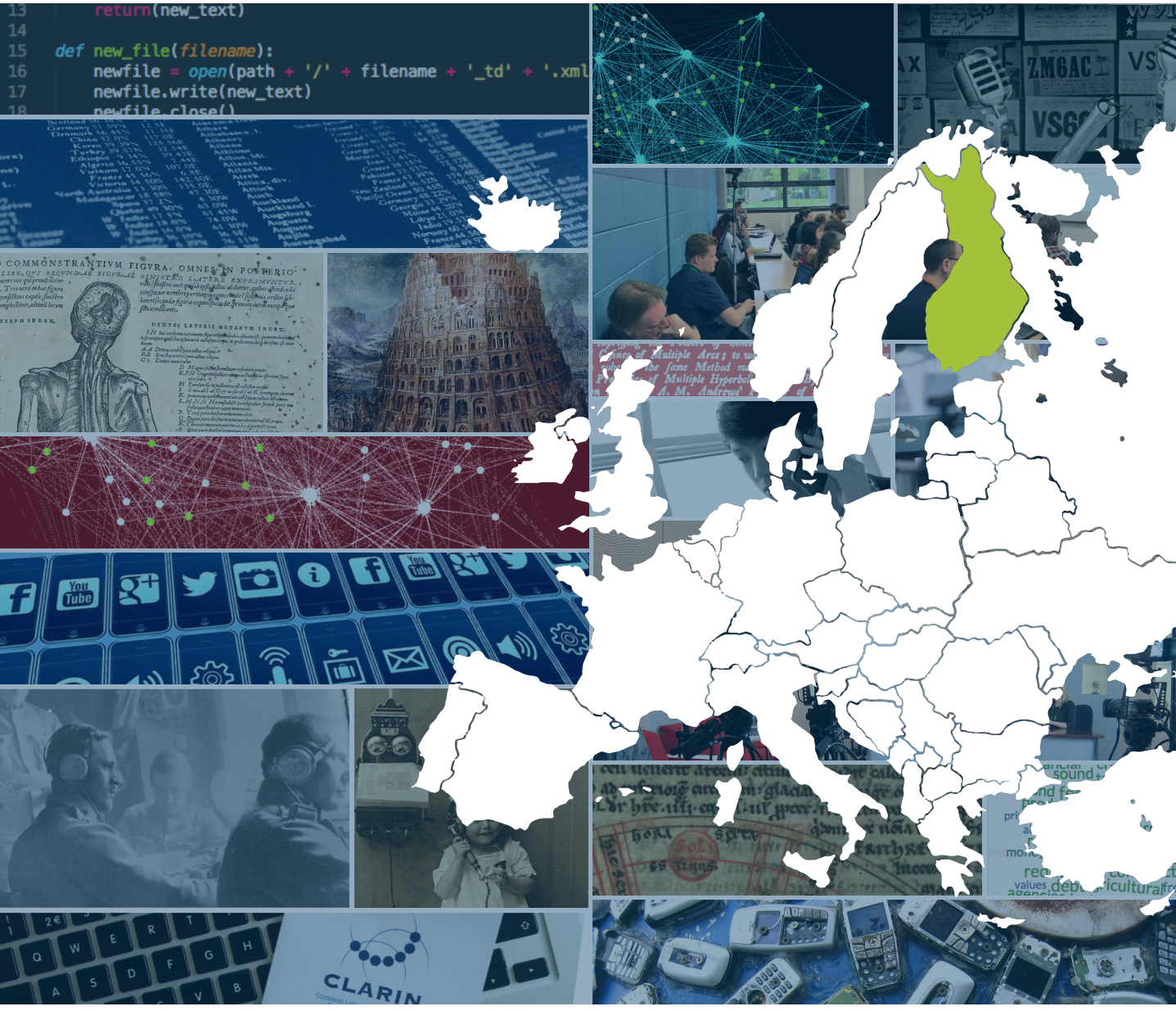
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Finland



