Lindén, Krister; Jauhiainen, Tommi; Lennes, Mietta; Kurimo, Mikko; Rossi, Aleksi; Kurki, Tommi; Pitkänen, Olli

Donate Speech: Collecting and Sharing a Large-Scale Speech Database for Social Sciences, Humanities and Artificial Intelligence Research and Innovation

Published in:
CLARIN : the infrastructure for language resources

DOI:
10.1515/9783110767377-019

Published: 01/10/2022

Document Version
Publisher's PDF, also known as Version of record

Krister Lindén, Tommi Jauhiainen, Mietta Lennes, Mikko Kurimo, Aleksi Rossi, Tommi Kurki, and Olli Pitkänen

Donate Speech

Collecting and Sharing a Large-Scale Speech Database for Social Sciences, Humanities and Artificial Intelligence Research and Innovation

Abstract: The Donate Speech campaign aimed to collect 10,000 hours of ordinary, casual Finnish speech to be used for studying language as well as for developing technology and services that can be readily used in the languages spoken in Finland. In this project, particular attention has been devoted to allowing for both academic and commercial use of the material. Even though this ambitious target currently seems likely to evade us, the Donate Speech campaign has managed to amass an extensive resource of more than 4,000 hours of Finnish colloquial speech comprising more than 220,000 speech recordings by more than 25,000 speakers from all over Finland in just a few months.

Keywords: speech resources, colloquial speech, large-scale data collection, academic and commercial use

Acknowledgements: We are grateful to the Vake Oy (currently Ilmastorahasto) for funding the initial survey and the development of the software platform for speech collection, to the national Finnish Broadcasting Company Yle for developing and advertising the collection campaign in national media, as well as to the Academy of Finland for funding the transcription of a substantial part of the speech data and the development of the framework to distribute the data through FIN-CLARIN and the Language Bank of Finland. In addition, we are grateful to researchers of the Aalto University, the University of Helsinki, and the University of Turku, as well as the staff of Yle and Vake for generously contributing their time to the project.

Krister Lindén, University of Helsinki, Helsinki, Finland, e-mail: krister.linden@helsinki.fi
Tommi Jauhiainen, University of Helsinki, Helsinki, Finland, e-mail: tommi.jauhiainen@helsinki.fi
Mietta Lennes, University of Helsinki, Helsinki, Finland, e-mail: mietta.lennes@helsinki.fi
Mikko Kurimo, Aalto University, Aalto, Finland, e-mail: mikko.kurimo@aalto.fi
Aleksi Rossi, Yle – Finnish Broadcasting Company, Yleisradio, Finland, e-mail: aleksi.rossi@yle.fi
Tommi Kurki, University of Turku, Turku, Finland, e-mail: tommi.kurki@utu.fi
Olli Pitkänen, 1001 Lakes Oy, Helsinki, Finland, e-mail: olli.pitkanen@1001lakes.com

https://doi.org/10.1515/9783110767377-019

1 Introduction

There are already several commercial systems utilizing AI with Finnish speech recognition in production use, but many more use cases are waiting to be successfully commercialized. To some extent this may be due to the fact that the demand for and supply of language resources do not always align, but the consensus of opinion among experts is that openly available large language resources will further accelerate the development and implementation of various language-based AI applications. Openly available speech processing components make it possible for many different actors wishing to test service ideas to pilot high-level services, while leaving the final decision on what technology to use in the production phase to a later stage. For example, automatic speech recognition (speech-to-text) and speech synthesis (text-to-speech) in Finnish have been available on a few devices and applications for several years (e.g., as speech capabilities in Apple and Google products), but many end-user services require better and more reliable processing support for colloquial Finnish.

A worldwide effort by the Mozilla Common Voice project[1] is ongoing, but their aim is to collect speech that has been read aloud. From previous projects, we know that prompted speech tends to bring people to use standardized and non-colloquial speech, and we specifically wanted everyday spontaneous speech from a large number of speakers.

In the remainder of Section 1, we will describe the process that led to the point where Vake, the Finnish State Development company (currently Ilmastorahasto Oy) was able to make the decision to fund the speech data collection campaign. We also offer a glimpse of the history of the Language Bank of Finland to explain why it was chosen as the distributor of the data, what speech material had previously been collected, why we still decided that we needed to collect new speech material for modern colloquial speech, and how CLARIN has prepared for the distribution of such large personal data collections.

The remainder of this chapter is structured as follows: in Section 2, we get an overview of a similar project (with a purely academic goal) which gave us valuable previous experience. In Section 3, we learn how the Finnish national broadcasting company Yle designed the media campaign to get people to donate speech. In Section 4, we describe the legal framework for collecting the data so that it can be reused by academia and industry alike. In Section 5, we take a look at the technical implementation and where to find the software for the speech collection platform. In Section 6, we overview the data we were able to collect,

1 https://commonvoice.mozilla.org/

and in Sections 7 and 8, we draw some conclusions and acknowledge the funders and the organizations who contributed to the implementation and running of the campaign.

1.1 The need for speech corpora

At the beginning of the 21st century, the efforts and resources of Finnish speech technology and spoken language research were scattered all over Finland and represented by relatively small teams. The USIX – Uusi käyttäjäkeskeinen tietotekniikka [New User-Centric Information Technology] technology programme was launched in 1999 and funded by the Finnish Technology Agency (Tekes, currently Business Finland). The programme, resulting in new projects and cooperation between research teams, boosted research in Finnish speech and language technology. With funding from the Ministry of Education, a survey on the state of the art of speech and spoken language research in Finland was published in 2001 (Toivanen and Miettinen 2001). One of the key findings of the survey was that investments in the availability of digital speech data were required to boost the development of research and technology in Finnish speech processing.

The availability of speech data is a prerequisite for both research in spoken language and the development of speech technological applications, including speech interfaces. The consortium project *Integrated Resources for Speech Technology and Spoken Language Research in Finland (SA-Puhe)*, funded by the Academy of Finland in 2003–2004, aimed to tackle the need for general guidelines and methods for researchers to collaboratively collect, annotate, and share speech corpora. During the project, phoneticians and language researchers at the University of Helsinki worked together with the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology and CSC – IT Center for Science.

The SA-Puhe project made a big effort to address the need for a centralized infrastructure in storing, sharing, and maintaining both speech data and the related annotations for research purposes. The platform was to be built on an object-oriented database system called QuickSig, which had been developed at the Helsinki University of Technology, including some further collaboration with the University of Helsinki, during the 1990s (Karjalainen and Altosaar 1993; Altosaar, Millar, and Vainio 1999). The database system was to provide efficient queries and access via a graphical query formation compiler (Altosaar and Lennes 2005). In order for researchers to be able to contribute, share, and maintain their transcripts and structured annotations for the speech recordings, the

first version of a collaborative annotation editor (Puh-Editor) was developed at CSC – IT Center for Science (Grönroos and Miettinen 2004).

Unfortunately, it was not possible to complete the integration of the components of the speech database platform during the funding period. Due to the lack of resources for further development and maintenance, the Puh-Editor software was discontinued after a couple of years in test use, and the database system remained a local development project. During the project, general speech annotation guidelines were developed with the help of the language researcher community (Lennes and Ahjoniemi 2005). These guidelines proved to be useful when the idea of big data for speech processing was revived inspired by recent progress in speech technology due to neural network technology.

The process that led to the launch of the *Donate Speech* campaign began with the meetings of an ad hoc group of companies and public organizations during 2018. In spring 2019, Vake commissioned a preliminary study for the needs of Finnish language resources for artificial intelligence from FIN-CLARIN and the Language Bank of Finland (Kielipankki).[2] The goal was to specify interventions that would enable wide usability of the languages spoken in Finland in various AI applications, beginning with Finnish as the most widely spoken language in Finland. The Language Bank collected opinions and conducted interviews with more than 50 commercial and public organizations in Finland. One of the eight identified development targets was a large corpus of spontaneous colloquial speech, as identified in the study published in October 2019.[3]

FIN-CLARIN, through the Language Bank of Finland, cooperated with the Finnish Broadcasting Company (Yle) and the Finnish State Development Company (Vake Oy, currently Ilmastorahasto Oy) in the Donate Speech campaign (Lahjoita puhetta). Experts from the University of Helsinki, Aalto University, and the University of Turku also participated in the project. Vake assigned the data protection analysis and the drafting of legal documents to 1001 Lakes Oy, and legal counsels from the University of Helsinki and from Yle participated in developing the legal framework of the collection campaign.

