Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation

Hanna Hedeland, Timm Lehmberg, Felix Rau, Sophie Salffner, Mandana Seyfeddinipur, Andreas Witt
HZSK/Universität Hamburg, INEL/Universität Hamburg, IfL/Universität zu Köln,
ELAR&SWLI/SOAS University of London, IfDH/Universität zu Köln
{hanna.hedeland, timm.lehmberg}@uni-hamburg.de
{f.rau, andreas.witt}@uni-koeln.de
elararchive@soas.ac.uk

Abstract

The European digital research infrastructure CLARIN (Common Language Resources and Technology Infrastructure) is building a Knowledge Sharing Infrastructure (KSI) to ensure that existing knowledge and expertise is easily available both for the CLARIN community and for the humanities research communities for which CLARIN is being developed. Within the Knowledge Sharing Infrastructure, so called Knowledge Centres comprise one or more physical institutions with particular expertise in certain areas and are committed to providing their expertise in the form of reliable knowledge-sharing services. In this paper, we present the ninth K Centre – the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD) – and the expertise and services provided by the member institutions at the Universities of London (ELAR/SWLI), Cologne (DCH/IfDH/IfL) and Hamburg (HZSK/INEL). The centre offers information on current best practices, available resources and tools, and gives advice on technological and methodological matters for researchers working within relevant fields.

Keywords: CLARIN Knowledge Sharing Infrastructure, linguistic diversity, language documentation

1. Introduction

The European digital research infrastructure CLARIN (Common Language Resources and Technology Infrastructure)[1] is not restricted to interconnected technical centres and service units; CLARIN is also actively building a Knowledge Sharing Infrastructure (KSI) to ensure that existing knowledge and expertise in individual centres all across Europe (and beyond) is easily available both for the CLARIN community and for the humanities research communities for which CLARIN is being developed. Within the Knowledge Sharing Infrastructure[2], so called Knowledge Centres[3], or K Centres, play an important role. The Knowledge Centres comprise one or more physical institutions with particular expertise in certain areas and are committed to providing their expertise in the form of reliable knowledge-sharing services.

In this paper, we present the ninth K Centre: CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[4], certified in September 2017. The CKLD is a virtual distributed centre comprising institutions at the Universities of London (ELAR/SWLI), Cologne (DCH/IfDH/IfL) and Hamburg (HZSK/INEL). The mission of the CKLD is to establish a single interface to reliably provide the expertise on various aspects of the thematic focus brought in by the founding members. Section 3. thus describes the areas of competence and specific services of the individual partners.

The knowledge-sharing services provided by the CKLD are the topic of section 4.: The center offers information on current best practices, available resources and tools, and gives advice in technological and methodological matters for researchers working within relevant fields. As a member of the European digital research infrastructure CLARIN, the aim of the CKLD is also to allow for a better integration of these research communities and the resources and tools traditionally used into existing and emerging digital research infrastructures. This also includes further coordinated work on relevant best practices.

Language documentation and research in linguistic diversity share a joint focus on less-widely studied languages. These languages and their varieties are often minority languages and frequently endangered. Further aspects relevant to most of these linguistic settings are multilingualism and language contact, since usually more than one (non-standard) language or variety interact with at least one other. The main objective of language documentation is an annotated multimodal audio-visual corpus as a comprehensive record of the linguistic practices characteristic of a given speech community. Research in linguistic diversity on the other hand comprises descriptive studies and analyses of individual under-resourced languages and comparative typological linguistics that takes into account the world-wide diversity of human languages.

While research questions and methods may differ, both disciplines share many methods, tools, and challenges. The documentation and analysis of smaller, lesser known and endangered languages around the world and related research – typically in a field-work setting – requires technical knowledge and specialized skills. The required expertise comprises recording of speech events, metadata man-

[1] https://www.clarin.eu/

[2] https://www.clarin.eu/content/knowledge-sharing

[3] https://www.clarin.eu/content/knowledge-centres

[4] http://ckld.uni-koeln.de

agement, data handling, (semi)-automatic analysis, documentary and typological corpus and database compilation, archiving of results, among others. Typological research in language diversity often requires additional layers of annotations and the application of search and clustering algorithms to typological corpora and databases.