Written and edited by Darja Fišer and Jakob Lenardič

Foreword

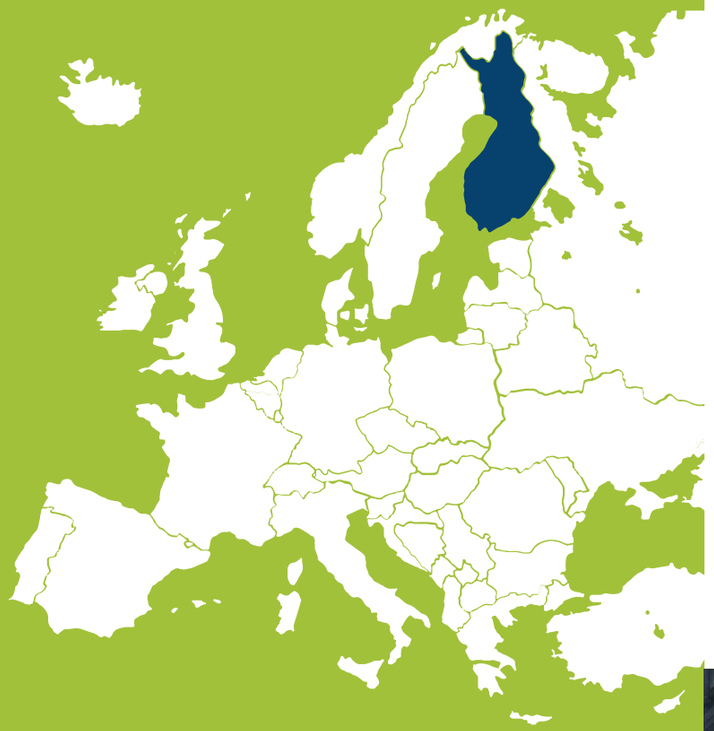
Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents Finland and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

Turku, Finland | photo by Patrick Selin | Unsplash

Finland



Written by Darja Fišer and Jakob Lenardič

The Finnish national consortium FIN-CLARIN[1] has been a CLARIN member since 2015. The members of the consortium are the University of Helsinki, the University of Eastern Finland, the University of Jyväskylä, the University of Oulu, the University of Tampere, the University of Turku, the University of Vaasa, the Institute for the Languages of Finland, the Helsinki Institute of Technology and the IT Centre for Science (CSC). The national coordinator for FIN-CLARIN is Research Director Krister Lindén.

FIN-CLARIN has been actively engaged in developing tools and resources that have become a staple of Finnish researchers working with language data. Through the Language Bank of Finland, which is a certified CLARIN B-Centre, researchers can access dozens of Finnish corpora, which are in most cases available through online interfaces such as Korp.

A flagship resource provided by the Finnish consortium is the Suomi 24 Sentences Corpus, a corpus that compiles texts from discussion forums of the popular Suomi24 online networking website. The data from the corpus is currently being analysed in the framework of the Citizen Mindscapes project, which seeks to uncover "trends and shifts in attitudes in connection to societal phenomena" in Finland, thus making the corpus an extremely important resource that highlights how corpus-based linguistics can lead to a greater understanding of society at large. Read more about this corpus on page 7.

The Finnish consortium is actively engaged with ground-breaking researchers working in Digital Humanities and Social Sciences who make use of the consortium's resources and tools. The Language Bank hosts a "Researcher of the Month" archive, intended to highlight both the work of the prominent researchers and the tools and resources of potential use to researchers.

In 2016, Finland organised 22 user involvement events. A very successful event was the Roadshow organised to celebrate 20th anniversary of the Language Bank of Finland. It consisted of a series of seminars at all the member organisations of the FIN-CLARIN consortium.

[1] https://www.kielipankki.fi/organization/

The Language Bank of Finland hosts a variety of language tools, for instance the following publicly available ones:

- Finnish Parse, which is a powerful dependency parser that is capable of tokenisation, sentence splitting, morpho-syntactic tagging and parsing, and can be applied to plain Finnish text with extremely high accuracy
- Aalto-ASR, a continuous speech recogniser that can handle a large amount of Finnish vocabulary
- the Helsinki Finite-State Transducer Technology that provides software for morphological analyses of various European languages, and
- the Proto-Indo-European Lexicon, which acts as a generative etymological dictionary, providing data on word origins and historical changes for the hundred most ancient Indo-European languages.