2 The FIN-CLARIN consortium (www.helsinki.fi/finclarin) is led by the University of Helsinki and the main service centre of FIN-CLARIN is the Language Bank (www.kielipankki.fi).

3 https://vake.fi/wp-content/uploads/Vaken-suomenkielisen-tekoälyn-kehittämisohjelma-Esiselvitys-2019.pdf

1.2 FIN-CLARIN and the Language Bank of Finland

Since 2009, FIN-CLARIN has been on the national research infrastructures roadmap maintained by the Academy of Finland. The FIN-CLARIN consortium consists of all Finnish universities engaged in linguistic and language technology research,[4] the Institute for the Languages of Finland (Kotus),[5] and CSC – IT Center for Science.[6] FIN-CLARIN maintains the Language Bank of Finland,[7] through which the members of the consortium make available various language resources, both corpora and tools.

From the beginnings of the Language Bank in 1996, the aim has been that both corpora and tools are made available to the research community in the most efficient way possible. Because little attention has been paid to making materials and tools available to companies, many language resources are licensed specifically with a non-commercial restriction. In many cases, copyright or data protection issues have also led to restricted licenses. In FIN-CLARIN, CSC is responsible for the technical maintenance and the University of Helsinki for the acquisition and curating of corpora and tools.

1.3 Potential applications for special needs

Searching speech recordings for content is error-prone, even if word-spotting techniques are available for locating likely speech segments. Another approach is to convert speech into textual transcripts and use existing tools for text analysis. One may wish to count how many recorded telephone calls mention certain issues in a robocall survey. Examples of more complex use cases are various analyses of telephone discussions and their post-processing solutions. Another application is the automatic transliteration of interviews conducted by journalists or researchers. Quickly finding a quote from the speech signal would considerably speed up the verification of the details of such interviews. Improving the searchability of speech recordings also improves the usability of video-recorded debates for later verification, for example, the debates associated with decisions made in the plenary of the Parliament.

4 The Aalto University and the universities of Eastern Finland, Helsinki, Jyväskylä, Oulu, Tampere, Turku, and Vaasa.
5 https://www.kotus.fi
6 https://www.csc.fi
7 https://www.kielipankki.fi/language-bank/

Automatic speech recognition (ASR) is frequently needed and used for traditional text dictation, for instance for drafting messages in situations where hands and eyes have other duties. Dictation that adapts to the speech of a single person already works reasonably well in Finnish, for example on mobile devices, especially in conditions where the amount of background noise is low and/or the speaker is close to the microphone.

With improved speech processing, television shows, lectures, and so on can be subtitled automatically, either directly from the original audio or from the dictation of a human subtitler. Special groups such as the hard of hearing would benefit greatly from near-real-time subtitling of speech. Reliably functioning, genre-independent subtitling of Finnish speech would also provide a basis for automatic translation and interpretation, which has innumerable uses in the globalizing world.

Society currently requires a number of digital user skills, such as the utilization of mobile devices. If a user's vision is impaired or their finger dexterity is insufficient for a device, a user may currently be excluded from many services. Often, however, these requirements can be bypassed with a voice-enabled user interface to services in the user's native language. For the elderly and disabled, intelligent applications may complement or even replace personal services and provide an opportunity to live independently while improving the quality of life. On the other hand, if a voice interface exists but works poorly, it creates distrust and the users may avoid using a service. In some cases, such as healthcare services, user interface deficiencies may also pose security risks.

In language learning applications, speech interfaces that adapt to specific users are more useful. Interactional and oral skills are often emphasized in today's society and working life, and they are becoming an increasingly important part of language learning. For immigrants in Finland, having good oral skills in Finnish can be a great advantage in the job market and in building their social networks. A large database of transcribed colloquial speech with known topics is a good reference point, but other types of data are also needed to reliably measure pronunciation features in the speech of individual language learners and to model their speech and communicative activities in real interactional situations.

There are use cases where the speech to be analysed does not need to be presented in textual form but the analysis can be inferred directly from the speech. Such functionalities are, for example, automatic speaker identification or the automated analysis of a user's age, state of alertness, or health. The latter are useful for customizing applications and various services provided to the user, even if the accuracy is less than 100%. Even when such applications do not require the speech to be presented in textual form, they require large training corpora of speech data annotated with personal and health-related features.

1.4 Speech data for commercial use

Transcribing speech into text is a subjective process. A transcript is produced for a particular purpose and it reflects the choices made by an individual annotator. Regardless of the selected transcription system, a written transcript is unable to reflect all features that are relevant to natural interaction and the meaning of speech. These include momentary variations in the production of speech sounds or other noises, as well as longer-term prosodic properties, for example, voice quality, pitch, intensity, speech rate, and pauses. These features contribute not only to the impressions of melody, accents, and rhythm but also to the perceived meanings, intentions, and attitudes that we hear and understand in each other's speech as well as gestures, expressions, gazes, and other activities related to the interaction situation and context. The primary objective for the transcription of the collected Donate Speech data is to provide a phonematically accurate transcription of the sounds in the signal that will later be mapped to standardized speech for searchability and for enabling further language processing research and development.

The construction of secure, privacy-friendly voice user interfaces may in some cases require that the components of an application can be used without the transfer of personal data from one service to another, to a third party, or to another state. These factors argue for the fact that the speech processing components should be openly accessible and open source.

Speech corpora previously distributed by the Language Bank of Finland, such as the "Plenary Sessions of the Parliament of Finland, Downloadable Version 1" containing recordings of Parliamentary Plenary Sessions from 10 September 2008 to 1 July 2016, as well as their transcripts, are licensed CC-BY-NC-ND. Here NC is an abbreviation for *non-commercial*, that is, the materials may not be used for commercial purposes. Renegotiating licenses for this and other similar corpora to allow business use is another way to add commercially usable speech material. While in the case of the Plenary Sessions of the Parliament it may still be possible, it is often not feasible to renegotiate access rights to speech material after it has been collected and licensed. For this reason, it was particularly important to make sure new speech material was collected in a targeted manner, specifically including the possibility of commercial use.

1.5 Legal considerations for sharing data within CLARIN

The legal framework in the EU is intended to provide an interoperable space for various activities. While the legal framework harmonizes many of the activities

in other parts of society, the research arena has at times been left for national consideration. This affects the sharing of research data and resources that can be achieved through a research infrastructure like CLARIN as we need to find common legal ground that is applicable to research in all EU countries. In addition, research is not only limited to academia, so to share resources within a country, we often need solutions that apply to industry as well.

The intellectual property aspect of the legal space has been extensively discussed in (Kelli, Lindén, Vider 2016; Kelli, Mets, Vider, et al. 2018; Kelli, Tavast, Lindén, et al. 2019) by members of the CLARIN Legal Issues Committee. CLARIN recommends using Creative Common licenses whenever possible (Oksanen and Lindén 2011). For all datasets, including those that cannot be made openly available, CLARIN offers a legal metadata classification system (Oksanen, Lindén, and Westerlund 2010) to inform the users of potential restrictions that they need to be aware of when accessing a dataset. For datasets that cannot be made openly and publicly available, CLARIN also offers standard license templates for depositing data to be shared through CLARIN Centres (Kelli, Lindén, Vider, et al. 2018). The IPR relevant for sharing research data has been extensively scrutinized by CLARIN over the last ten years, which is documented in Kamocki, Kelli, and Lindén (2022) Section 3.5 of this book, and we are eagerly awaiting new opportunities provided by the EU text and data mining directive (Kelli, Tavast, Lindén, Vider, et al. 2020).

