FIN-CLARIAH

WP 2.3 TRANSLATION AND INTERPRETATION

Authors:
Erik Axelson
FIN-CLARIN, University of Helsinki
Mietta Lennes
FIN-CLARIN, University of Helsinki
Jyrki Niemi
FIN-CLARIN, University of Helsinki

THE GOALS

The foreseen impact is to provide infrastructure for translation and interpretation research both in machine translation and 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.

CURRENT SITUATION

We currently have reasonably large parallel samples of text in the Korp version of the Opus corpus (2.7 billion tokens, 16 languages). Other similar corpora include ParFin, ParRus, CEAL, HeKo-EuroParl, HeKo-JRC-Acquis, Semfinlex, MULCOLD, and Kotus Finnish-Swedish Parallel Corpus, all available in the Korp service. Aligned speech and transcript are also available in the Korp version of the Plenary Sessions of the Parliament of Finland, where links have been added from each utterance in the transcript to the corresponding portion of the video recording (22 million tokens, 3700 hours).

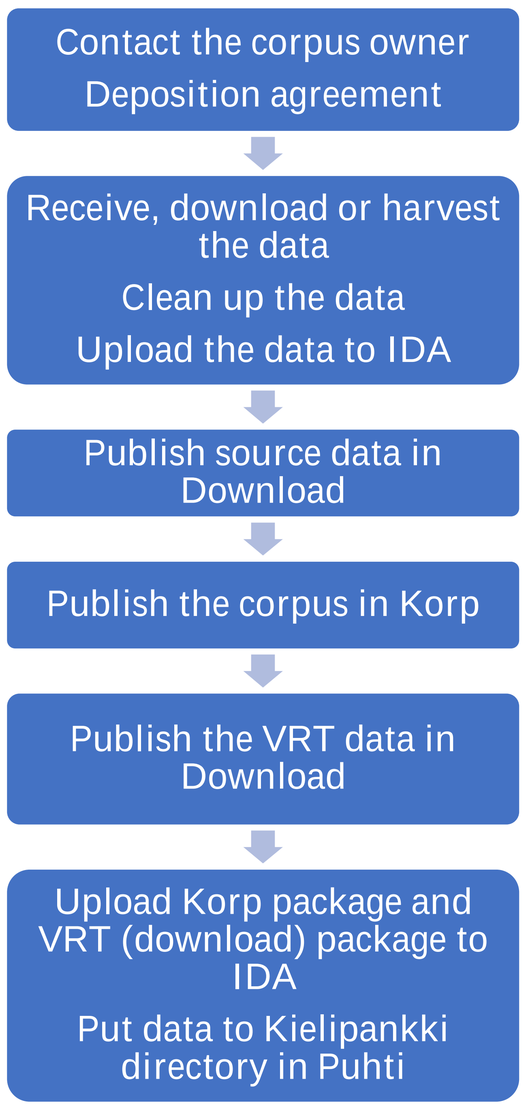
FUTURE TASKS

We need to provide access to speech data in other languages. The current project will upgrade the Language Bank with a service for retrieving interpretation samples for Finnish to English by licensing, aligning and sharing interpretation training sessions. We need to develop further the features supporting the use of parallel/translation corpora in Korp.

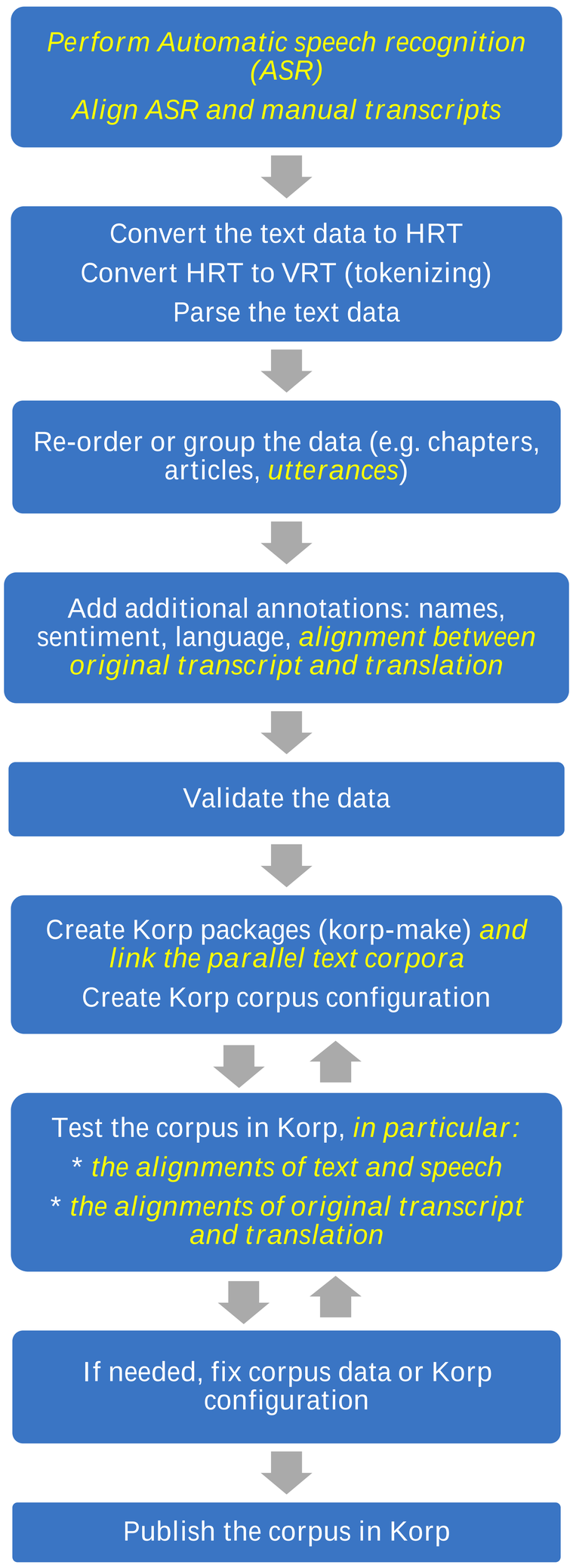
- License and obtain sets of data that contain original speech audio and the corresponding audio for (simultaneous) interpretation
- Perform automatic speech recognition of both audio signals (or create manual or semi-automatic transcriptions for both)
- Process the transcriptions and publish the original and interpreted transcripts as an aligned parallel corpus in Korp

We need to find a replacement for the discontinued LAT service and convert the existing resources to the new service. Features no longer available (at least not to the same extent) after LAT was discontinued include:

- browsing individual audio and video files along with their annotations (multiple tiers of time-aligned annotations)
- simple and complex queries within time-aligned annotation files (concordance including the matching items linked to the original annotated media)
- downloading individual media and annotation files or zip files of entire sessions (media+annotations)
- support for tier types and hierarchies, closed vocabularies and other structured annotations created in ELAN



The standard corpus workflow from raw data to publication in Korp or download service.



A detailed workflow of the phase “Publish the corpus in Korp” where special issues concerning translation/interpretation corpora are highlighted.