Building Multilingual Comparable Corpora

Rebecca Paterson

CNRS-LLACAN

SIL International

becky.d.paterson@gmail.com

Abbie Hantgan

CNRS-LLACAN

Christian Chanard

CNRS-LLACAN

Presented at
The 7th International Conference on Language Documentation & Conservation (ICLDC)
4–7 March 2021

Keywords: *ELAN, ELAN-CorpA, Language Documentation, West Africa, Corpora, Reported Speech*

Building on existing corpora and new audio/video documentary fieldwork from 12+ languages from across West Africa, we are creating a multilingual comparative corpus with input from 20+ collaborating researchers (Nikitina et al. 2020). We present a toolkit of technologies and three parallel workflows that can be used to mobilize language materials from diverse sources for a variety of purposes, particularly for the discovery of discourse patterns in legacy materials that could then be used in revitalization efforts.

Our toolkit includes the following technologies: ELAN-CorpA[1] (Chanard 2015; 2019), Fieldworks Language Explorer[2] (FLEx, SIL International), Toolbox[3] (SIL International) , ELAN Tools (Chanard et al. 2020, under development), SpeechReporting Template (Nikitina et al. 2019) and Tsakorpus[4] (Arkhangelskiy 2019). The three workflows differ with regard to the initial file format and the software platform that is to be used for parsing and glossing of texts. All three workflows lead to a collection of annotated files that can be queried with ELAN-CorpA (Hantgan 2019).

In the first workflow, (1) ELAN-CorpA is used for time-aligned translation and transcription of a recorded text; (2) FLEx is used to parse and gloss the text, and (3) ELAN Tools converts the `.flextext` export into the project template for use in ELAN-CorpA. (4) Once in the project template, the text is annotated for project categories and complex queries

[1] http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php
[2] https://software.sil.org/fieldworks
[3] https://software.sil.org/toolbox
[4] https://bitbucket.org/tsakorpus/tsakorpus see also: https://bitbucket.org/tsakorpus/tsakorpus/src/master/docs/pipeline.md

can be run across all texts in any language using search features of ELAN-CorpA. In the second workflow, (1) transcription, translation, parsing and glossing is done in Toolbox, (2) ELAN Tools converts a Toolbox file to the project template; (3) ELAN-CorpA is used to time align and annotate for the project. In the third workflow, translation, transcription, parsing and glossing is all done in ELAN-CorpA using the project template from initial stages. Using these three workflows, the data from various source file types is processed into a shared format that will be displayed in an online platform via Tsakorpus.

This methodology may be of interest to community members looking for ways to prepare already-collected language materials in order to display them on the internet, those interested in specific questions regarding discourse phenomena, typologists, and linguists in general.

Suggested Citation

Paterson, Rebecca, Abbie Hantgan & Christian Chanard. 2021. Building Multilingual Comparable Corpora. Paper presented at: The 7th International Conference on Language Documentation & Conservation (ICLDC), 4–7 March. University of Hawai'i at Mānoa. http://ling.lll.hawaii.edu/sites/icldc

References

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. *Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages.* 125–140. Tartu, Estonia. http://volgakama.web-corpora.net/Social_media_corpora_IWCLUL2019_final.pdf (2020-10-05.)

Chanard, Christian. 2019. ELAN-CorpA. Villejuif-Paris, France: CNRS-LLACAN. http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php (23 July 2019.)

Chanard, Christian. 2015. ELAN-CorpA: Lexicon-aided annotation in ELAN. In Amina Mettouchi, Martine Vanhove and Dominique Caubet, eds. *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages,* 311–332. Studies in Corpus Linguistics. 68. Amsterdam, Netherlands: John Benjamins Publishing Company. https://benjamins.com/catalog/scl.68.10cha (2020-10-02 12:43:12.)

Chanard, Christian. 2020. ELAN Tools. Villejuif-Paris, France: CNRS-LLACAN.

Hantgan, Abbie. 2019. La Construction d'un Corpus Comparative: Une Méthodologie Pour L'étude Des Langues Parlées en Afrique de l'Ouest. Paper presented at: LIFT 2019: Journées scientifiques "Linguistique informatique, formelle et de terrain" / Scientific meeting of the "Computational, formal and field linguistics" research group, 28–19 November. Orléans, France. https://lift2019.sciencesconf.org

Nikitina, Tatiana, Abbie Hantgan & Christian Chanard. 2019. Reported speech annotation template for ELAN. Villejuif-Paris, France: CNRS-LLACAN.

Nikitina, Tatiana, Elena Perekhvalskaya, Abbie Hantgan, Rebecca Voll, Ekaterina Aplonova & Rebecca Paterson. 2020. *SpeechReporting Corpus: Discourse Reporting in Storytelling.* Villejuif-Paris, France: LLACAN.