During the last few years, the consequences of EU's General Data Protection Regulation (GDPR) has been widely recognized (Kelli et al. 2021). Some leeway was given to individual EU member countries to implement exceptions for research, and this has led to differing practices for sharing personal data for academic research purposes (Kelli, Lindén, Vider, et al. 2019; Lindén et al. 2020). Resources containing personal data are among the resources that cannot be made available without protective measures, and CLARIN is in the process of updating its license templates to reflect how personal data can still be shared in safe and controlled ways for academic research (Kelli, Lindén, Vider, et al. 2020).

Despite the fact that not all data can be made openly accessible, it is possible to use data to which one has legal access for creating openly accessible language models. For a detailed discussion of this, see Kelli, Tavast, Lindén, Bristonas, et al. (2020). To illustrate how personal data can be collected and shared within the EU, we will present the legal underpinnings of the Donate Speech campaign in Section 4. The campaign involved more than 25,000 citizens in Finland donating more than 220,000 speech samples comprising roughly 4,000 hours of colloquial speech to be used by academia and industry for developing and researching language and AI applications. The fact that the data was collected to be used by industry as well makes it particularly relevant for CLARIN as industry use is regulated by the common EU ground in the GDPR.

2 Earlier speech collections in Finland

In Finland, there are several extensive speech databases previously collected for linguistic research by the Institute for the Languages of Finland, the universities, and memory organizations, but for commercial purposes access to them is limited. In addition, plenty of linguistic research has been done, over a long period, from the perspective of dialectology, sociolinguistics, and interactional linguistics, and there are exceptionally extensive dialectological corpora (most of them representing the regional variation of Finnish in the 1960s and 1970s) and large sociolinguistic corpora representing social variation on the segmental levels of Finnish. However, a new extensive speech database representing large-scale regional and social variation in contemporary Finnish is potentially a valuable new asset also in linguistics. Collecting dialectological and sociolinguistic speech data has typically been done through fieldwork and face-to-face interaction. Due to this aspect, compiling such a database has typically required vast resources of time and funding.

The Donate Speech Campaign is associated, on the one hand, with dialectology and sociolinguistics and their long traditions in obtaining data by doing extensive fieldwork, and on the other hand with phonetics and speech technology, which obtain data in laboratory settings. Both fields are largely empirical in practice. As dialectology and sociolinguistics aim for naturalness, with a focus on conversational speech and representativeness of speakers within communities, phonetics holds the replicability of experiments in high esteem and focuses on speech in laboratory settings (Thomas 2013: 108). In this project, collecting speech data over the internet needed to strike a balance between the two and at the same time take into account the possibilities and limitations of the digital environment.

Collecting speech data over the internet is a faster and more economic method than traditional fieldwork, and it makes it possible to reach a large number of potential participants who would not necessarily otherwise participate. Meanwhile, several questions arise: how can we collect controlled data with elicited tasks that represent speech as naturally or as spontaneously as possible and cover current regional and social variation as widely as possible? How can we obtain a large database that also represents functionally different speech samples (statements, commands, questions, echo questions, etc.)? A dialectologist or sociolinguist seeks ways to grasp the variation of language, and in practice will inevitably face the observer's paradox as Labov (1972: 209) has phrased it: "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation."

Whether a scholar records an interview or a conversation in which informants are involved or an informant makes a recording alone – in other words, whether the scholar collecting data is present or not – Labov's paradox holds. The same paradox applies to data collection over the internet, especially when the goal of the campaign is not to collect read speech data. When interacting with a computer instead of another human being, how can we overcome the potential distraction that participants are self-consciously aware that they are recorded and, due to this, carefully watch their language use?

2.1 Previous lessons from the Prosovar project

The Donate Speech campaign had a Finnish predecessor that incorporated new methodology and new ways of obtaining speech data over the internet, implementing a crowdsourcing approach. The multidisciplinary project *The Regional and Social Variation of Finnish Prosody* (Prosovar) was conducted by the University of Turku and financed by the Kone foundation (2013–2015; see also Kurki et al. 2014; Nieminen and Kurki 2017). The objectives of this project included (a) the formation of a speech corpus particularly for the study of Finnish prosody and its regional and social variation (The Corpus of Prosodic Variation in Finnish) and (b) the development and testing of a method for data collection and analysis for the study of natural spoken language.

As a complement to old fieldwork for obtaining speech in dialectology and sociolinguistics, a new, partially crowdsourced method for collecting sociolinguistic and sociophonetic data via the internet was developed and tested in the Prosovar project. There was also a precedent for collecting sociolinguistic data on the internet (in particular, Dialect Topography by Professor J. K. Chambers; cf. Chambers 1994), but to our knowledge, Prosovar was one of the first attempts in dialectology, sociolinguistics, and sociophonetics (cf. computational linguistics; e.g., Lane et al. 2010; McGraw 2013) to collect speech data over the internet. The development of data collection methods in Prosovar required a multidisciplinary approach, where dialectological, sociolinguistic, (socio)phonetic, computer science, and Finnish language expertise was needed.

The idea was to motivate non-linguists to participate in data collection by completing recording tasks with a web application created for the Prosovar project. From the beginning of the project, it was crucially important to find ways to attract voluntary participants willing to record their speech samples for linguistic research purposes. The goal of giving public presentations, interviews to newspapers, and campaigning in social media was to arouse public interest. The possibility of listening to anonymous speech samples from other participants and

implementing the elements of a game-like design in developing the application were also found to be good ways to attract interest.

Participants were able to make recordings with their personal computers, (Android) tablet computers, and (Android) cellular phones, as long as their device had a microphone and they created a user account. At the same time, this presented a way for them to further participate in the research; as long as they made recordings for the database, they were allowed to listen to randomly selected anonymous voice clips from the database and evaluate them in a folk linguistic manner. For example, a participant was asked to listen to a clip and locate the speaker's dialect on a map, or he/she was asked to describe, using a few adjectives, what the speaker in a clip sounded like. This information was and is possible to investigate from a folk linguistic perspective by analysing the language with regard to respondents and from a computer science perspective by applying dialect recognition techniques (e.g., how humans and computers perceive sounds differently).

Unregistered guest users were only able to listen to a few selected anonymous samples and obtain general information about Finnish colloquial speech and dialect samples in the data obtained so far in the project. In order to access the recording tasks and the "game", in which one listened to short audio clips and tried to locate their speakers, one had to (1) create a user account, (2) accept the conditions and terms of use, and (3) finish at least one recording task in order to access the game. All the data and the background information about the participants were moved to a separate server for privacy and security reasons. By the end of November 2015, there were approximately 1,000 registered users, of whom 395 had made recordings for the project, producing a total of over 9,300 recorded samples.

Inventing and designing suitable elicitation tasks was of crucial importance to the Prosovar project (see also Nieminen and Kurki 2015; Nieminen and Kurki 2017). The objective was to obtain comparable utterances, that is, the same thing in different dialects. In the very first tasks designed for the pilot stage, the participants were just prompted to read out loud the text on a screen; consecutive sentences of the same paragraph one by one, or simply disjointed phrases without any further context. Soon, it became clear that this might lead participants to use standard Finnish and thus obfuscate regional and social variation. The shorter the task and the more time for a participant to react, the more likely it was – at least for some informants – to become notably aware of their own language use; this was not ideal, since the idea was to collect spontaneous verbal reactions and not performances consciously planned to be recorded.

Due to this, tweaks were made to the old tasks and new tasks were designed. For instance, the same phrase was shown in two or three distinct dialects at the

same time on a screen and participants were asked to consider how they would express the same phrase in their own way. In another task, a participant was told to list months and weekdays. In addition, tasks with visual, auditory, or audio-visual stimuli were devised. It still seemed that in tasks with textual stimuli, participants paid close attention to their language. Especially if the time to react to a stimulus was unlimited, some participants consciously paid attention to their language use, and as a consequence tended to either exaggerate dialectal forms or strive for perfection speaking in a very standard Finnish manner.

In the end, it was best to have various tasks with different stimuli; in most cases, the task instructions or cues were kept out of the way as much as possible while ensuring decent predictability in what the informant would ultimately say. Thus, the participants were required to react to assignments of various kinds. For instance, there was a task where the participants were shown two pictures with minor differences; their task was to spot the differences and report them verbally. In another task, the participant was shown a map of a fictional town and asked to guide a stranger from one point to another. In some tasks, participants were instructed to speak simultaneously when they saw a stimulus or when they were watching it. For instance, participants were shown a short animated video and, instead of watching it first and then summarizing the plot of the video, they were asked to describe and explain what was happening in the video.

