Creating a Learner Corpus Infrastructure: Experiences from making learner corpora available

Alexander König (CLARIN ERIC)
Jennifer-Carmen Frey (Eurac Research)

ICTeSSH 2020

1st July 2020

Learner Corpora

- Gathering data from language learners
- Quite a large community (LCA) with their own conference series
- Gives insights into the language acquisition process
- Mostly L2 (learning a 2nd/foreign language)
- ...but sometimes also L1 (measuring development in the native language)

Learner Corpora – The Data

- Often written texts and no audio recordings
- Often not born digital, but manually transcribed
- Heavily annotated, e.g. for errors
- Non-standard language (wich can make use of standard NLP tools problematic)

Learner Corpora – The Problem

- Corpora are highly indivual
- No common standards for either file format or annotations
- Annotation practices are not always well documented
- Even corpora collected by the same (team of) researcher(s) can look very different

The Learner Corpus Infrastructure (LCI)

- Institute for Applied Linguistics at Eurac Research
- Making all their learner corpora available online to the scientific community
- (Try to) harmonize file formats, annotations, licensing
- *Still a work in progress*

LCI - The Approach

- According to the FAIR Principles
- Covering Findability and Accessibility first
- Addressing Interoperability and Reusability by harmonizing
  - file formats,
  - query interfaces
  - annotation standards
  - metadata

LCI – The Data

- **KoKo**: L1, DE, essays, upper secondary schools
- **LEONIDE**: longitudinal, DE/EN/IT, lower secondary school
- **Merlin**: CZ/DE/IT, adults
- **Kolipsi-1**: L2 with additional L1 texts, DE/IT, upper secondary schools
- **Kolipsi-2**: L2, DE/IT, upper secondary school
- **LEKO**: L2, DE/IT, based on Kolipsi-1, annotated for collocations and formulaic language

ERCC Repository

- Data Repository based on CLARIN-DSpace
- Part of the CLARIN Infrastructure
- Covers
  - Findability (OAI-PMH, VLO, OLAC, PIDs)
  - Accessibility (CLARIN Federated Login, clear licensing)

Interoperability/Reusability

All corpora

- will be provided as XML and in Annis format
- can be searched through Annis
- use CMDI metadata to facilitate CLARIN integration
- will have additional metadata on the text and learner level (as CMDI or CSV)

Outlook

- Some corpora are already available, others will follow in the near future
- Learner corpus metadata is being worked on together with CLARIN
- There are also talks with other LC research groups to increase the chance of adoption outside of Eurac
- Provide guides to help others in FAIRifying their learner corpora

Thank you for your attention!

alex@clarin.eu

JenniferCarmen.Frey@eurac.edu

https://porta.eurac.edu