2. Building a Distributed Centre

Building a sustainable distributed support and knowledge infrastructure for linguistic diversity and language documentation is an organisational and academic challenge. Linguistic diversity research and language documentation are by definition global endeavours and at the same time highly localized. Methodologically, it includes the classical issues of linguistic fieldwork as well as the application of the full range of language-related technologies from the areas of speech technology and digital humanities in well-equipped academic settings and remote field sites. The CKLD approaches this challenge by combining the expertise of language typologists, field linguists, sociolinguists, computer linguists, computer scientists, data curators as well as language archivists from institutions in several geographic locations into a single digital institution. The geographical distribution and variation in object languages and research paradigms of research projects is often an obstacle when it comes to creating synergies and common solutions to similar problems. It is therefore more than desirable that joint initiatives with a technical/infrastructural focus bring together researchers from different disciplines under one subject.

3. Participating Institutions

The CKLD is a joint endeavour of three main founding members located at SOAS University of London (ELAR/SWLI), the University of Hamburg (HZSK/INEL), and the University of Cologne (DCH/IfDH/IfL). Since the founding of the centre, the newly established Leibniz-Centre General Linguistics (ZAS) at the Humboldt University in Berlin has recently joined as a full member, and further partners with suitable profiles are likely to become a part of the centre in the future.

3.1. London (ELAR/SWLI)

The Endangered Languages Archive (ELAR) and the SOAS World Languages Institute (SWLI) provide training in language documentation, data collection, data management and archiving. Training has been a core activity of ELAR and SWLI since their inception in 2002 and 2016 respectively.

The Endangered Languages Archive (ELAR) is a digital repository specialising in preserving and publishing endangered language documentation materials from around the world. ELAR currently holds multimedia collections of endangered languages worldwide (more than 420 languages), with regional strongholds in Africa, Middle East, Asia, Australia, Oceania and Meso- and South America. The collections can be browsed and accessed through the ELAR online catalogue[5]. All materials are digital, free to access, openly available (after free registration) and have a worldwide coverage. ELAR is seen as a reputable and reliable repository and as one of the leading language and culture archives worldwide, being approached by more and more external projects (e.g. American NSF- funded projects, EU projects) within and beyond linguistics as a host repository now that data depositing is seen as an integral part of good practice research. ELAR has a small team of highly specialised archivists, data scientists, audiovisual experts, and linguists, pioneering in the areas of user interface, managed access and accessibility. The archive was originally funded by Arcadia and is since 2014 part of the National Research Library of SOAS University of London.

Starting out 100 years ago with teaching languages of Asia, SOAS now combines language scholarship and disciplinary expertise with a regional focus. It has the largest concentration in Europe of academic staff concerned with Africa, Asia and the Middle East and offers unparalleled expertise in a wide range of non-European languages, e.g. Somali, Khmer or Syriac. Building on SOAS's tradition of non-European language teaching and research, the SOAS World Languages Institute (SWLI) was created in 2016, bring together a wide range of affiliated scholars and research students from across SOAS to make the expertise of its members more widely available to the wider world. A further focus of the Institute is the digital humanities. The institute supports researchers in planning projects with digital components and aims to strengthen SOAS scholars' expertise in the methods of the digital humanities. The SWLI runs various events throughout the year, including talks, seminars, workshops, conferences and round table discussions, as well as book launches and film screenings. It provides bespoke training programmes related to languages and language-based research. The specialist courses can be tailor-made for individuals and groups of different sizes from a range of different backgrounds – including the public and private sector, business, charities, government and NGOs – and include topic such as Language Documentation: theory and methods, annotation software for audio and video data, video recording for scientific analysis, or data management. The SWLI partners with several institutions and projects around the world, as for instance Docip (a Swiss Foundation based in Geneva and Brussels which documents and provides access to records of the UN representation of indigenous peoples from all around the world before international organisations) and Memrise (a language learning company based in East London). At the moment SWLI is part of the UK CLARIN-Consortium and aims at being a C- and possibly a B-Centre.

3.2. Hamburg (HZSK/INEL)

The members of the CKLD within the University of Hamburg are the Hamburg Centre for Language Corpora (Hamburger Zentrum fur Sprachkorpora, HZSK), which is a certified CLARIN B Centre with a thematic focus on spoken and multilingual corpora, and the long-term (18-year) project INEL (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages), which builds on the existing HZSK infrastructure while creating and making available more than ten cor-

[5] https://elar.soas.ac.uk/

pora for endangered and/or lesser documented languages from the Uralic and Altaic language families.

The HZSK was launched in 2011 to create a sustainable institutional framework for the software, corpora and expertise developed within the research data managing project of the Special Research Centre on Multilingualism in Hamburg during twelve years of funding.