Obtaining functionally different speech samples was one of the most challenging parts in creating the Prosovar database. It was much easier to develop tasks that reached narratives and even declaratives than interrogatives. Some sound samples of tasks in which a participant was asking questions without an actual interlocutor in the scene appeared awkward or unnatural. To mitigate this, tasks were created in which the research group tried to create an illusion of mutual interaction. For instance, there was a setting for social interaction in a marketplace where the participant was instructed to either buy berries from a salesman or to sell berries to a client, while the other party's line was provided by a pre-recorded sound sample on the site. This solution helped to construct a substantially more vivid setting; inevitably, however, it was only slightly reminiscent of actual human interaction. A more functional solution would presumably have been to have two or more participants online simultaneously in the same recording task. In addition to the limited technical resources at the time in the Prosovar project, as well as the risk of data abuse or pestering of other participants prevented the implementation of this collaborative type of task.

Previous experiences in collecting speech data over the internet also showed that sound quality has to be taken into consideration. In the Prosovar project, there was a need to find a balance between catching as many potential participants as possible and setting the system requirements for the devices of potential

participants. Overly sophisticated system requirements would have decreased the number of potential users. For the same reason, it was decided that collecting data would be carried out without asking the potential participants to install any application on their device. Because of this, the minimum requirement for a device was basically a microphone (built-in or external). Since the recordings were fully carried out by the participants, it was not possible to control the recording settings. Participants had a varying range of computers with varying quality of sound equipment. The website provided information on how to use the mixer and how to ensure eligible recording conditions, but few participants seem to have made use of them.

This tended to leave the Prosovar research group at the mercy of the web browsers and their plug-ins and add-ons. As a consequence, the quality of speech data was very variable. Still, the majority of the samples were actually of good enough quality as the objective in Prosovar was to study the prosodic features of speech, which are generally more robust than the spectral features.

3 Designing the Donate Speech campaign

This section describes the process that was used to design and launch the Donate Speech campaign. The initial objective of the Donate Speech campaign was to collect data for all languages spoken in Finland. However, the first phase of the campaign focused on Finnish with the objective to obtain 10,000 hours of colloquial Finnish representing the wide variety of ways the Finns currently speak it in everyday settings. The data is intended for linguistic research and development of technology for both academic and commercial purposes. We also describe what kind of meta-information was collected from the participants and how.

The goal of the campaign was not merely to collect a vast amount of any kind of speech, but to reach out to as many different groups of Finnish speakers and to as many individuals as possible. In marketing the campaign to citizens, it was emphasized that all variants of spoken Finnish are welcome, including speech from second-language Finnish learners. However, in order to understand the privacy notice and the instructions, a certain level of language proficiency was required from the speech donors.

In order to strike a balance between the material goals, the technical possibilities, and the resources that were available, design workshops were organized for all interested parties. During these events, general ideas were collected from both industry and academia on the different uses for the collected speech, while most of the planning of the thematic tasks to elicit speech was left to the staff

of the national broadcasting company Yle with advice collected from previous efforts like the Prosovar project. Yle was in charge of the public outreach through its radio and TV channels. Yle designed pictures, videos, and texts that were presented to speakers in the web application and the downloadable apps. A number of technical templates were designed to allow the design of themes with various types of content in order to target a desired speaker group.

The workshops to determine potential use cases, target audiences, and required and optional features were conducted in autumn 2019 with key research stakeholders, following up during spring 2020. The workshops were facilitated by the solution developer company Solita and were loosely based on the Design Thinking methodology. Later a series of key features were also tested with quick paper prototypes, and in succession with semi-interactive tools. A multitude of design suggestions were made by professional service designers guided by their experience, and a few crucial ones were also tested in practice.

Key issues and challenges for the design of the user interface were in determining elicitation methods that entice a person to speak freely, gaining the trust of the speaker, making him feel comfortable while also satisfying legal constraints for presenting enough required information in an easy-to-understand format, as well as more technical choices of supported platforms, presentation forms, visual and auditory feedback of the ongoing recording or its quality. After some ideas for themes had been formulated and tested, Yle settled on the fail-safe recurring functions of showing a video, a picture, or some textual content, enticing a person to speak with a single, easy-to-use button for starting and stopping the recording.

There were a number of discussions about whether and how to introduce gamification elements similar to the ones suggested by the Prosovar project, such as telling the user how much he had donated, or elements like scoreboards to compare results and maintain user interest, or social elements like sharing results or collecting teams. Eventually, only the amount of total time donated was included as a gamification element, leaving room for further improvements.

The opening theme *Harjoitellaan ensin* (Let's practice) started by test-driving the recording with the user, and assuring them that mainly AI researchers would use the recordings and reminding them about the privacy notice. The technical platform also presented metadata questions for the user to answer, for example about dialect background (the location of the phone is neither queried nor transmitted), basic demographics like age group, gender, mother tongue, the current county a person lives in or was born in, and their profession and education level. In addition, the technical platform was also collected for statistical purposes.

In the end, Yle developed around 40 rather straightforward themes for stimulating the collecting of speech data. In addition to the opening practice theme,

the 12 most popular themes, through which almost half of the data was collected, were: *Rakkain eläimeni* (My dearest pet), *Mistä kodikkuus syntyy?* (What makes a cozy home?), *Tärkeä esineeni* (An important object of mine), *Lempivaate* (My favourite piece of clothing), *Mikä suututtaa?* (What's infuriating?), *Turhat tavarani* (My superfluous things), *Mitä opimme?* (What did we learn?), *Entisajan lemmikit* (Old time pets), *Katson ikkunasta* (While I am looking out the window), *Kuva-arvoitus* (Picture riddle), *Kerro aamiaisesta* (What was your breakfast like?).

As part of the campaign, Yle made comical infomercials with requests to the general public to donate speech. These were broadcast during programme breaks in national radio and TV channels during the summer and autumn of 2020, during the Covid-19 pandemic, with some trailing reruns during spring 2021.

4 Legal aspects

From the beginning, it was clear that the processing of data must be conducted in a legally and ethically sound way. All the central actors in the project – Kielipankki at the University of Helsinki, Vake, and Yle – are public organizations that cannot ignore these aspects.

The speech material donated during the campaign will be stored in the Language Bank of Finland (Kielipankki), coordinated by the University of Helsinki. It was noted that the material may contain subject matter protected by several legal rights (Alen-Savikko and Pitkänen 2016), such as:

- data protection rights (Wrigley, Alen-Savikko, and Pitkänen 2019)
- copyright and neighbouring rights (e.g., the right of the producer of a sound recording, database sui generis right) (Pitkänen 2017)
- patents (Ballardini et al. 2013)
- trademarks (Weckström 2012)
- trade secrets (Schröder 2018).

In particular, the personal data protected by European and national data protection legislation, most notably by the General Data Protection Regulation (GDPR),[8] is considered to be essential from the campaign's viewpoint. The definition of the personal data is very broad and therefore significant parts of the speech material can be considered personal data for various reasons:

- metadata about the speaker, his or her identification, name, etc., can be linked directly to a person.

8 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

- the recognizable voice of a speaker may also be linked to a person, at least if there is some other information about the speaker available.
- the content of the speech may include personal information, e.g., if the speaker reveals what he was doing with his friends last weekend.

According to the GDPR, it is important, inter alia:
- to define the purpose of the processing of personal data;
- to inform the data subjects about the processing of personal data in a concise, transparent, intelligible, and easily accessible form, using clear and plain language;
- to define a lawful basis to cover data processing, i.e., consent, contract, legal obligation, vital interest, public interest, or legitimate interest;
- to analyse and mitigate the potential risks of personal data processing to individuals.

These requirements were taken very seriously from the beginning.

The speech material can be shared with individual researchers, universities and research organizations or private companies that need it for studying language or artificial intelligence, for developing AI solutions, or for higher education purposes related to the aforementioned areas. During and after the campaign, the privacy practices of the Language Bank of Finland have been developed in accordance with the GDPR.