FIN-CLARIN Team | Back row: Atro Voutilainen, Senka Drobac, Pekka Kauppinen. Middle row: Maria Palolahti, Erik Axelson, Jyrki Niemi; front row: Krister Lindén (Research Director), Mietta Lennes, Jussi Piitulainen. Not in the photo: Tero Aalto, Imre Bartis, Ute Dieckmann, Sam Hardwick, Martin Matthiesen.

Helsinki, Finland | photo by Alexandr Bormotin | Unsplash

AaltoASR

Written by Darja Fišer and Jakob Lenardič

The AaltoASR project,[2] which is led by Professor Mikko Kurimo at Aalto University, focuses on the development of an Automatic Speech Recognition system that is able to transcribe spoken Finnish language with a very high accuracy rate. The system, which started as a relatively simple spoken-language recogniser in the 1980s that was at first capable of handling around 1,000 Finnish words, is today a complex piece of software that can recognise and transcribe not only isolated words but also spontaneous speech. The AaltoASR system comprises of complex procedures that accurately transform audio signals into linguistically-modelled speech units on the basis of a complex network of probabilistic distributions, thus making the system easily adaptable to various domains and styles.

By focusing on complex agglutinative languages such as Finnish and Estonian and under-resourced languages such as the Sami ones, the AaltoASR team continues to make groundbreaking progress in the development of a successful large vocabulary speech recogniser that is able to tackle complex inflectional and compounding systems, which otherwise make it difficult to perform rule-based morphology analysis and, by extension, speech recognition. The AaltoASR tool is open source, and the developer version can be found on the tool's GitHub page. AaltoASR is available for research use via the Language Bank. ASR systems built on top of AaltoASR tools are also used by companies for subtitling TV broadcasts in Finland and Sweden.

The most recent papers by Prof. Kurimo and his colleagues include:

Mansikkaniemi, A., Smit, P., and Kurimo, M. (2017). Automatic Construction of the Finnish Parliament Speech Corpus. In Interspeech 2017. http://dx.doi.org/10.21437/Interspeech.2017-1115.

Smit, P., Virpioja, S., and Kurimo, M. (2017). Improved subword modeling for WFST-based speech recognition. In Interspeech 2017. http://dx.doi.org/10.21437/Interspeech.2017-103.

[2] https://github.com/aalto-speech

The Suomi24 Corpus

Written by Darja Fišer and Jakob Lenardič

The Suomi24 corpus[3] is a comprehensive collection of texts from discussion forums of Suomi24, which is Finland's largest and most popular social media website and is used by 86% of Finns every month. The corpus contains more than 2.6 million tokens of texts from 2001 to 2016 and is tokenised and morpho-syntactically tagged with the Turku Dependency Parser. A version of the corpus where the sentences are scrambled is publicly available through the web interface Korp under the CLARIN ACA - NC licence, while researchers who have a username and a password can download the entire corpus in the VRT format.

The corpus is used by researchers working in the Citizen Mindscapes project (2016–2019), funded by the Academy of Finland. The aim of the project is a far-reaching socio-political and linguistic analysis of the everyday discourse that is part and parcel of the Finnish society. By applying a wide range of quantitative and qualitative methods such as statistical data analysis and thematic interviews and by making use of advanced language tools to process the data within the corpus, Citizen Mindscapes researchers, who are led by Professors Jussi Pakkasvirta and Krista Lagus, seek not only to uncover the current societal and political trends in Finland, but also pinpoint those features of the online discourse that may very well hint at the prospective evolution of the Finnish society as a whole.

To make the presentation of the complex data within the Suomi24 corpus as optimal as possible for socio-political analysis, the Citizen Mindscapes team is developing the Social Thermometer. This novel visualisation method helps researchers detect deep-rooted views related to complex issues such as nationalism, which often begin in and are shaped by discussions on the Internet.