The software suite EXMARaLDA (Schmidt and Wörner, 2014) for spoken corpora now developed jointly by the HZSK and the Institute for the German Language (IDS)[6] includes a transcription and annotation editor (Partitur-Editor), a corpus and metadata management tool (Coma) and a corpus search and analysis tool (EXAKT). The CLARIN compatible EXMARaLDA framework allows for import and export of most common transcription formats and for multimedia visualisation of transcription data in various layouts. Parts of the framework are also integrated as web services into the Data Seal of Approval (DSA)[7] certified HZSK Repository[8](Jettka and Stein, 2014), through which the corpora hosted at the HZSK are made available to the academic community.

The corpora at the HZSK (Hedeland et al., 2014) traditionally cover various aspects of individual and societal multilingualism, but also document lesser known regional varieties, such as Porteño Spanish or Faroese Danish, or endangered languages such as Nganasan (Wagner-Nagy and Szeverényi, 2015). For the transparent and efficient curation of corpora, several customized workflows have been developed and implemented at the HZSK using the open source distributed version control system Git[9] and the open source project management software Redmine[10].

The expertise gathered at the HZSK is provided as consulting, support and training on best practices, standards and tools for researchers and research projects working with corpora. The HZSK also developed and implemented the CLARIN-D Helpdesk(Lehmberg, 2015), which is now managed by the HZSK and used to coordinate user support for CLARIN-D centres and CLARIN services.

The INEL project (Hedeland et al., 2016), which started in February 2016, is organized jointly in the framework of the Academies' program of the Federal Republic of Germany by the Union of the German Academies of Sciences and Humanities and the University of Hamburg and institutionally located at the Institute for Finno-ugric/Uralic Languages (IFUU).

Since INEL is making use of the HZSK software and infrastructure, several components are being adapted and extended to meet the needs of the linguistic subprojects, in which richly annotated corpora based on existing archival material and partly supplementing new recordings are being created and made available to the academic community. The extensions include interoperability with commonly used tools such as FLEx[11] and the definition of streamlined collaborative Git-based workflows for the curation of archival materials and the publication of corpora of glossed data with rich sociolinguistic metadata. The INEL project also provides specific expertise on fieldwork and curation of archival materials for indigenous languages.

The synergetic alliance of the HZSK as an infrastructure unit and INEL as a long term project with their particular expertise in a broad spectrum of aspects in linguistic diversity and language documentation enables both of them to provide solid support for the linguistic community in these overlapping areas.

3.3. Cologne (DCH/IfDH/IfL)

The Department of Linguistics (IfL), the Data Center for the Humanities (DCH), and the Digital Humanities Department (IfDH) of the University of Cologne are partners in CKLD. The Department of Linguistic is a center for research and teaching in the areas of language documentation, methodologies of language documentation, and language diversity (Himmelmann, 2008; Himmelmann, 2012). It has been active in the documentation of endangered languages since the very beginning of the field in the early 1990s (Himmelmann, 1998). Since then, numerous research projects with a focus on language documentation were and are based in Cologne. Language documentation and language diversity are the core of the linguistics curriculum in Cologne. Most of the teaching and research staff of the general linguistics department are working on either linguistic diversity or language documentation or both and the department offers several courses related to theory and methodology of language documentation and linguistic diversity research every year. The Data Centre for the Humanities (DCH) is a facility of the faculty of arts and humanities. The DCH provides methodological and technical support to research projects and provides services in the areas of data management, curation and archival storage (Sahle and Kronenwett, 2013). DCH and IfL have developed software for language documentation, archiving[12] and analysis. (Blumtritt et al., 2013) The Language Archive Cologne[13] is a joint endeavour of the linguistics department and the DCH. Together with the linguistics department and other partners, the DCH is extending the Language Archive Cologne in the KA$\hat{3}$ project (Cologne Center Analysis and Archiving of Audio-Visual Data)[14] funded by the German Ministry of Research (BMBF). The newly established Digital Humanities Department features one chair explicitly dedicated to digital linguistics. The Digital Humanities Department will also contribute expertise and staff to the K-Centre.

4. Services

Building and running a distributed knowledge centre for linguistic diversity and language documentation requires a flexible and reliable support and knowledge-sharing infrastructure. The services offered by the knowledge centre consist of a web presence providing information on the relevant

[6] http://www.ids-mannheim.de

[7] https://www.datasealofapproval.org/

[8] https://corpora.uni-hamburg.de/repository

[9] https://git-scm.com/

[10] www.redmine.org/

[11] https://software.sil.org/fieldworks/

[12] http://cmdi-maker.uni-koeln.de

[13] https://lac.uni-koeln.de/en/

[14] http://dch.phil-fak.uni-koeln.de/ka3.html

topics, a helpdesk allowing researcher to interact directly with the centre, as well as training courses offered on a regular basis.

