FIN-CLARIAH 2024/25

WP1.2 Speech processing and annotation

Author:
Martin Matthiesen, CSC
Collaborators:
Sjur Moshagen, UiT
Katri Hiovain-Asikainen, UiT
Jo Raknes, NRK
Moa Frygell, SVT
Mikko Kurimo, Aalto
Krister Lindén, UH

DESCRIPTION OF WORK

The foreseen ***impact*** is to provide automated speech recognition with an emphasis on recognizing and classifying everyday speech and dialects furthering Goal 1. We currently ***have*** the Donate Speech data for Finnish. Based on this data set, we have developed a web service for Finnish speech transcription. For Goal 4, we ***need*** (D:1.2.1) to collect colloquial speech for Swedish and the Sámi languages and (D:1.2.2) to provide transcription services for interviews in these languages by SSH scholars based on transformer technology.

TIMELINE

Quarter	1	2	3	4	5	6	7	8
			D 1					
							D 2	

Status and Plan

- Donera Prat raw data is collected, awaits processing: Plan is to partially transscribe the data to be used as ASR training data.
- Access to Finnish Sámi data for research is already possible. NRK and SVT now working on legal frameworks.
- For Sámi the plan is to make the data available according to language variant, regardless of originating institution or country. Data needing extra protection will be protected using the Language Bank of Finlands ”Language Bank Rights” process.

OUTCOME / Deliverable #1

Donera Prata dataset and Sami language data provided via the Language Bank of Finland. Especially in the case of Sámi we will have an understanding of the Legal and technical challenges of sharing partially copyrighted data from three difference jurisdictions (FI,SE,NO) and multiple sources (NRK,SVT,YLE, Sametinget).

- Public Sámi audio data from two different sources packaged for download with proper descriptive metadata and license
- Donera Prat audio data (Swedish spoken in Finland) partially manually transcribed and ready for download
- Broadcasters (NRK, SVT, YLE) Sámi audio data available for download with proper metadata and license

OUTCOME / Deliverable #2

A Whisper based speech recognizer for at least colloquial Swedish spoken in Finland and North Sámi.

- Sámi speech recognizer model available for North Sámi at least for Whisper
- Speech recognizer model for Swedish spoken in Finland available at least for Whisper