Text as an entryway to speech

– a journey into the most inaccessible areas of the archives

Rickard Domeij, Gunnar Eriksson, Eva Lindström, Erik Magnusson Petzell, Susanne Nylund Skog, Fredrik Skott, Jenny Öqvist

Institute for Language and Folklore, Sweden

Abstract. In our poster we account for our first experiences and results from the *Tilltal* project, exploring how different types of text material from the archives can be used to find relevant sections of the recordings, as a complement to speech technological methods..

Keywords: archives, recordings, speech technology, found data.

1 Introduction

Tilltal (*Tillgängligt kulturarv för forskning i tal*, 'Accessible cultural heritage for speech research') is a multidisciplinary and methodological project undertaken by the Institute of Language and Folklore (Isof), KTH Royal Institute of Technology (KTH), and The Swedish National Archives (RA).

The *Tilltal* project is applying speech technology methods to archival material, with the overall goal of making speech archives more accessible to humanities and social science (HS) research. There are immense amounts of recorded interviews which currently have to be played in real time in order to be analysed. With digital tools, we see possibilities to explore the recordings in new ways. (Berg et al. 2016.)

Tilltal comprises three case studies and one user study. In the case studies, three research agendas from different fields (ethnology, sociolinguistics and interaction analysis), aimed at different types of speech analysis, are being pursued (Malisz et al. 2017).

In the user study we are applying an activity-theoretical approach with the aim of involving researchers, and investigating how they use – and would like to be able to use – the archival resources at Isof. Taking into account the researchers' needs, digital solutions will be suggested and their usefulness and applicability will be judged in practice. (Berg et al. 2017.)

The archives at Isof also contain a wide range of information in written form, including descriptions of recording situations and manual transcripts. These written data sources provide further possible pathways into the speech materials.

In our poster we account for our first experiences and results, exploring how different types of text material from the archives can be used to find relevant sections of the recordings, as a complement to speech technological methods. So far, our focus

has been one of the case studies, which explores how personal experience narratives are transformed into cultural heritage (Nylund Skog 2018).

2 Experiences and results from our explorations of the archives

Initially, we examined different categories of written material in order to find entry points to the recordings in the folklore collections, for example: various kinds of records, letters and transcriptions. Yet, even though we were prepared for the fact that gaining access to material would be time consuming, we had underestimated the complexity. The metadata available about the recordings at Isof is very heterogeneous both in both form and content. Finding information about transcriptions and reports of recordings, for example, can be surprisingly complicated. In some cases we even found that after having generated a large number of written documents, the actual recording was discarded without comment.

Owing to the problems with finding recordings through written material, we proceeded to instead use recordings as starting points, and to look for written materials associated with them. Work is currently in progress to create ”bunches” of such material, and by means of different methods, to link the texts within these materials to the timeline of the recording.

In this way, the written materials belonging to a certain recording are collected in a digitally accessible bunch. Such bunches are valuable tools for the case study within Tilltal which explores the transformation from personal experience narratives into cultural heritage, but also for completely different research where archive materials are used.

By bringing together the different material categories of the archive and making them available as digital bunches, the archive’s collections are, on the whole, also more transparent and accessible (c.f. Prescott & Hughes 2018). From a user study perspective, the work on bunches brings new approaches for supporting researchers’ analysis work through digital methods.

3 Suggested digital solutions

The aim of our work is to develop prototypes for, on the one hand, a search system which directs the researcher straight into relevant bits of a recorded interview, and, on the other hand, a tool which enables the researcher to explore the other materials while listening. Furthermore, it will be made possible to add other information about a recording, such as laughter, or marking sections with fast or otherwise intensive exchange. We have begun outlining what this might look like, using annotation tools such as ELAN.

