FIN-CLARIAH 2024/25

W.P 2.3. TRANSLATION AND INTERPRETATION

Authors:
Tommi Jauhiainen
University of Helsinki
ChatGPT 4o
OpenAI

IMPACT

The foreseen impact is to provide infrastructure for translation and interpretation research both in machine translation as well as in translation studies. An important aspect of this is the search and retrieval of translation samples, i.e. bilingual samples in parallel corpora and monolingual samples in related corpora in different languages furthering Goal 1 and 2.

WE HAVE

We currently have reasonably large parallel samples of text in our Opus service. We have also developed the provision of access to parallel speech data.

WE NEED

We currently need (D:2.3.1) to upgrade our access to remote text data repositories like the Parliament and the National Broadcasting company (YLE). The current project will also (D:2.3.2) upgrade the LBF with a service for accessing spoken data from the Finnish Parliament as well as data in the Living Archive of YLE.

D:2.3.1. ACCESS TO REMOTE TEXT DATA REPOSITORIES

The concept of a "remote text data repository" typically evokes images of large, structured databases accessible via the internet, housing vast amounts of textual data for various applications. While the Finnish Parliament or the YLE may not traditionally consider themselves as such, they inherently embody the characteristics and functionalities of a remote text data repository offering valuable resources for translation and interpretation research, machine translation development, and broader translation studies.

We currently have the “Plenary Sessions of the Parliament of Finland” version 1.5 available via the Korp service as well as a downloadable version. It contains the transcriptions of of the plenary sessions of the Parliament of Finland from 10.09.2008 to 1.7.2016. (http://urn.fi/urn:nbn:fi:lb-2019101621).

The dataset needs to be updated and currently we are lacking the means to perform routine updates. Additionally, we also only have the original, multilingual, versions of the proceedings even though the Parliament is providing translations to Finnish and Swedish for every speech.

The Parliament of Finland hosts an open data portal at “avoindata.eduskunta.fi”. There are various ways of accessing the data, but we have to figure a workflow/pipeline that does not get broken easily and would allow us to update the resource with minimal effort.

List of action points:

- Contact the Parliament and discuss the possibility to access the translations currently available only to the MPs.
- Study the possibilities provided by the current API to directly access the text of the speeches.
- Make a new version of the dataset using the excel files that the Parliament has published containing the original speeches until the end of 2023.

We currently have several text corpora orinating from the YLE textual news content:

- Yle Finnish News Archive 2011-2021
- Yle News Archive Easy-to-read Finnish 2011-2018 and 2019-2021
- Yle Swedish News Archive 2012-2018 and 2019-2020

We already have an access to the YLE API, through which we can download the needed data. However, the pipeline of producing new versions of the dataset still proves to be too cumbersome as is evidenced by the fact that we do not have newer versions available. We have a general resource publishing pipeline tasklist that we use for processing all corpora.

List of action points:

- Create a custom resource publishing pipeline focusing on producing new version of the Yle datasets.
- Publish new versions of all the existing Yle datasets with data until the end of 2023.
- Yle has textual news also in Northern Sami, Inari Sami, Skolt Sami, Livvi-Karelian Ukrainian, Russian, and English. We should consider making similar corpora of them as well.
- Investigate possibilities to create parallel corpora using all the previous datasets.

D:2.3.2. ACCESSING SPOKEN DATA FROM:

THE FINNISH PARLIAMENT

We already have mappings for the parliamentary audio data in Korppi. It fits more into the package 2.3.1. Mapping of new parliamentary materials' audio/video and session transcripts. What is needed for that?

The alignments should be managed, or alternatively, obtain timestamped data directly from the Parliament? A transcription made with a speech recognizer as the base, from which the found segments are replaced with text received from the Parliament through alignment.

THE LIVING ARCHIVE OF YLE

Investigate the possibility of obtaining subtitles for the Living Archive (OPUS? Yle?) and how to link from Korppi to the correct part of the video.