CLARIN

Common Language Resources and
Technology Infrastructure

Tour de CLARIN

CLARIN Knowledge Centre for the Languages of Sweden

PERVI AKT.
PERVI NASTOP.
Dvoriſhe pred Shupanovo hiſho ; doli na uni plati ſe sazhne en borſht , na ti drugi plati je pole , inu ſe delezh vidi.
Jaka , Mizka.
Jak. Lubesniva Mizka, povęj mi vùnder, sakaj Ansheta nozhesh? — ali ni on en lepi sali fant? — ali nima lepo hiſho , lępe njive ; shivinzo pak tako , de ſe enimu ſerzę ſmeja , kader jo vidi.
Miz. Ozha, povejte meni , sakaj vi kiſlu vinu radi nepijete?
Jak. Shema ! sato, ke mi ne duſhy.
Miz. Prov ! jeſt tudi Ansheta nozhem, sato, ke mi ne dopade.
Jak. Ti ſhentanu Deklé , sdej ſi me plazhala — — ampak zhakej , sdej jeſt tebe praſham, sakaj tebi Anshe nedopade ?
Miz. Sa to, ke mi nedopade — — —
Jak. Imaſh ſpet prov. Deklè, ti imaſh vezh pameti , koker tvoj Ozha — no ,
A 2 jeſt
Lanfranc ['lænfræŋk] id.; Langland ['læŋlənd] id.
langrage, -edge ['læŋgridʒ] kartets.
lang syne ['læŋ'sain] (Sc) langgeleden; (goede) oude tijd; zie auld.
language ['læŋgwidʒ] taal, spraak;
bad ~; use (bad) ~, vloeken, gemeene
gebruiken; '~-master, '~-teacher, leeraar, -onderwijzer.
languid ['læŋgwid] kwijnend, zwak, flauw
loom, lusteloos, mat, slap,
languish ['læŋgwiʃ] (weg)kwijnen; smachten, verslappen;
74ACT244
27C512-120
ELCAP1
220uF/16V
PAL10C
getElementsByTagName
querySelectorAll
appendChild
toLowerCase
firstChild

Edited by **Darja Fišer** and **Jakob Lenardič**

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, 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 and knowledge centres 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.

The brochure presents the CLARIN Knowledge Centre for the Languages of Sweden and is organized in two sections:

- Section One presents the members of the Knowledge Centre and their work
- Section Two includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the Knowledge Centre’s infrastructure in their research

CLARIN Knowledge Centre for the Languages of Sweden

CLARIN Knowledge Centre for the Languages of Sweden

Introduction

Written by **Rickard Domeij**

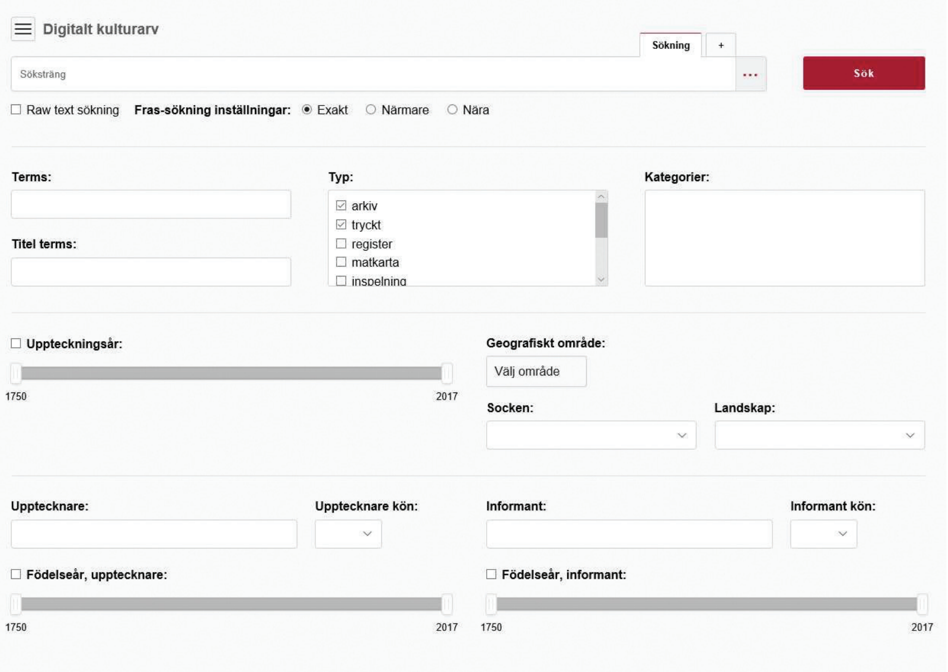
The SWELANG Knowledge Centre is an information service offering advice on the use of digital language resources and tools for Swedish and other languages spoken in Sweden, as well as other parts of the intangible cultural heritage of Sweden. The centre is based at the Language Council of Sweden (Stockholm) and is run in cooperation with the other sections of the Institute of Language and Folklore (ISOF) in Uppsala and Gothenburg. The institute is sanctioned by the Swedish government to collect, preserve, process and disseminate scientific knowledge and material concerning the Swedish language, the national minority languages, the Swedish sign language and Swedish dialects, as well as other parts of the intangible cultural heritage of Sweden.

