Towards Continuous Quality Control for Spoken Language Corpora

Anne Ferger and Hanna Hedeland
University of Hamburg

INEL
Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages

HZSK
Hamburg Center for Language Corpora

Akademie der Wissenschaften in Hamburg
Union of the German Academies of Sciences and Humanities in Hamburg

CLARIN
Common Language Resources and Technology Infrastructure
Aim of the Presentation

- Our approach on optimizing a **linguistic data creation and curation workflow** aiming towards **continuous integration** of speech corpora
- The **data** we work with
- Our **framework** and **practical issues** we overcame
- Our **perspective**
Language Corpora

Structure of a spoken language corpus
Exemplary Data in the INEL Project

Language Data
• with and without audio and video
• transcriptions in XML format

Metadata
• diverse corpus-wide metadata
• in XML and CMDI format
Quality Management for Spoken Language Corpora

- Existing linguistic data creation and curation workflow: mostly manually and non-reproducible
- In our case: Creating searchable, consistent language corpora that can be used for quantitative or qualitative analysis
- Publishing that corpora in a Fedora repository
- Completely automated curation is not possible because it would require unacceptable constraints on the creation of the data
Users and Maintainers of Our Approach

- Technical staff in infrastructure projects at the research data centre
- Non-technical users in inhouse research projects
- Non-technical users in external research projects
Our Approach

Why do we need continuous quality control for spoken language corpora?

• Limiting (expensive) manual work
• Avoid unnecessary data curation
• Increasing the amount of automatic enhancements of the data
• Creating high quality resources suitable for different research needs
• Making the publishing of the resources as fast, spontaneous and easy as possible
• Enabling the conversion of the data into various different formats
Formalizing the Workflows

**Corpus Initialisation**
- Prepare files and folders
- Set up Git repository
- Validate and assess data
- Create error reports
- Create curation tasks

**Corpus Integration**
- Create derived corpus file formats
- Clean up file references within corpus
- Complete metadata and documentation
- Generate generic and specific metadata formats
- Finalise and tag publication version

Diagram:
- Corpus Initialisation → Data Curation → Corpus Integration → Repository Ingest
- Corpus Data Version 0.1 → Data Curation → Corpus Integration → Repository Ingest
- Corpus Data Version 0.2 → Data Curation → Corpus Integration → Repository Ingest

Corpus Version:
- Version 0.1
- Version 0.2
- Version 0.3
Using Git for Version Control

Using a git workflow

planned creation/curation

ad hoc corrections

publication with generated files

version 0.1
The Modules Used for Our Approach
Specific Git Solutions I

- Using different branches for publication

### Revisions

<table>
<thead>
<tr>
<th>#</th>
<th>Date</th>
<th>Author</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>187a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>285a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>485a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>654a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>655a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>656a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>657a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
<tr>
<td>658a</td>
<td>12/04/2018</td>
<td>Langue1</td>
<td></td>
</tr>
</tbody>
</table>

### Technical staff in infrastructure projects at the research data centre

- Non-technical users in inhouse research projects
- Non-technical users in external research projects

Version Control
(Git, Redmine integration)
Displaying the git repository as a folder on a shared drive
Specific Git Solutions III

Scripts to let users work with git without noticing it

- **Version Control (Git, Redmine integration)**
- **Synchronization Services**
- **“Hidden Versioning”**

- Technical staff in infrastructure projects at the research data centre
- Non-technical users in inhouse research projects
- Non-technical users in external research projects
Automatically Supported Workflows I

Using a plugin in the project management software git to automatically create issues to be carried out.

- Custom Workflows (Redmine)
- Version Control (Git, Redmine integration)
- Synchronization Services
- "Hidden Versioning"
- Technical staff in infrastructure projects at the research data centre
- Non-technical users in inhouse research projects
- Non-technical users in external research projects
Automatically Supported Workflows II

Using a plugin in the project management software git to automatically create issues to be carried out.
Automatically Supported Workflows III

Not only create, but also (partly) carry out the required issues automatically for users in another infrastructure.
Quality Control I

A framework to gather existing checks and fixes in a consistent and reusable way.
Quality Control II

Html error list along with XML error lists that can be opened in the software used to produce the data.
Quality Control III
Quality Control III - Example
Conclusions

- Technical solutions for non-technical users are needed
- Git for Humanities
- Technical support will still be needed for Humanities projects

Adaptability to other data

- Technical support will be needed for the project
- Resources should be versionable using Git
Perspective

- Enhance the hidden versioning
- Make the workflows more open to external projects/users
- Enhance the GUI options
- Adapt Framework to be even more user-friendly and robust
Acknowledgements

**INEL**
https://inel.corpora.uni-hamburg.de/

**Project leader**
Prof. Dr. Beáta Wagner-Nagy (IFUU, Universität Hamburg)

**Applicants**
Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler,
Hanna Hedeland, Timm Lehmberg

**Researchers/Developers**
Dr. Alexandre Arkhipov (Research Coordinator),
Timm Lehmberg (Technical Coordinator),
Dr. Maria Brykina, Chris Lasse Däbritz, Anne Ferger,
Dr. Valentin Gusev, Daniel Jettka, Dr. Svetlana Orlova

**HZSK/CLARIN**
https://corpora.uni-hamburg.de

**Spokesperson/Project leader**
Prof. Dr. Kristin Bührig

**Managing director**
Hanna Hedeland, M.A.

**Deputy Managing Director**
Daniel Jettka, M.A.

**Researchers/Developers**
Tommi Pirinen, PhD
Hanna Hedeland, M.A.
Thank you!

Contact:
anne.ferger@uni-hamburg.de
hanna.hedeland@uni-hamburg.de
inel@uni-hamburg.de
corpora@uni-hamburg.de
Optional additional Information

Links:

• https://corpora.uni-hamburg.de
• https://inel.corpora.uni-hamburg.de
• https://exmaralda.org/en/