According to the GDPR, personal data shall be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes.[9] Therefore, it was essential to define the purpose as clearly as possible. In general, it is very difficult to avoid some vagueness when trying to define forthcoming undertakings. However, the following definition is as accurate and comprehensible as it was possible to come up with: "Personal data is processed for the development and research of applications and services that understand and produce speech, as well as for language research and higher education related to these purposes."

According to GDPR Article 6, the processing of personal data is lawful only if and to the extent that at least one of the lawful bases applies:
- consent
- contract
- legal obligation
- vital interest

9 GDPR Article 5(1)(b).

- public interest
- legitimate interest.

In this case, there is no legal obligation or vital interest to collect speech. Public interest could be applicable to scientific research, but it is too restrictive considering that the material should also benefit commercial product development. To use a contract as a legal basis would require that processing is necessary for the performance of a contract to which the data subject is party. That was not the case. In principle, it would have been possible to use consent as the legal basis, but that was considered impractical, because the consent must be specific and the data subjects have the right to withdraw their consents at any time.

Therefore, legitimate interest to collect and process speech to be used for studying language as well as for developing technology and services that can be readily used in the languages spoken in Finland was chosen to be the best basis for the processing of personal data in the campaign. However, it was recognized that if it becomes necessary to also process special categories of personal data, like racial or ethnic origin, political opinions, religious or philosophical beliefs, data concerning health, or data concerning a natural person's sex life or sexual orientation, the explicit consent to the processing of such personal data is needed in accordance with GDPR Article 9. Until then, the controllers strive not to collect and process any personal data in these special categories.

To inform the data subjects (i.e., the individuals who donate their speech to the campaign), two essential documents were drafted:[10]

- A short *information page* including simple conditions of participation. It briefly describes the campaign and the responsible organizations, emphasizes that the donation is completely voluntary, explains that the individual may have copyright or other rights in the speech and he/she will need to assign those rights to the extent necessary, asks not to provide any personal data or intellectual property of others, provides links to the data protection policy and some additional information, and finally asks the person to accept these terms. It should be noted that this is not consent to process personal data as discussed above; rather, the lawful basis is a legitimate interest to process personal data.
- A more comprehensive *data protection policy*, titled "Tietosuoja" (Data Protection). The policy aims to describe, in a comprehensible and clear way, how personal data are processed in the campaign. It gives some basic information on data protection and describes how the donor can remove the donated

10 https://lahjoitapuhetta.fi/

speech from the campaign. Furthermore, it attempts to fulfil the data subject's right to be informed, as prescribed in the GDPR, Articles 12 and 13. The controllers (University of Helsinki, Yle, and Vake) are identified, their contact information and the contact details of their data protection officers are disclosed, and the controllers' responsibilities specified; the purpose of the processing of personal data is explained, the legitimate interest as the lawful basis of processing is specified and justified, the categories of personal data are listed, the principles to whom the personal data can be transferred are stated, and it is explained for how long the data is stored. The data subject's applicable rights are explained: the right to be informed and to get access to data, the right to request rectification or erasure of personal data or restriction of processing concerning the data subject and to object to processing, and the right to lodge a complaint with a supervisory authority. It is also noted that personal data is not used for automatic decision-making nor for direct marketing.

In order to use legitimate interest as the lawful basis for the processing of personal data, it was necessary to accomplish a balance test to ensure that the legitimate interests are not overridden by the interests or fundamental rights and freedoms of the data subject. The Finnish Data Protection Authority has published a model balance test, which was carefully applied. The model consists of six steps:

1. Is legitimate interest the most appropriate basis for processing?
2. Are the basic requirements (legal, clearly stated, representing a genuine and direct need) met?
3. Is the processing of personal data necessary for pursuing the interest?
4. Does the interest truly override the rights and interests of the data subject?
5. How are additional guarantees for data protection ensured?
6. How is the legality and transparency of the operations demonstrated?

To better understand the risks and possible problems that the processing of personal data may cause to individuals, a careful risk assessment was also performed. After completing all six steps, it seemed clear that a legitimate interest existed, met the legal requirements, and was not overridden by the interests or fundamental rights and freedoms of the data subject.

It was also considered that the risks to the rights and freedoms of natural persons were not very high. However, just to be sure, it was decided, in accordance with GDPR Article 35, to carry out a data protection impact assessment (DPIA) as well. The above-mentioned balance test to ensure that the legitimate interests are not overridden by the interests or fundamental rights and freedoms of the data subject – especially when complemented with a significant risk assessment – is

not very different from a data protection impact assessment. Therefore, it was possible to reuse most of the balance test in the DPIA and only complement it as required by the GDPR.

4.1 Data protection impact assessment

A data protection impact assessment (DPIA) was carried out because of possible risks related to the processing of data. In particular, the extensive processing as well as the new technologies and innovation development related to the purpose of processing were taken into account.[11] The University of Helsinki and Yle have data protection officers and they were involved in the data protection impact assessment as required by the GDPR.[12]

In the DPIA, the processing of personal data in the campaign was described in line with the discussion above. The purpose of the processing was described, the controllers and their responsibilities were specified, and the subcontractors were listed. It was explained who may receive the data, and it was noted that they can be located outside the European Union and the European Economic Area. The different phases of the processing were described, and the data that was to be processed, the sources of the data, and the purpose of processing were defined. The assessment of the necessity and proportionality of the processing operations in relation to the purposes was included. An essential part of the DPIA was the listing and the analyses of the recognized risks to the rights and freedoms of data subjects.[13]

The outcome of the DPIA was that the processing does not result in a high risk after the measures taken by the controllers to mitigate the risks. The DPIA will be updated as needed, if for example processing of special categories of personal data becomes necessary.

4.2 Communicating the data to the public

The Language Bank Rights (LBR) is an electronic application system for managing access to language resources. It is based on the Resource Entitlement Management System (REMS) developed by CSC for research data. A solution is being designed for how the LBR REMS will be accessible by private companies as well.

11 GDPR Article 35(1).
12 GDPR Article 35(2).
13 GDPR Article 35(7).

The Language Bank of Finland will begin redistributing the speech data when a sufficient amount of material has been donated and when the appropriate rights application process is in place in the beginning of 2022. For academic researchers, the use of the data will be free of charge, like the rest of the services of the Language Bank of Finland. For commercial use, a fee will probably be charged in order to cover handling costs.

5 Technical implementation

Speech for the Donate Speech campaign[14] could be donated via a web browser or mobile app, both of which offered a selection of tasks with light-hearted themes that aimed to inspire and encourage the user to talk about a particular topic. Representatives from both industry and academia developed the general specifications for the app. The software solution development company Solita developed the apps. The software platform has been published as open-source software,[15] allowing other organizations to build their own systems for collecting similar speech material or to enable specialized collection campaigns by researchers, or similar campaigns in other countries.

Technical voice quality is a complicated topic of its own. Having the microphone near the user is imperative, so advising more relaxed use, like leaving the phone on the table, would introduce more echoes and weaker signal. A discussion format with a group of people was also ruled out. There would have been obvious benefits, like the free-flowing, back-and-forth dialogue that characterizes a group discussion but does not exist in a single-speaker situation. However, that would have presented technical challenges rendering it hard to use when everybody should be close to a single microphone, or far away from each other with everyone having his own device to minimize cross-feeds and echoes in the signal. In addition, multiple signals would need to be synchronized in the backend system, or there would be a need to register which phones were co-recording the multi-mic discussion. For this reason, no user testing was conducted on which styles of dialogue triggers would work best for yielding interesting, differing flows of dialogue.

The recordings were kept simple by recording the speech signal in the highest lossless formats possible and accompanying them with metadata about the system, phone type, and version. The metadata therefore allowed for some post-processing

14 https://lahjoitapuhetta.fi/

15 https://github.com/CSCfi/Kielipankki-donatespeech-backend

corrections using, for example, sound equalization according to microphone type. A rudimentary VU-meter to give feedback to the user about an acceptable signal level was considered but not implemented, to conserve battery and diminish the development burden. Based on user testing, the usefulness of this feedback was also in doubt. First, the meter would provide a distraction or most likely be ignored; second, educating the user on how to interpret this additional information would encumber the user interface; and third, the improvement of the signal would not be substantial as the user would mainly move closer to the microphone for some time.