The Citizen Mindscapes project is thus a pivotal multidisciplinary endeavour that has brought together and established long-term collaborations between researchers from diverse fields such as natural language processing, sociolinguistics, sociology and political studies. Wishing to promote open science and open data and thereby support the development of novel approaches in social sciences and NLP, the researchers in Citizen Mindscapes plan to make their data sets and tools available within the Language Bank of Finland in collaboration with FIN-CLARIN and the Centre for Science.

Follow Citizen Mindscapes on Twitter: @mindscapes24

Researcher and Helsinki Challenge semi-finalist Krista Lagus with the Citizen Mindscapes research project team (photo by Linda Tammisto).

[3] http://urn.fi/urn:nbn:fi:lb-2017021506

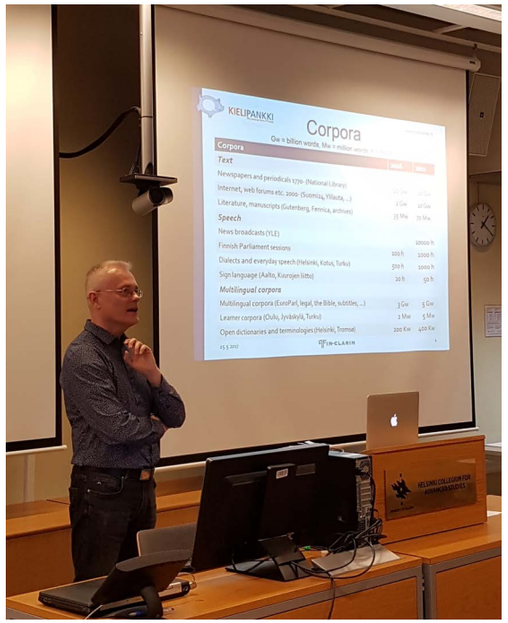
The Language Bank of Finland's Roadshows

Written by Darja Fišer and Jakob Lenardič

The Finnish consortium FIN-CLARIN is one of the most active ones in the CLARIN network in the field of user involvement. In 2016, they organised the greatest number of user involvement events. Among these, their roadshow events, which were organised to celebrate the 20th anniversary of the founding of the Language Bank of Finland, stand out in particular. Through the course of the roadshows, FIN-CLARIN presented the tools and resources of the Language Bank at all the Finnish universities that are members of the Finnish consortium, beginning with the University of Turku on 28 April 2016 and ending with the University of Jyväskylä on 15 November 2016. Each presentation began with a demonstration of the Korp interface and of various ways in which researchers can use it to search through the text and speech corpora that are available in the Language Bank. The presentation materials are available both in English and Finnish.

To conclude each event, presentations were given by researchers who work at the hosting universities and who have used the tools and resources of the Language Bank. For instance, when the roadshow stopped at Aalto University on 27 September, André Mansikkamäki presented speech recognition tools developed within the AaltoASR project, which we highlighted on page 6. When the roadshow concluded for 2016 at the University of Jyväskylä, Tommi Jantunen, whose interview can be read on page 9, gave a presentation on the Finnish Sign Language and Jarmo Jantunen on the usage of the Suomi24 corpus, which we present on page 7.

Because of the success of the 2016 events, FIN-CLARIN continued the roadshows in 2017. After a brief pause in 2018 when FIN-CLARIN was busy carrying out other UI initiatives, the roadshows are expected to continue in 2019 with a focus on introducing newly developed tools and corpora.



Speaker: Krister Lindén, National Coordinator of FIN-CLARIN at the Helsinki Collegium of Advanced Studies, 15 May 2017 (photo by Mietta Lennes).

Tommi Jantunen

Tommi Jantunen is a linguist specialising in the Finnish Sign Language working at the University of Jyväskylä. The following interview took place via Skype on Tuesday, 23 May 2017 and was conducted and transcribed by Jakob Lenardič and edited by Darja Fišer.