During 2019, we will also further develop the system *Digitalt kulturarv* (Digital cultural heritage; Dagsson & Skott 2018). This system was designed for the purpose of digitising folklore records from Isof’s collections and making them accessible to

the public. In the next step, we focus specifically on researchers. Our aim is to increase the content and bring together related materials (recordings, reports of recordings etc.) through digital methods, making it possible to follow the recording trips through time and place, as well as acquaint oneself with all the documents created along the way. This is done in collaboration with the recently established *Nationella Språkbanken* (National Language Bank) and Swe-Clarin (Borin et al. 2018). We also plan to develop crowdsourcing tools for transcription and improvement of archive material.

Our plan in the long run is to integrate these different parts to a complex digital tool box, through which we can offer researchers new possibilities to work with Isof's archival materials.

4 Closing words

By focusing in depth on archive materials for the purpose of making them more accessible, we have discovered a number of difficulties that we did not foresee, and that would otherwise not have been known. This is important, since an increased use of the archive resources of folklore institutions – especially the many unexplored speech recordings – has great potential gains for research in the humanities and social sciences. Not least, we want to make it easier for researchers to work with the actual recordings, and not only with transcriptions of them. As we proceed with our plans and prototypes, we would be very grateful for comments and suggestions from researchers and others with similar projects and experiences.

5 Acknowledgement

The *Tilltal* project is financed by the Royal Swedish Academy of Letters, History and Antiquities and the Swedish Foundation for Humanities and Social Sciences (Riksbankens Jubuleumsfond; SAF16-0917:1). The projects *Språkbanken* and *Swe-Clarin* are financed by The Swedish Research Council (Vetenskapsrådet; 2017-00626).

References

1. Borin, Lars, Forsberg, Markus, Edlund, Jens & Domeij, Rickard. 2018. Språkbanken 2018: Research Resources for Text, Speech, & Society. Poster DHN I: Mäkelä, Eetu, Tolonen, Mikko & Tuominen, Jouni (eds.) Digital Humanities in the Nordic Countries 3rd Conference. 504–506. Retrieved from: http://ceur-ws.org/Vol-2084/poster7.pdf
2. Berg, Johanna, Domeij, Rickard, Edlund, Jens, Eriksson, Gunnar, House, David, Malisz, Zofia, Nylund Skog, Susanne & Öqvist, Jenny. 2016. Tilltal – making cultural heritage accessible for speech research. Paper presented at CLARIN Annual Conference 26–28 October 2016, Aix-en-Provence, France.
3. Berg, Johanna, Domeij, Rickard, Edlund, Jens, Eriksson, Gunnar, House, David, Malisz, Zofia, Nylund Skog, Susanne & Öqvist, Jenny. 2017. Involving users and collaborating be-

tween disciplines in making cultural heritage accessible for research. Paper presented at CLARIN Annual Conference 18–20 september 2017, Budapest, Hungary.

4. Dagsson, Trausti & Skott, Fredrik. 2018. Digitalt Kulturarv – ett digitalt folkarkiv [Blog post]. Retrieved from: https://sweclarin.se/swe/digitalt-kulturarv-–-ett-digitalt-folkminnesarkiv
5. Malisz, Zofia, Öqvist, Jenny, Fallgren, Per, Edlund, Jens & House, David. 2017. Visualising vocalic variability in space and time – automatic exploration of "found data". Paper presenterat vid 47th Poznań Linguistic Meeting, 18–20 September 2017, Adam Mickiewicz University, Poznań, Polen.
6. Nylund Skog, Susanne 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.* Lauri Harvilahti, Audun Kjus, Cliona O'Carroll, Susanne Österlund-Pötzsch, Fredrik Skott and Rita Treija (eds.) Helsinki: Academia Scientiarum Fennica, FFC 315.
7. Prescott, Andrew & Hughes, Lorna. 2018. Why do we digitize? The case for slow digitization. In: Archive journal (Special issue: Digital Medieval Manuscript Cultures, edited by Michael Hanrahan & Bridget Whearty, September 2018.) Retrieved from: http://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/