In the end, users were instructed to speak freely in their own environment. A clear signal in a noise-free environment is often preferable, but currently the recordings have a bit more variety as they also contain some noise, such as people in the background or wind in outdoor settings. In any case, according to the user tests, most people did the recording sessions on their own in rather quiet indoor settings. A delayed transmit in the background of locally stored recordings for uploading to the cloud was prepared in case the user did not have a steady internet connection, but it was probably not that important a feature.

The web, Android and iOS were chosen as platforms for smartphones, tablets, and computers with microphones. There was also an associated website informing users about the campaign and Yle published its own articles and campaign site. The apps were released from the Yle account instead of using separate dedicated or campaign-specific accounts to lend trust in an established entity to the campaign.

The solution architecture consisted of multiple frontends on different platforms, backend services and databases to collect data in the cloud, the web hosting, and the analytics. By splitting responsibilities for analytics and backend hosting, the visibility of the legal entities could be limited, so the party driving the campaigns had the option to access usage data to focus the campaign efforts without access to the raw speech donations. The system was developed for monolingual use, but further adaptation and localization to other languages and other themes was kept in mind.

To comply with the GDPR and to enable deletion of contributions, the backend allows easy deletion of user submissions through a long random identifier given to the user at the time of speech donation. There are no other user-specific identifiers in the backend data. One still needs to consider that individual users may be identifiable by their metadata in the case that the participating group is small or a combination of metadata very specific. For example, men of a certain age bracket in a small geographical area with a particular dialect background could potentially result in a tiny group of people both in the collected data and the real world. The technical platform as such does not restrict the collection of overly

specific metadata, as the GDPR-compliant processing of data is the responsibility of the controller and the processors when further processing the data or publishing findings in a way that is anonymous.

In spring 2021, the Android and iOS mobile application versions of Donate Speech were submitted to the annual marketing competition GrandOne, for web applications launched during the previous year. The Donate Speech applications won the prestigious first prize[16] in the mobile service category and an honourable mention in the category for best data use. Yle also submitted the Donate Speech campaign to the annual Prix Europa competition for European broadcasters, and in autumn 2021, after a thorough evaluation, the Donate speech campaign won the category of Best European Digital Audio Project 2021[17] in the highly prestigious TV, radio, and online product competition, chosen from among 684 entries from 26 countries. The award recognized a fresh way to conceptualize broadcasting and its output; the new cooperation model between commercial and public service entities and a broadcasting company like Yle; and a great web service accompanied by a light-hearted and humorous campaign.

6 Characteristics of the Donated Speech data

The objective for the Donate Speech campaign was to collect 10,000 hours of speech during half a year of campaigning. That would have meant about obtaining 8.5 seconds from each 10- to 70-year-old person in Finland, or getting 600,000 persons to donate a minute each, or 120,000 persons to donate 5 minutes each. The objective was considered quite a stretch but attainable in an optimal situation.

The campaign collected about 3,500 hours in half a year. The launch on national TV in June 2020 inspired the biggest number of contributions, but as can be seen in Figure 1, the summer of 2020 during the Covid-19 pandemic was quite active. The campaign was able to reach new audiences throughout the autumn but at a considerably slower pace. Towards the end of the campaign, there was a push on regional radio to collect dialects and the last 10% was collected in a week around Christmas 2020. Yle had a campaign page for its campaign events.[18] The campaign had officially ended by New Year 2021, but trailing infomercials

16 https://grandone.fi/kilpailutyo/?entry=lahjoita-puhetta-siivittaeae-suomenkielistae-puheentunnistusta

17 PRIX EUROPA 2021 Winners – PRIX EUROPA (https://www.prixeuropa.eu/news/2021/10/15winners-y4emh).

18 https://yle.fi/aihe/lahjoita-puhetta

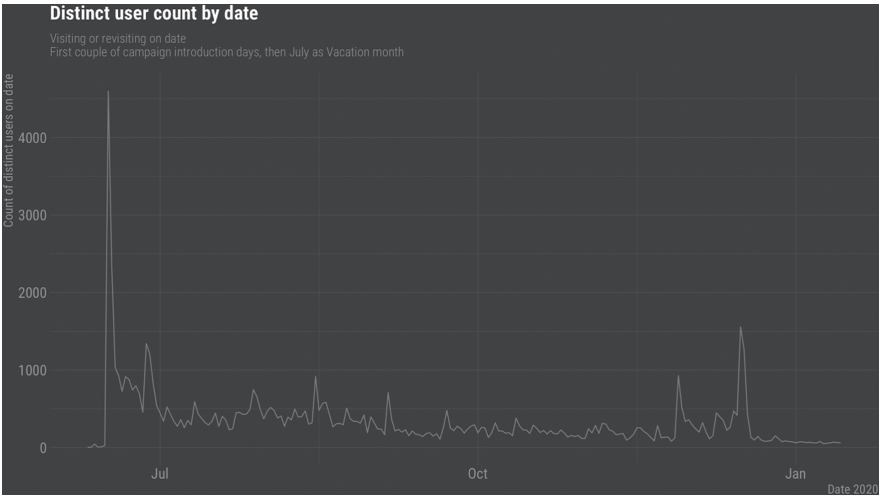

Figure 1: Distinct user count by date.

and reruns were still broadcast during the spring of 2021, resulting in a trickle of additional contributions.

Figure 2 breaks down the speech donations by age group. There are hours of data representing a wide range of age brackets. Perhaps surprisingly, 21- to 30-year-old females, unfazed by the somewhat technical set-up, donated most of the speech. The smallest amount of speech was donated by very young participants (1–10 years old) and very old participants (80 years or more). Two groups to consider for future focus activities are teens around 11–20, and retired people around 71–80. Both have distinctive characteristics from an AI development point of view, speaking with different pitch, vocabulary, pace, breaks, and potentially with interleaving and heavier breathing. One industry partner considers developing AI-powered elderly care systems, and specific modes like talking while lying down would also be useful.

Not everyone provided all the metadata, but among those who provided metadata, we can make some interesting observations. People between 20–60 years old made around three quarters of the donations. More than 70% of the donors were women. As expected, almost half of the donations were from the four regions with the largest Finnish cities: Uusimaa (including Helsinki and Espoo), Pohjois-Pohjanmaa (including Oulu), Varsinais-Suomi (including Turku), and Pirkanmaa (including Tampere), but donations were made from all the regions of Finland and 50 different counties, with 95% of the donors being native speakers. We note that the geographic areas have about the same amount of donations per 100,000 inhabitants, with approximately 60% to 150% deviation from the mean. A consid-

erably larger share of Swedish and Saami minority speakers in some areas probably explains a couple of outliers with smaller contributions. More than two thirds of the data was donated by students, retired persons, teachers, entrepreneurs, experts, and nurses (in descending order of contributor number) with the remainder contributed by more than 30 other professions from diverse areas of society. Approximately 62% had a higher education and 28% a secondary education.

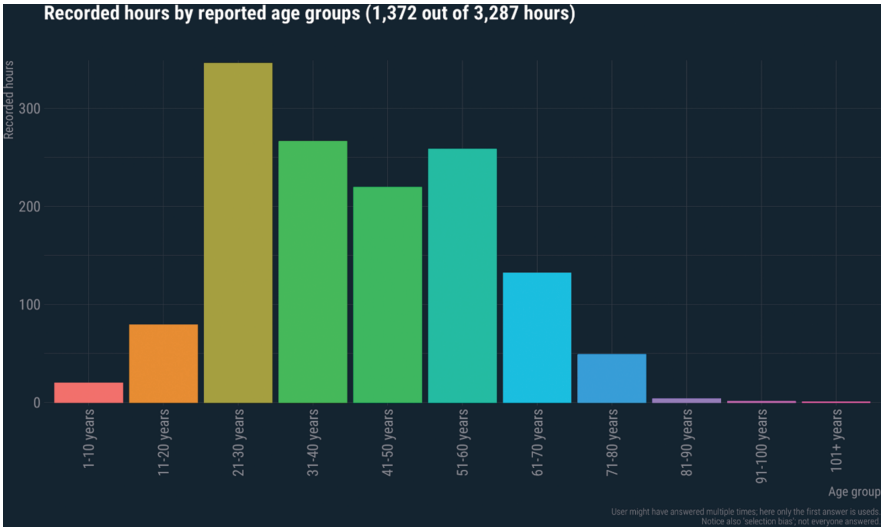

Figure 2: Recorded hours by age group.

