpus of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland. pp. 259--263. URL: http://nkjp.pl/.

Schiel, F.; Draxler Chr. and Tillmann H. G. (1997). The Bavarian Archive for Speech Signals: Resources for the Speech Community. In *Proceedings of the EUROSPEECH 1997*, Rhodos, Greece, pp. 1687--1690.

Sesink, L.; van Horik, R. and Harmsen, H. (Eds.), Data Seal of Approval - Quality guidelines for digital research data in the Netherlands. The Hague, Data Archiving and Networked Services - 2nd ed. DANS,

2010. ISBN 978 9490 531 027. URL: http://datasealofapproval.org/en/

Simov, K.; Osenova, P.; Kolkovska, S.; Balabanova, E.; Doikoff, D.; Ivanova, K.; Simov, A. and Kouylekov, M.. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proceedings of LREC 2002,* Canary Islands, Spain, pp. 1729--736.

Telljohann, H., Hinrichs, E., Kübler, S. (2004). The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004),* pp. 2229--2235.

van der Beek, L.; Bouma, G.; Malouf, R. and van Noord, J. (2002). The Alpino Dependency Treebank. In M. Theune; A. Nijholt & H. Hondorp (Eds.), *Computational Linguistics in the Netherlands 2001. Selected Papers from the Twelfth CLIN Meeting*. Amsterdam/New York: Rodopi, pp. 8--22.

van Uytvanck, D.; Zinn, C.; Broeder, D.; Wittenburg, P. and Gardelleni, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari; B. Maegaard; J. Mariani; J. Odjik; K. Choukri; S. Piperidis; M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation* (*LREC 2010),* La Valetta, Malta. European Language Resources Association (ELRA), pp. 900--903.

Vossen, P.; Maks, I.; Segers, R.; van der Vliet, H.; Moens, M-F.; Hofmann, K.; Tjong Kim Sang, E. and de Rijke, M.. (2013). Cornetto: A Combinatorial Lexical Semantic Database for Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme.* Theory and Applications of Natural Language Processing, XVII, URL: http://www2.let.vu.nl/oz/cltl/cornetto/.

Zastrow, T.; Hinrichs, E.; Hinrichs, M. and Beck, K. (2013) Scientific Visualization for the Digital Humanities as CLARIN-D Web Applications. In *Digital Humanities 2013 Conference Abstracts*, Center for Digital Research in the Humanities, University of Nebraska-Lincoln, Lincoln, Nebraska, USA, July 16-19, 2013, pp. 466--469. URL: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/ index.php/CiNaViz