Development of Digital Tools and Services

The SWELANG knowledge centre cooperates closely with SWE-CLARIN and the National Language Bank of Sweden (Nationella språkbanken). The knowledge centre focuses on developing methods for collecting two types of data:

- Official texts and terminology for research in official communication and social conditions. The material is multilingual with parallel texts in Swedish and translations into easy-to-read, plain language of the five national minority languages (Finnish, Sami, Romani, Yiddish and Meänkieli), as well as other minority languages used in official communication.
- Folk narratives, as well as other text and speech material from the dialect and folklore archives. The material consists of inventories, dialect word databases, letters, recordings, transcriptions, etc. It is important both in terms of content and linguistic quality, as it includes a large number of geographical, social, and stylistic varieties.

In addition, the centre is developing methods to manage and make widely available contextualized digital archive material through a map-based research interface called Digitalt kulturarv (Digital Cultural Heritage). The interface is connected to a database of 16,000 complete records. Apart from text material, consisting of transcribed records that were scanned using OCR or HTR, the database also contains metadata, such as year of recording, categories and location, as well as information about the person recording and informants (i.e. name, year of birth, gender). The interface shows not only a list with search results, but also visualizes statistics from the metadata. For example, a map illustrates the geographical distribution of the records. A limited public version of Digitalt kulturarv called Sägenkartan (Map of Legends) can be accessed on the web (in Swedish only). A log-in version with richer content for researchers is on its way.



☐ Raw text sökning   Fras-sökning inställningar: ● Exakt ○ Närmare ○ Nära

Terms:

Titel terms:

Typ: ☑ arkiv ☑ tryckt ☐ register ☐ matkarta ☐ inspelning

Kategorier:

☐ Uppteckningsår: 1750 2017

Geografiskt område: Välj område

Socken:

Landskap:

Upptecknare:

Upptecknare kön:

Informant:

Informant kön:

☐ Födelseår, upptecknare: 1750 2017

☐ Födelseår, informant: 1750 2017

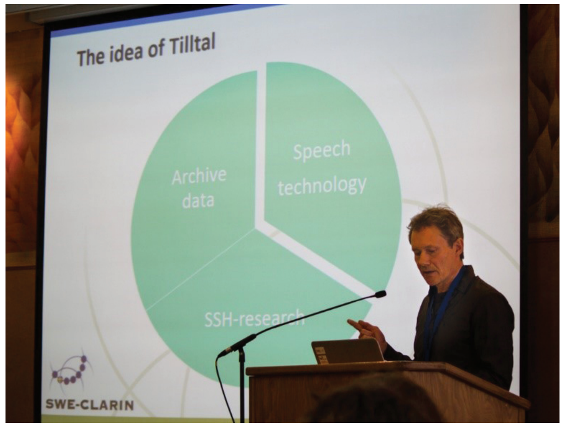
Figure 1: *The search interface of the Digital Cultural Heritage*

The knowledge centre is also developing an infrastructure for dictionaries, the aim of which is to store and make available official terminology and dialect words in collaboration with Språkbanken (a CLARIN B centre). Resources offered by the SWELANG Knowledge Centre are already available through the SWE-CLARIN catalogue. This mostly includes bilingual dictionaries that pair Swedish with other languages spoken in the country, such as the Swedish-Bosnian dictionary and the Swedish-Azerbaijani dictionary.

