Technological and Methodological Aspects of Research Data Quality Control

Anne Ferger
Universität Hamburg
anne.ferger@uni-hamburg.de

Hanna Hedeland
Universität Hamburg
hanna.hedeland@uni-hamburg.de

Despite recent technological advances, the creation of linguistic corpora based on spoken language still include time-consuming manual tasks. Transcription of lesser-resourced varieties or other types of non-standard language and several types of linguistic annotation must be carried out manually. Due to the great variety of research questions and theoretical frameworks, the creation of spoken corpora can not be standardized. At the same time, spoken language corpora are highly complex resources, comprising recordings, transcripts, annotations, further related files and metadata on recorded sessions and persons. The required flexibility and the resource's complexity often result in insufficiently controlled workflows and thus major problems with data integrity and consistency. These problems become evident when project data has to comply with quality standards to enable wide re-use in other contexts, but also prevent reliable searches on the data throughout the project duration. Our poster will illustrate our work aimed at developing the necessary technology and support for researchers to move towards a workflow which integrates Git-based versioning and automatic quality control tests into the corpus creation process.

Since the idea of bringing the concept of continuous quality control (or integration) into research data management is rather new to the Humanities (cf. Almas & Clérice (2017) and Ayer et al. (2017)), a major challenge is to overcome the impression of negative methodological impact through technology imposed on researchers' current workflows. Our focus has thus been on achieving acceptance among mainly non-technical users. For the automated consistency checks and fixes for (mainly EXMARaLDA (Schmidt & Wörner, 2014)) corpus data and metadata we have developed, we provide the results and further diagnostic overviews of the data in familiar sort- and filterable HTML data tables and error lists. To support manual correction of the data, we also use interactive EXMARaLDA error lists allowing users to display the needed changes for each transcript and correct these from within the familiar transcription editor. Apart from intuitive graphical user interfaces and familiar contexts, comprehensible, non-technical error messages have proven crucial for acceptance. With our poster and a demonstration of our system we hope to continue the dialogue with researchers on how to best exploit these technologies within various methodological approaches.

References: • Almas, B. & Clérice, T. (2017): Continuous Integration and Unit Testing of Digital Editions. *Digital Humanities Quaterly*, 11(4). • Ayer, V., Pietsch, C., Vompras, J., Schirrwagen, J., Wiljes, C., Jahn, N., & Cimiano, P. (2017). Conquaire: Towards an architecture supporting continuous quality control to ensure reproducibility of research. *D-Lib Magazine*, 23(1/2). • Schmidt, T. & Wörner, K. (2014): EXMARaLDA. *Handbook on Corpus Phonology*, 402-419.