Paper Session: Ethical Workflows

Time: Thursday, 20/June/2024: 9:00am - 11:00am
Session Chair: **Andrew Janco**, University of Pennsylvania

Closing the loop: integrating students and the community in the Creolistic research workflow

Carlos Silva[1,2,4], Luís Trigo[1,3,4], Vera Moitinho de Almeida[3,4]
[1]CLUP - Centre of Linguistics of the University of Porto; [2]DEPER - Department of Portuguese and Romance Studies; [3]CODA - Centre for Digital Culture and Innovation; [4]FLUP - Faculty of Arts and Humanities, University of Porto, Portugal;
cssilva@letras.up.pt

CreoPhonPt is an interdisciplinary and collaborative research project on phonological, lexical and sociolinguistic information of Portuguese Creoles [1][2]; and based on the integration of Open Science, Citizen Science, and FAIR [3] and CARE [4] principles, into the flow of scientific production in the humanities and social sciences fields [5]. To foster students and citizen engagement, we integrated this project in the classroom environment.

The main goals of the Phonology seminar of the Master in Linguistics, at FLUP, are that students (I) acquire knowledge inherent to phonological theory and (II) can apply phonological theory to the description of natural languages. To this end, Portuguese-based creoles and African languages were selected for teaching how to build a linguistic-analysis workflow from bottom-up, while enhancing and taking advantage of CreoPhonPt's potentialities.

Across two years, we applied project-based learning [6] and cooperative learning [7] methods to encourage creative and critical thinking in the classroom and to foster collaboration and social responsibility outside the classroom. Seminars were split into ten blocks, each one comprising theoretical exposure and practical work. All stages of the workflow (Fig.1) were monitored by the teaching team. A set of freeware and opensource online tools, also easily available outside the classroom, were used. Students archived the data in GitHub's CreoPhonPt open access and version-controlled repository.

Year 1: focused on language data and metadata from existing bibliography. Each student was given a set of raw data, which had to be structured, manually/automatically processed (incl. data wrangling and reconciling) and archived.
Year 2: we took students a step back and focused on phonological data and metadata from field work. Students were divided in groups of four and made responsible for collecting, structuring, manually/automatically processing and archiving phonological/audio data from fellow colleagues Guinean-Bissau creole speakers, who assessed them throughout the workflow. Two interdisciplinary open seminars were organised, bringing together researchers from sociology, anthropology, education science, culture and communication, as well as technology.

RESULTS:

1) Learning: students acquired a full set of skills (incl. soft/hard, theoretical, technical, research, management) that can be useful beyond the scope of this Master course.
2) Community: Afro-Portuguese communities, which have these languages as their cultural and social asset, were actively involved in field data collection, authority/responsibility, archiving, (re)use and dissemination.
3) Research: the open access CreoPhonPt repository was largely enriched with further socio-historical, cultural and linguistic data/metadata for preservation, dissemination and further investigations.

The lack of previous studies and the fact that these languages are endangered was per se a motivation for all the participants. Likewise, the co-creation of scientific data and its collective benefit were very stimulating, as evidenced by the pedagogical surveys' results. Recently, the CreoPhonPt incursion in the classroom was awarded an honourable mention "Innovative Pedagogical Practice", by UPorto. More recently, students presented their work at a public event [8], the collected corpus was submitted to an international workshop [9], while the collective written essays were submitted to a special issue of *elingUP* (online open access journal of UPorto's Linguistics students).

INEL workflows for creating digital corpora of minority languages: Lessons learned

Alexandre Arkhipov, Elena Lazarenko, Aleksandr Riaposov
Universität Hamburg, Germany; aleksandr.riaposov@uni-hamburg.de

Since 2016, the INEL project has been working on creating richly annotated XML-based corpora of various minority languages of Northern Eurasia. Four corpora have been published by 2023 – Kamas, Selkup, Dolgan, and Evenki – with several more currently under development.

As we are dealing with severely endangered, low-resource languages, the source materials tend to be of diverse form and origin, including data from fieldwork archives containing audio recordings and handwritten texts. The texts that end up in the corpus encompass various genres such as folklore, personal narratives, and conversations on everyday topics. As a starting point, the available data need to be digitized and/or converted to a format suitable for import into SIL FLEx, a linguistic analysis software. Digitization may involve transcribing audio with a native speaker consultant (ELAN), manual typing text from manuscripts, Handwritten Text Recognition or OCR (Transkribus, ABBYY FineReader), transcoding into a unified Latin-based transcription in Unicode, etc. The next step in our workflow is linguistic annotation, which is divided in two stages: first, interlinear morpheme glossing in SIL FLEx; second, providing additional layers of annotation (e.g., syntactical functions, information structure) and corpus metadata via the EXMARaLDA package. At the latter stage, corpus data start being version-controlled in Git and undergo continuous curation through Corpus Services, an in-house developed Java package that helps maintain data consistency by pinpointing errors in the files and automatically correcting them if possible. The finalized corpora are provided in several XML-