Interdisciplinary collaboration within the TillTal project

In the TillTal project we examine how speech and language technology methods can make the historical speech recordings more accessible for research in cooperation with data holders, researchers and speech and language technologists. For instance, there are immense amounts of recorded interviews which currently have to be played in real time in order to be analysed. These materials conceal a wealth of information of great interest for the Humanities and Social Sciences.

With digital tools we see possibilities to explore the recordings in new ways. For example, we are exploring methods to visualize and browse large amounts of audio data together with the CLARIN Knowledge Centre of Speech Analysis at KTH (Malisz et al. 2017). This is done by projecting sound segments on a two-dimensional plane with a technique used to find similarities in images, so that representations of similar sounds are clustered together. We hope that this will make it possible to find interesting features in audio files without actually listening to them one by one, for example to identify applause and singing from speech, or even find similar vowel pronunciations. Our archives also include a wide range of information in written form, including descriptions of recording situations and manual transcripts, which we use to provide further pathways into the speech materials (Domeij et al. 2019).



Rickard Domeij presenting the TillTal project

Associated project collaborations

The K-Centre is part of the following national and international language infrastructure collaborations:

- CLARIN — the European research infrastructure for language resources and technology
- ELRC — European Language Resource Coordination
- eTranslation TermBank — collection and provision of terminological resources for machine translation within the EU
- META-NET, a Network of Excellence consisting of 60 research centres from 34 countries, is dedicated to building the technological foundations of a multilingual European information society
- SWE-CLARIN
- TillTal project

References:

Borin, L., Forsberg, M., Edlund, J., and Domeij, R. 2018. Språkbanken 2018: Research Resources for Text, Speech, and Society. Poster DHN I: Mäkelä, Eetu, Tolonen, Mikko and Tuominen, Jouni (eds.) *Digital Humanities in the Nordic Countries 3rd Conference,* 504–506. http://ceur-ws.org/Vol-2084/poster7.pdf.

Berg, J., Domeij, R., Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S., and Öqvist, J. 2016. Tilltal – making cultural heritage accessible for speech research. Paper presented at the CLARIN Annual Conference 26–28 October 2016, Aix-en-Provence, France.

Berg, J, Domeij, R, Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S., AND Öqvist, J. 2017. Involving users and collaborating between disciplines in making cultural heritage accessible for research. Paper presented at the CLARIN Annual Conference 18–20 September 2017, Budapest, Hungary.

Dagsson, T. and Skott, F. 2018. Digital Cultural Heritage — a Digital Folklore Archive [Blog post]. https://sweclarin.se/eng/digital-cultural-heritage-%E2%80%94-digital-folklore-archive.

Domeij, R. and Eriksson, G. 2018. Språkbanken Sam. A CLARIN knowledge center for the languages of Sweden. Poster presented at SLTC 2018 , 20–22 November at Stockholm University.

Domeij, R., Eriksson, G., Lindström, E., Magnusson Petzell, E., Nylund Skog, S., Skott, F., and Öqvist, J. 2019. Text as an entry point to speech – a journey into the most inaccessible areas of the archives. Book of abstracts 4th Conference of the Association Digital Humanities in the Nordic Countries Copenhagen, March 6–8 2019.

Nylund Skog, S. 2018. From personal letters to scientific knowledge: The creation of archived records in a tradition archive. In *Visions and Traditions: Knowledge Production and Tradition Archives*. Helsinki: Academia Scientiarum Fennica, FFC 315.

Malisz, Z., Öqvist, J, Fallgren, P., Edlund, J., and House, D. 2017. Visualizing vocalic variability in space and time – automatic exploration of "found data". Paper presented at the 47th Poznań Linguistic Meeting, 18–20 September 2017, Adam Mickiewicz University, Poznań, Polen.

Interview | **Susanne Nylund Skog**

Susanne Nylund Skog is an ethnologist and folklore researcher who collaborates with the SWELANG K-Centre in the TillTal project.

Please describe your academic background

<

I am a researcher at the Institute for Language and Folklore in Uppsala, Sweden, where I work at the Department of Dialectology and Folklore Research, and an Associate Professor of ethnology at Uppsala University and of Nordic folklore studies at Åbo Akademi University, Finland.