Interestingly, the web interface was used by two thirds of the donors, and only 20% used the Android app with the rest using the iPhone app. Close to 90% of the more than 220,000 recordings were between 10 seconds and 3 minutes, with the median length being 30–60 seconds, in the end totalling roughly 4,000 hours.

There are a couple of limitations as to the reliability of these figures. The analytics data consist of a sequence of events of donations and interleaved metadata questions. Some users have not answered all the demographic questions. Other users might have multiple differing answers so the attribution of donation hours per metadata subcategory remains an estimate. In addition, the analytics system missed about 10% of the user events. Still, we believe that the figures paint quite a good initial picture of the success of the campaign.

After 80 hours of an initial random sample of the speech data was quality checked and manually transliterated, the initial impression was quite positive. Small random samples (1, 10, and 80 hours) of manually transcribed data were

evaluated by the current automatic speech recognition technology group at Aalto University to assess how accurately this material can be automatically transcribed, what kind of errors occur, and how the accuracy varies according to the conditions and given metadata. The initial impressions were rather positive: the material is on average not harder to recognize than previously recorded conversations at the Aalto University, despite being more diverse in terms of speakers, ages, dialects, and topics, as well as recording devices and conditions.

7 Conclusion

Even though the target of 10,000 hours was ambitious, the Donate Speech campaign has managed to collect an extensive resource of Finnish colloquial speech from a large number of speakers in just a few months. The campaign was implemented by Yle (the National Broadcasting Company of Finland) in cooperation with Ilmastorahasto (former state development company Vake) and the University of Helsinki. The University of Helsinki represented FIN-CLARIN and its service centre Kielipankki (the Language Bank of Finland), through which the FIN-CLARIN members make available various language resources, both corpora and tools.

Society currently requires a number of digital user skills, such as the utilization of mobile devices. If a user's vision is impaired or their finger dexterity is insufficient for a device, a user may currently be excluded from many services. To develop such services, speech data that is also available for commercial purposes was needed. At the beginning of the 21st century, the efforts and resources of Finnish speech technology and spoken language research were scattered all over Finland and represented by relatively small teams or researchers or public bodies. While automatic speech recognition (speech-to-text) and speech synthesis (text-to-speech) in Finnish have been available in a few devices and applications for several years (e.g., as speech capabilities in Apple and Google products), implementing or enhancing many end-user services still requires better and more reliable processing support for colloquial Finnish. To remedy this there was a need for collecting and making available a sizable amount of speech data that could also be used for commercial purposes.

In Finland, there are several extensive speech databases that were previously collected for linguistic research by the Institute for the Languages of Finland, the universities, and memory organizations, but for commercial purposes access to them is limited. Renegotiating licenses for corpora to allow business use is one way to add commercially usable speech material, but it is often not feasible to renegotiate access rights after data has already been collected and licensed.

The Donate Speech campaign had a Finnish predecessor called Prosovar as regards new methodology and new ways of obtaining speech data over the internet, implementing a crowdsourcing approach. The goal of the Donate Speech campaign was not merely to collect a vast amount of any kind of speech, but to reach out to as many different groups of Finnish speakers and to as many individuals as possible. In marketing the campaign to citizens, it was emphasized that all variants of spoken Finnish are welcome, including speech from second-language Finnish learners. However, in order to understand the privacy notice and the instructions, a certain level of language proficiency was required from the speech donors. In order to strike a balance between the material goals, the technical possibilities, and the resources that were available, design workshops were organized for all interested parties.

From the beginning, it was clear that the processing of data must be conducted in a legally and ethically sound way. All the central actors in the project (Kielipankki at the University of Helsinki, Vake, and Yle) are public organizations that cannot ignore these aspects. To better understand the risks and possible problems that the processing of personal data may cause to individuals, a careful risk assessment was also performed. After completing all the six steps of the balance test, it seemed clear that a legitimate interest existed, met the legal requirements, and was not overridden by the interests or fundamental rights and freedoms of the data subject. A data protection impact assessment (DPIA) was carried out because of possible risks related to the processing of data. In particular, the extensive processing as well as the new technologies and innovation development related to the purpose of processing were considered. The Language Bank Rights (LBR) is an electronic application system for managing access to language resources. The Language Bank of Finland will begin redistributing the speech data when a sufficient amount of material has been donated and when the appropriate rights application process is in place in the beginning of 2022.

In the end, Yle developed around 40 rather straightforward themes for stimulating the collecting of speech data. As part of the campaign, Yle made comical infomercials with requests to the general public to donate speech. These were broadcast during programme breaks in national radio and TV channels during the summer and autumn of 2020, during the Covid-19 pandemic, with some trailing reruns during spring 2021. Speech for the Donate Speech campaign (Lahjoita puhetta) could be donated via a web browser or mobile app, both of which offered a selection of tasks with light-hearted themes that aimed to inspire and encourage the user to talk about a particular topic. To comply with the GDPR and to enable deletion of contributions, the backend allows easy deletion of user submissions through a long random identifier given to the user at the time of speech donation.

Not everyone provided all the metadata, but among those who provided metadata, we can make some interesting observations. People between 20 and 60 years old made around three quarters of the donations. More than 70% of the donors were women. As expected, almost half of the donations were from the four regions with the largest Finnish cities: Uusimaa (including Helsinki and Espoo), Pohjois-Pohjanmaa (including Oulu), Varsinais-Suomi (including Turku), and Pirkanmaa (including Tampere), but donations were made from all the regions of Finland – 50 different counties – with 95% of the donors being native speakers. We note that the geographic areas have about the same amount of donations per 100,000 inhabitants, with approximately 60% to 150% deviation from the mean. A considerably larger share of Swedish and Saami minority speakers in some areas probably explains a couple of outliers with smaller contributions. More than two thirds of the data was donated by students, retired persons, teachers, entrepreneurs, experts, and nurses (in descending order of contributor number) with the remainder contributed by more than 30 other professions from diverse areas of society. Approximately 62% had a higher education and 28% a secondary education. Interestingly, two thirds of the donors used the web interface for donating speech, and only 20% used the Android app with the rest using the iPhone app. Close to 90% of the more than 220,000 recordings were between 10 seconds and 3 minutes, with the median length being 30–60 seconds, totalling roughly 4,000 hours.

After 80 hours of an initial random sample of the speech data was quality checked and manually transliterated, the initial impression of the collected data was quite positive. At the time of writing, 1,500 hours of speech has been transliterated, which will allow much more precise training of speaker independent supervised speech recognition, as well as new directions in research in unsupervised or minimally supervised machine learning of speech processing using current neural network technology.

Bibliography

Alen-Savikko, Anette K. & Olli Pitkänen. 2016. Rights and entitlements in information: Proprietary perspectives and beyond. In Tobias Bräutigam & Samuli Miettinen (eds.), *Data protection, privacy and European regulation in the digital age*, 3–33. Helsinki: Forum Iuris.

Altosaar, Toomas & Mietta Lennes. 2005. A graphical query formation compiler for speech database access. In Margit Langemets & Priit Penjam (eds.), *The Second Baltic Conference on Human Language Technologies, Tallinn, Estonia, April 4–5, 2005*, 209–218.

Altosaar, Toomas, Bruce Millar & Martti Vainio. 1999. Relational vs. object-oriented models for representing speech: A comparison using ANDOSL data. In *Proceedings of EUROSPEECH'99, Budapest, Hungary, 5–9 Sept. 1991*, Vol. 2, 915–918.