4.1. The Website as a Portal to the Centre

The CKLD website[15] operated by the University of Cologne is the main portal to the Knowledge Centre. The site offers a description of the centre and a list and explanation of its services. Besides providing a collection of informational material and further relevant links to software tools, language resources, manuals, and documentation of best practices, the website is also the entrance point into the interaction with the CKLD Helpdesk.

4.2. Support through the CKLD Helpdesk

The CKLD Helpdesk is integrated into the CLARIN-D Helpdesk operated by the HZSK in Hamburg. The use of a ticketing system (OTRS)[16] and proven support workflows allows for efficient handling of incoming inquiries. For each question, an automatic confirmation mail is sent to the user. Within two working days, questions will either be answered directly by the first-line support at the HZSK, or, if needed, forwarded to experts within the CKLD while keeping the user updated on the status of the inquiry. The incoming inquiries also provide the CKLD with valuable information on which topics need to be covered more in detail on the website or in dedicated trainings. The CKLD Helpdesk provides information and guidance to researchers in the preparation and during the execution of language documentation and other field-work based linguistic research projects as well as typological research. This includes questions relating to equipment, digital tools, methods, where to find data and information, whom to contact for specialist information on particular regions or language families.

4.3. Teaching and Training Modules

The participating institutions offer courses in language documentation, corpus creation, data analysis and research data management. ELAR and SWLI are one of the leading providers in trainings in theory and method in modern language documentation. The HZSK regularly offers trainings for the EXMARaLDA software or project specific trainings on research data management. With the distributed expertise, the CKLD can offer various training modules on topics such as research data management, publication and sustainability, including ethical and legal issues, but also hands-on trainings that provide researchers and students with the practical knowledge required to actually create high quality language resources. The partners coordinate their training activity and communicate their activities via the CKLD website.

5. Outlook

The newly established Knowledge Centre CKLD aims to act as a contact point for all researchers in the fields of linguistic diversity and language documentation. While working to provide researchers and students with relevant information, support and trainings, the CKLD is also a cooperation between centres in different geographical locations, which also on the level of object languages focus on different geographical areas respectively, but which can align their work on best practices for the creation and publication of language resources to jointly work towards better interoperability and integration into existing and emerging international digital research infrastructures.

[15] http://ckld.uni-koeln.de

[16] https://www.otrs.com/

6. Bibliographical References

Blumtritt, J., Bouda, P., and Rau, F. (2013). Poio API and GraF-XML: A radical stand-off approach in language documentation and language typology. In *Balisage: The Markup Conference 2013, Montréal, Canada, August 6 - 9, 2013. In Proceedings of Balisage: The Markup Conference 2013*.

Hedeland, H., Lehmberg, T., Schmidt, T., and Wörner, K. (2014). Multilingual corpora at the hamburg centre for language corpora. In Michael Haugh, et al., editors, *Best Practices for Speech Corpora in Linguistic Research*, pages pp. 208–224. Cambridge Scholars Publishing.

Hedeland, H., Lehmberg, T., and Wagner-Nagy, B. (2016). Digitale workflows in langzeitprojekten am beispiel einer infrastruktur zur dokumentation indigener nordeurpäischer sprachen (inel). In *DHd 2016: Modellierung - Vernetzung - Visualisierung, (Leipzig, 7-12 March, 2016)*, pages 152–155, Leipizg. Digital Humanities im deutschprachigen Raum e.V.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36(1):161–196.

Himmelmann, N. P. (2008). Reproduction and Preservation of Linguistic Knowledge: Linguistics' Response to Language Endangerment. *Annual Review of Anthropology*, 37(1):337–350.

Himmelmann, N. P. (2012). Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation*, (6):187–207.

Jettka, D. and Stein, D. (2014). The HZSK repository: Implementation, features, and use cases of a repository for spoken language corpora. *D-Lib Magazine*, 20(9/10).

Lehmberg, T. (2015). Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities. In *Forschungsdaten in den Geisteswissenschaften (FORGE) 2015, (Hamburg, 5-18 September, 2015)*, pages 25–27, Hamburg.

Sahle, P. and Kronenwett, S. (2013). Jenseits der Daten. *Libreas*, (23).

Schmidt, T. and Wörner, K. (2014). Exmaralda. In Ulrike Gut Jacques Durand et al., editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Wagner-Nagy, B. and Szeverényi, S. (2015). Linguistically annotated spoken nganasan corpus. *Tomsk Journal of Linguistics and Anthropology*, 2:25–33.