I defended my doctoral dissertation in ethnology and childbirth stories at the University of Stockholm in 2002, and have since then done extensive ethnographic research on Jewish life in Sweden and on stories by birdwatchers. With performance and narrativity in focus, I have explored issues such as anti-Semitism, whiteness, intertextuality, emotions and materiality. I am currently doing research on archive collections within the project TillTal aimed at making spoken cultural heritage accessible for research, which is funded by Riksbankens Jubileumsfond, the Swedish Foundation for Humanities and Social Sciences.

>

How did you get involved with the K-Centre for the Languages of Sweden? What is the main goal of the Tilltal project?

<

I first came in contact with Rickard Domeij and Gunnar Eriksson from the K-Centre at the SWE-CLARIN exploratory workshop for researching audio materials from a cross-disciplinary perspective. The workshop ended in a joint research grant proposal for the multidisciplinary project Tilltal by the Institute for Language and Folklore, KTH Royal Institute of Technology and the Swedish National Archives.

The overall goal of the project is to make Sweden's archive of recorded speech more accessible for Humanities and Social Science research, which is also one of the main goals of the K-Centre. I am involved in the project as a qualitative researcher who studies the recordings, and I collaborate with language technologists like Gunnar Eriksson who help me with technological solutions for my research questions.

>

How do speech recordings differ from other materials used in Digital Humanities research? What does the TillTal project do to promote the use of speech recordings in multidisciplinary approaches?

<

Speech recordings represent a seriously underutilized resource of the Swedish memory institutions, at least for Humanities and Social Sciences purposes, where researchers often only work with secondary materials, such as transcriptions of the spoken materials, instead of investigating the recordings themselves. One problem is that the number of speech recordings is very large. The archives of the Institute for Language and Folklore alone contain around 25,000 hours of recorded speech. Paradoxically, this contributes to the fact that such materials are not often used by Humanities researchers, as speech is extremely challenging and time-consuming to work with and can be quite unmanageable without appropriate tools.

To help overcome this problem, the TillTal project has established three different case studies and one user study.[1] In the case studies, research agendas from three different Humanities and Social Sciences fields are being pursued with the help of speech technologies. These are case 1: from personal experience narratives to cultural heritage, which focuses on speech recordings in ethnology, case 2: linguistic variation in time and space, which involves collaboration between speech and language technologists and sociolinguists, and case 3: interaction patterns over time and type of conversation, which extends previous work within interaction analysis. In the user study, we are applying an activity-theoretical approach with the aim of involving researchers, such as me, and investigating how we use – and would like to be able to use – these archival speech resources.

>

[1] http://www.sprakochfolkminnen.se/download/18.46a737b116a496e255833f9/1556021072709/Domeij pres.pdf

Could you describe your research in collaboration with the K-Centre? Have there been any prominent results from this inter-disciplinary collaboration?

<

I am directly involved in case 1: from personal experience narratives to cultural heritage, where I mostly work with a collection of Swedish folklore that was created by Karl Gösta Gilstring, a clergyman and high school teacher who lived in Sweden between 1915 and 1986. Gilstring worked on his collection for more than fifty years, and the result is regarded as the largest folklore collection assembled by a single Nordic researcher in modern times. It consists of more than 8,000 original letters, as well as 250 hours of recordings (mainly interviews conducted by Gilstring himself), from which Gilstring made 70,000 folklore records, divided into approximately one hundred parish collections and organized by subject matter, which aside from folk tales also includes descriptions of rural daily life and traditions.

In our case study, I am interested in establishing the motivations and scientific premises that Gilstring used to create his collection of folktales and to investigate the reasons as to why it has become an integral part of the cultural heritage of Sweden. In the TillTal project, I explore the differences between the unedited audio interviews and his edited written versions that later appeared in the collection. A prominent finding in this respect is that when Gilstring wrote down the folktales he had collected from letters and by conducting oral interviews, he sometimes omitted parts of the story that he felt were his informant's modern interpretations and not part of a "traditional" incarnation of the folktale. This goes to show that cultural heritage is socially constructed, in that Gilstring's rather conservative attitude, which involved a rejection of modern ideals, directly influenced the content of what we nowadays perceive as our "traditional" folklore in Sweden.