Ballardini, Rosa Maria, Pamela Lönnqvist, Perttu Virtanen, Nari Lee, Marcus Norrgård & Olli Pitkänen. 2013. The "one-size fits all" European patent system: Challenges in the software context. In Katja Weckström (ed.), *Governing innovation and expression: New regimes, strategies and techniques*, 327–350. Turku: University of Turku.

Chambers, J. K. 1994. An introduction to dialect topography. *English World-Wide* 15 (1). 35–53. https://doi.org/10.1075/eww.15.1.03cha.

Grönroos, Mickel & Manne Miettinen. 2004. Infrastructure for collaborative annotation of speech. *International Conference on Language Resources and Evaluation (LREC)*. 4. 543–546.

Kamocki, Paweł, Aleksei Kelli & Krister Lindén. 2022. The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

Karjalainen, Matti, and Toomas Altosaar. 1993. An object-oriented database for speech processing. In *Proceedings of Eurospeech 1993*, 183–186. Madrid.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Paweł Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavrilidou & Pavel Straňák. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In Inguna Skadin & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2018: Pisa, 8–10 October 2018* (Linköping Electronic Conference Proceedings 159), 72–82. Linköping: Linköping University Electronic Press.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Paweł Kamocki, Arvi Tavast, Ramūnas Birštonas, Gaabriel Tavits, Mari Keskküla & Penny Labropoulou. 2020. CLARIN contractual framework for sharing language data: The perspective of personal data protection. In Costanza Navarretta & Maria Eskevich (eds.), *Proceedings of CLARIN Annual Conference 2020. 5–7 October 2020*, online edition, 171–177. Utrecht: CLARIN ERIC.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Paweł Kamocki, Arvi Tavast, Ramūnas Birštonas, Gaabriel Tavits, Mari Keskküla, Penny Labropoulou, Irene Kull, Age Värv, Merle Erikson, Andres Vutt & Silvia Calamai. 2021. Sharing is caring: A legal perspective on sharing language data containing personal data and the division of liability between researchers and research organisations. In Costanza Navarretta & Maria Eskevich (eds.), *Selected Papers from the CLARIN Annual Conference 2020: 5–7 October* (Linköping Electronic Conference Proceedings 180), 129–147. Linköping: Linköping University Electronic Press.

Kelli, Aleksei, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki & Pavel Stranák. 2018. Implementation of an Open Science Policy in the context of management of CLARIN language resources: A need for changes?. In *Selected papers from the CLARIN Annual Conference 2017: Budapest, 18–20 September 2017* (Linköping University Electronic Press 147), 102–111. Linköping: Linköping University Electronic Press

Kelli, Aleksei, Tönis Mets, Kadri Vider, Age Värv, Lars Jonsson, Krister Lindén & Ramūnas Birštonas. 2018. Challenges of transformation of research data into open data: The perspective of social sciences and humanities. *International Journal of Technology Management & Sustainable Development* 17 (3). 227–251.

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits & Age Värv. 2019. The extent of legal control over language data: the case of language technologies. In Kiril Simov & Maria Eskevich (eds.), *Proceedings of CLARIN Annual Conference 2019*, 69–74. Utrecht: CLARIN ERIC.

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Ramūnas Bristonas, Penny Labropoulou, Kadri Vider, Irene Kull, Gaabriel Tavits, Age Värv & Vadim Mantrov. 2020. Impact of legal status of data on development of data-intensive products: Example of language technologies. In A. Damberga (ed.), *Legal Science: Functions, Significance and Future in Legal Systems II. Collection of Research Papers in Conjunction with the 7th International Scientific Conference of the Faculty of Law of the University of Latvia*, 383–400. Riga: University of Latvia Press.

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Straňák & Jan Hajic. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. In Kiril Simov & Maria Eskevich (eds.), *Selected Papers from the CLARIN Annual Conference 2019*. (Linköping Electronic Conference Proceedings 172), 53–65. Linköping: Linköping University Electronic Press.

Kelli, Aleksei, Kadri Vider & Krister Lindén. 2016. The regulatory and contractual framework as an integral part of the CLARIN infrastructure. In Koenraad De Smedt (ed.), *Selected papers from the CLARIN Annual Conference 2015: October 14–16, 2015, Wroclaw, Poland* (Linköping Electronic Conference Proceedings 123), 13–24. Linköping: Linköping University Electronic Press.

Kurki, Tommi, Tommi Nieminen, Heini Kallio & Hamid Behravan. 2014. Uusi puhesuomen variaatiota tarkasteleva hanke: Katse kohti prosodisia ilmiöitä [A new project considering variation in spoken Finnish. A view towards prosodical phenomena]. *Sananjalka* 56. 186–195.

Labov, William, 1972. *Sociolinguistic patterns*. Oxford: Blackwell

Lane, Ian, Alex Waibel, Matthias Eck & Kay Rottmann. 2010. Tools for collecting speech corpora via Mechanical-Turk. In Chris Callison-Burch & Mark Dredze (eds.), *Proceedings of the NAACL HLT 2010, Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 184–187. Stroudsberg, PA: ACL.

Lennes, Mietta & Sanna Ahjoniemi. 2005. Puheaineiston annotaatio eli nimikointi (Version 1.01) [Annotating speech data (Version 1.0)]. Zenodo. http://doi.org/10.5281/zenodo.1205453.

Lindén, Krister, Aleksei Kelli & Alexandros Nousias. 2020. A CLARIN Contractual Framework for Sharing Personal Data for Scientific Research. In Kiril Simov & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2019* (Linköping Electronic Conference Proceedings 172), 75–84. Linköping: Linköping University Electronic Press.

McGraw, Ian. 2013. Collecting speech from crowds. In Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent & David Suendermann (eds.), *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*, 37–71. Chichester: Wiley.

Nieminen, Tommi & Tommi Kurki. 2015. Prosovar-hankkeen väliraportti. Puheaineiston keruusta verkossa sekä havaintoja aineistosta [Collecting speech data on the internet and observations about the data]. In Mona Lehtinen & Unto K. Laine (eds.), *XXIX Fonetiikan päivät, Espoo 20.–21.3.2015, Julkaisut – Papers* (Tiede + Teknologia 7/2015), 29–38. Espoo: Aalto University. http://fp2015.aalto.fi/Fonetiikan_Paivat-2015_Aalto-yliopisto.pdf (accessed 1 March 2022)

Nieminen, Tommi & Tommi Kurki. 2017. Collecting speech data over the internet: Web 2.0 and speech corpora. In *Digital humaniora i Norden / Digital Humanities in the Nordic*

Countries, Göteborg, March 14–16 2017. Book of abstracts, 159–160. Gothenburg: Gothenburg University.

Oksanen, Ville & Krister Lindén. 2011. Open content licenses: How to choose the right one. *NEALT Proceedings Series* 13. 11–18.

Oksanen, Ville, Krister Lindén & Hanna Westerlund. 2010. Laundry symbols and license management: Practical considerations for the distribution of LRs based on experiences from CLARIN. LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Valletta, Malta, 23 May 2010.

Pitkänen, Olli. 2017. Mitä lähioikeus suojaa? [What does related rights protect?]. *Lakimies* 115 (5). 580–602.

Schröder, Vilhelm. 2018. *Legislative update: Implementation of the Trade Secrets Directive – A new trade secrets act proposed in Finland*. Liikejuridiikka 2018/1.

Toivanen, Juhani & Manne Miettinen. 2001. *Puheentutkimuksen resurssit Suomessa* [Resources for speech research in Finland]. Espoo: CSC.

Thomas, Erik R. 2013. Sociophonetics. In J. K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change*, 2nd edn., 108–127. Oxford: John Wiley and Sons.

Weckström, Katja. 2012. Trademarks in virtual worlds: Law, outlaws or new in-laws? *Journal of International Commercial Law and Technology* 7 (2). 112–120.

Wrigley, Sam, Anette Alen-Savikko & Olli Pitkänen. 2019. Finding the 'personal' in the industrial internet: Why data protection law still matters. In Rosa Maria Ballardini, Olli Pitkänen & Petri Kuoppamäki (eds.), *Regulating industrial internet through IPR, data protection and competition law*, 235–252. Alphen aan den Rijn: Kluwer Law.