1. Could you tell us a little bit about yourself, your background and your current work?

My name is Tommi Jantunen and I am an academic research fellow at the Department of Language and Communication Studies at the University of Jyväskylä. I have an MA in General Linguistics from the University of Helsinki and a PhD in Finnish Sign Language from the University of Jyväskylä. I am currently running the last year of my five-year-long project ProGram, which is funded by the Academy of Finland and in which my colleagues and I are investigating the grammar and prosody of the Finnish Sign Language. At the University of Jyväskylä, the Finnish Sign Language is one of the main fields that people can get a degree in, and I was one of the first people to obtain a PhD degree in Finnish Sign Language at our university almost ten years ago in 2008. I don't think that there are many universities in Europe and even in the world in general where one can fully focus on studying sign languages as a major subject, on par with more traditional fields like Finnish or English studies, so it's really excellent that our university offers this opportunity.

2. How did you hear about CLARIN and how did you get involved?

I think I knew about CLARIN ERIC even before I became directly involved with our national consortium. This is probably so because all Finnish universities are part of the FIN-CLARIN network so if you are a researcher working with languages it's almost impossible not to hear about CLARIN. I think I became involved a few years ago at the start of our current project ProGram. We needed to publish the data that we were starting to work on, so I contacted our FIN-CLARIN UI representative Mietta Lennes from the University of Helsinki and she was happy to get our project involved with the consortium.

3. How has CLARIN influenced your way of working and how was this received in your research community?

I think the easiest way to approach this question is from the following point of view. If you're a researcher and you are in the process of applying for research funding, one of the prerequisites for obtaining funding is that you draft a plan of how the data that you are working on can be related to and included within research infrastructures such as FIN-CLARIN. In other words, it is the funding agencies that require researchers to connect their data with a ready-made infrastructure so that the data become openly available. Related to this, FIN-CLARIN has proven itself to be an exceptionally good collaborator – whenever we write new applications for research funding, it is extremely easy for us to connect our work with the FIN-CLARIN and CLARIN ERIC infrastructures.

When it comes to sign language, we unfortunately can't use much of the existing services within the infrastructures because of the specificity of our field, which requires specialised tools and resources. However, we are building our own tools and services in collaboration with FIN-CLARIN, which we of course plan to make available within the FIN-CLARIN repository – that is, the Language Bank of Finland. For instance, my colleagues are currently compiling a very exhaustive Finnish Sign Language corpus, which is planned to be made available through FIN-CLARIN and which will mark the end of the project.

[4] http://users.jyu.fi/~tojantun/ProGram/ProGram.html

4. Could you describe the status of the Finnish Sign Language?

I think its status is relatively good, especially in comparison with some other countries. The Finnish Sign Language is recognised in the Finnish Constitution as a national minority language, and there's a sign language law that clarifies certain linguistic etc. rights for sign language users. The Finnish Sign Language came into the spotlight when a deaf rap artists called Signmark[5] won second place in Finland's national qualifications for Eurovision in 2009. Apart from the Finnish Sign Language, there's also the Finland-Swedish Sign Language, which is a kind of minority-within-a-minority language spoken by relatively few people in the Swedish speaking areas of Finland. Linguistically, these two sign languages are very close to each other. Ultimately, it is a political decision that governs the question what is a specific language and what isn't.

5. To what extent in your opinion is the sign language research community benefiting from digital tools and resources, such as the ones provided by CLARIN infrastructure?

I think one of the most important, and also the most invisible, aspects of what CLARIN ERIC and the national consortia have been doing is the work related to standardisation and the resolution of metadata issues. I don't know if we would even be able to do our corpus work if it weren't for the standards created in FIN-CLARIN. The LAT platform is also extremely important, as it allows us to publish our visual data. For instance, on the LAT platform we have already published a richly-annotated video corpus of the Finnish Sign Language. This corpus comprises a rather small sample that is much more deeply annotated than our final exhaustive corpus is going to be, and I also use this corpus to showcase Finnish Sign Language data to my students and colleagues.

6. What kind of specific methodological and technical challenges does a researcher working on sign language face with respect to the available infrastructure?