The collection is also valuable because of the geographic distribution of the materials. Gilstring's approximately 700 informants not only came from all over Sweden, but also from the Åland Islands and Finland, while around 60 of them were Swedes who had emigrated to America. This is important with respect to the map-based interface Digitalt kulturarv, which the SWELANG K-Centre is developing, since the interface allows me to trace the geographic origins of the letters that were sent to Gilstring by his informants. For instance, I have been able to observe – on the basis of the geocoded information specifying the location of an informant at the time he or she sent the letter – that after emigrating Swedish Americans typically did not stay at a particular place in North America for a long time, but rather moved all over the country, and sometimes even came back to Sweden for a time. Additionally, it was possible for me to observe that the emigrants often presented Sweden in a romanticized manner in their letters to Gilstring, painting the country in broader strokes in comparison to the descriptions in the letters by their compatriots who never left Sweden. This highlights the fact that the ways in which people perceive and remember a particular place (Sweden in the case of the Swedish Americans) are always socially and culturally constructed, and shaped by the individual who reports them.

>

What are the main obstacles of working with audio data? How does the K-Centre help you overcome them?

<

Just recently, I was conducting research on an audio recording that was made with one of Gilstring's informants – a Swedish American called Carl Nelson, who came to America in 1896 when he was 18 years old. What's interesting about the interview is that in certain parts Nelson repeats the same folk stories that he had already described to Gilstring in their previous written correspondences. Additionally, Nelson often jumps from one story to another and then later on returns to comment on a story he's already told. Aside from Nelson's rather messy narration with frequent digressions, the recording is 10 hours long in total, so it took me weeks to go through it. This shows that it is time consuming to analyse audio recordings, so it is incredibly important for me that TillTal gives me the opportunity to collaborate with language technologists like SWELANG's Gunnar Eriksson, who provides me with guidance on the use of automatic speech-analysis methods with which I am able to go back and forth between the different segments of a long audio interview in a time-efficient manner and to interlink them with other related materials in different formats and secondary sources. Indeed, one of the plans of SWELANG is to make available to the research community an environment in which various kinds of materials (e.g., audio recordings, written letters) can be combined so that, for example, dynamic links can be made from a recorded interview to a letter where the same subject or narrative is mentioned twice.

>

As a qualitative researcher, do you think there's any room for improvement in the way data is presented and made available by large-scale research infrastructures?

<

I often feel as though the various domain-specific resources (e.g., historical corpora) available through the repositories are mostly intended for large-scale projects that deal with quantitative "big-data" questions, but it isn't obvious to me how they are suitable for qualitative research. The problem is that many resources contain metadata describing only surface-level features, such as size and linguistic annotation, but lack metadata that are specific to the needs of my field, such as detailed descriptions of the collection process itself, information on who the contributors were in the case of folklore resources, where they came from, when they lived, and so on.

Nowadays, it is easier to get grant money if you propose a humanities project that will – aside from solving research questions that are intrinsic to the field – also involve digitization and collaboration with researchers working in computational fields. While I of course agree that it's extremely valuable to make the data that you're working on accessible in online environments through such collaborations, it often feels as though only the quantity of the data is seen as a measure of success, rather than the presentation of the content of the materials themselves.

I therefore think that it's important for such collaborative projects to re-focus, at least in part, on improving access to and the presentation of the resources that are already available, which is precisely what we are doing in the TillTal project by creating a user-friendly environment for the speech analysis of audio data where the presentation and accessibility of the recordings is tailored to the needs of researchers outside computational fields, like myself.

>

What are the future goals of the TillTal project and the SWELANG K-Centre?

<

One of the future aims of the TillTal project – and by extension the K-Centre – is to increase the amount of available content and bring together related materials (recordings, reports of recordings etc.) through digital methods, which will be done in collaboration with the National Language Bank and SWE-CLARIN. We also plan to release a search system tailored specifically to working with recorded interviews. The system will be accompanied by a tool that will enable us to explore other related non-audio materials while listening to the recording. With this tool, we'll also be able to add additional information about a recording on the fly, such as laughter, or mark sections with fast or otherwise intensive dialogue.

We also plan to develop crowdsourcing tools for transcription and improvement of archive materials, and further work on the mapping interface Digitalt kulturarv, with which researchers will be able to follow audio recordings through time and place, and thereby efficiently study all the documents that were created along the way. In the long run, the plan is to integrate these different technologies in a rich digital tool box, which will offer new possibilities to work with the archival materials of the Institute for Language and Folklore.

>

The National Language Bank and SWE-CLARIN is funded by the Swedish Research Council (2017-00626).

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

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

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

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 2019