Lifting the stepchild out of poverty: Text collections as a complement to grammars and dictionaries

keywords: text collection, interlinear glossed text, grammar writing, open access, FAIR principles

Franz Boas established the "Boasian Trilogy" in language documentation and description (Himmelmann 1998), consisting of a grammatical description, a dictionary, and a text collection. All three are necessary to get a comprehensive overview of a language, and they complement each other. While we have good outlets for grammars (eg Comprehensive Grammar Library) and dictionaries (eg Dictionaria), such is not the case for text collections. This means that only few of them are published, and even fewer follow the FAIR principles of findability, accessibility, interoperability, and reusability (Wilkinson 2016).

The project Open Text Collections (`http://opentextcollections.org`) will remedy this by making high quality text collections from endangered languages available in an open interoperable format. Next to providing pdfs or printed books to the communities themselves, this setup will also provide the data in CLDF format (Forkel et al. 2018) for downstream use in NLP applications.

Most reference grammars published today are the result of a language documentation project, often part of authors' dissertation projects. These grammars should be data-driven and accompanied by a corpus in order to facilitate the verification or falsification of the analysis (Mosel 2012). While countless hours are invested into the structuring and glossing of texts, in many cases, however, these texts are not made available in a reusable way. Linguists tend to have them somewhere on their hard drive, or uploaded to an archive but there is no generally established way of publishing them, at least not in a format which would feed further research downstream (e.g. linguistic typology, corpus-based language description, or NLP). This means that these valuable results of language documentation often fail to be discovered.

Open Text Collections will provide a quality venue for publishing text collections, following the setup established by Language Science Press. The platform is community-driven and aims at being attractive to both data producers

(ie language documenters) as well as data users (language communities, typologists, NLP practitioners). For data producers, the platform will provide rigorous peer review, quality control, and top-notch publishing (pdf and print-on-demand), making sure that the time invested in a text collection will not harm job prospects. For data consumers, different outlets will be available to suit different needs: printed books will be available for communities; a search interface (prototype available at `https://imtvault.org`) will be available for typologists, and all data will be available as CLDF dump for NLP practitioners. By making reuse easy, the research will spread more widely, which in turn is very attractive for the data producers.

As of today, there are 5 regional boards and 40 proposed text collections. This presentation will showcase the platform, its motivations, and its benefits for data producers and consumers.

References

Forkel, Robert, Johann-Mattis List, Simon Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci Data* 5. DOI: 10.1038/sdata.2018.205.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Mosel, Ulrike. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. *Language Documentation & Conservation Special Publication* 4. 235–250.

Wilkinson, M. et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3. 160018. DOI: 10.1038/sdata.2016.18.