The biggest challenge is related to the videos, which serve as basically the only way you can record and compile sign language data. On the one hand, working with videos is challenging from a technical perspective; it is difficult to search through visual data and they take up a lot of storage space. On the other hand, videos raise many privacy and legal issues, and we have to be extremely careful when it comes to getting informed consent. One of the problems we had in connection with this is that there was very little information about the necessary steps we needed to take in order to fulfil all the legal requirements. Additionally, it has been my experience that annotating the sign language data is extremely time consuming, more so than typical speech data.

7. How would you recommend your colleagues to get involved with CLARIN and start using the available infrastructure?

My advice to all researchers who are interested in working with languages is that they contact their respective UI representatives. In Finland, Mietta Lennes has been doing an excellent job in promoting FIN-CLARIN and CLARIN-ERIC. She has taken part of practically every linguistic conference in Finland where she has kept FIN-CLARIN very visible. In terms of visibility, I think we introduce CLARIN even to our BA students, so a researcher in Finland can't really avoid knowing something about CLARIN. I think a more important question here is when is it that a researcher needs to know something about CLARIN. This is at the point when you start doing your own research and when you apply for funding – as I've said, the funding agencies require that the data obtained in your research get connected with the existing infrastructures, and FIN-CLARIN, as well as CLARIN ERIC in general, provides an excellent infrastructure for this.

[5] https://en.wikipedia.org/wiki/Signmark

8. What resources, tools and services from CLARIN would you recommend to your colleagues? What would you recommend CLARIN to do in order to attract more researchers from your community?

I'm mostly familiar with the previously mentioned LAT platform and the ELAN tools, which we have been using to annotate our visual data. It is also great that such tools are available as web services, which makes them very easy to use. As to the second question, I think that FIN-CLARIN has been doing an exemplary job already, so I think it's impossible to recommend anything in terms of improvement. Last year, the roadshow event organised by FIN-CLARIN, which took place in autumn, also reached our university and I gave a presentation on Finnish Sign Language then.

9. What's your vision for CLARIN 10 years from now?

I would like to see more and more video materials and tools related to sign language processing in the repository.

10. Describe CLARIN in three words.

Internationality, openness and user-friendliness!

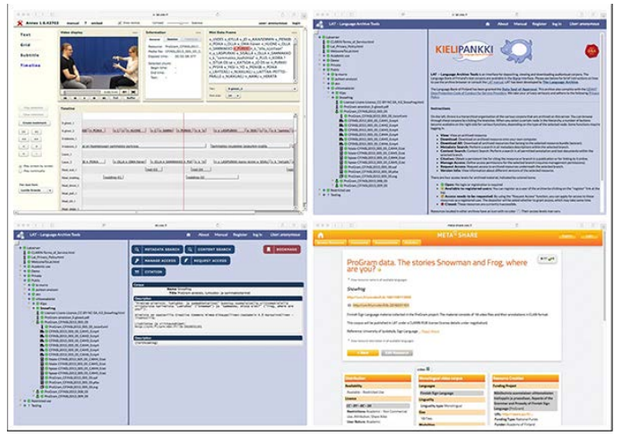


Figure 1: The ProGram corpus accessed via the FIN-CLARIN infrastructure.

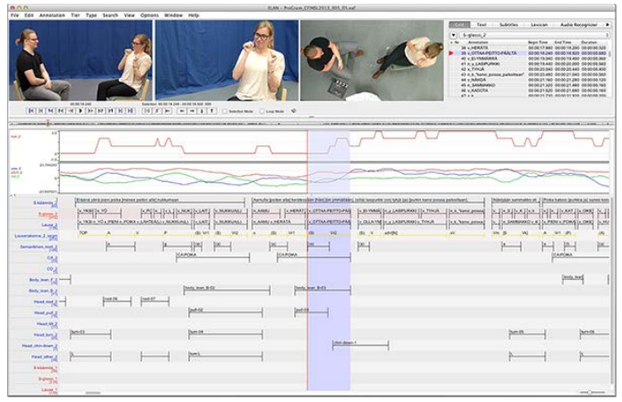


Figure 2: An annotated video corpus of the Finnish Sign Language.

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number: CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Written by
Darja Fišer and Jakob Lenardič

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN 2018