Spoken Data on Corpus Platforms

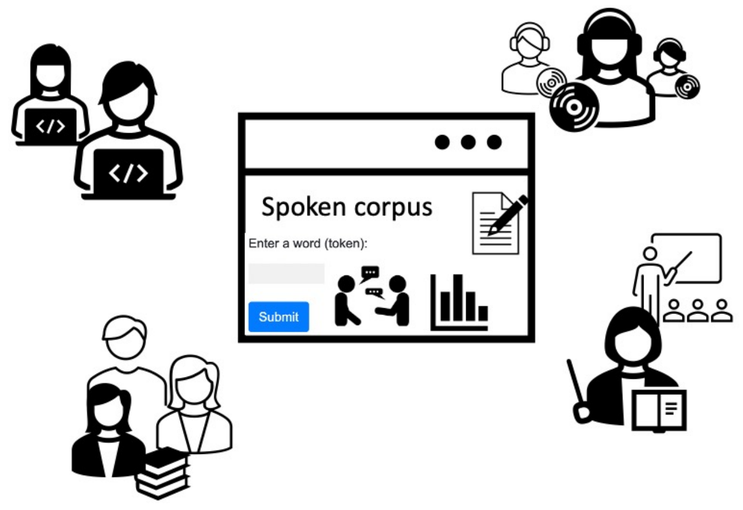
User-specific Views and CLARIN Concordancers

Dolores Lemmenmeier
ZHAW School of Applied Linguistics
leme@zhaw.ch

CLARIN
Common Language Resources and Technology Infrastructure

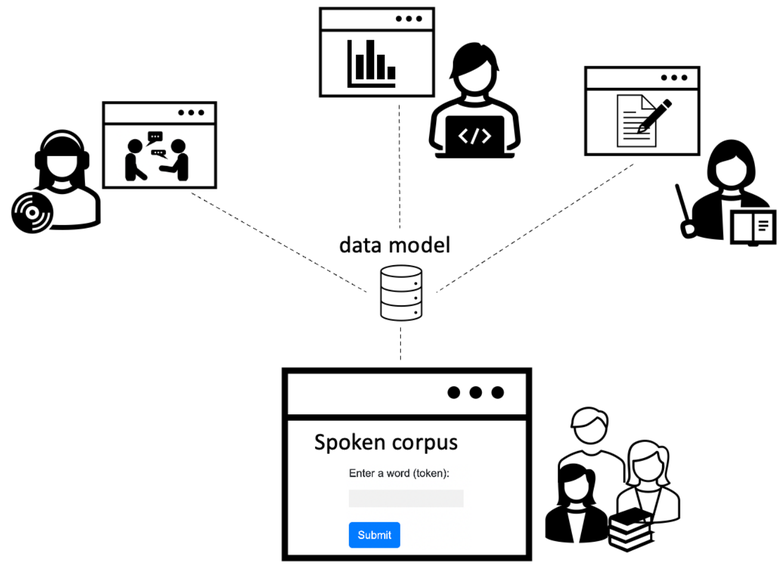
Problem: Presenting Spoken Data on Corpus Platforms is Challenging

- **Heterogeneous user groups**
  - Spoken corpora are used in different disciplines, e.g. by lexicographers, phoneticians, teachers, etc.
- **Different technological affinity of the users**
  - Many users are not familiar with query languages and standards from the field of corpus linguistics.
- **Versatility and usability**
  - Users expect spoken corpora to be multi-functional and user-friendly at the same time.



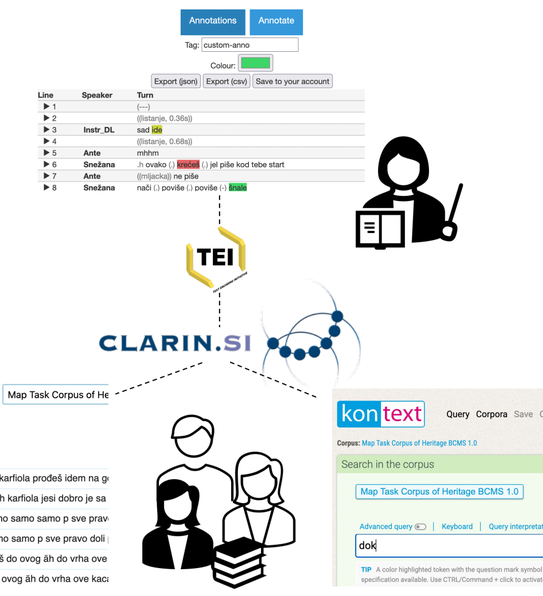
Proposed Solution: Creating User-group Specific Corpus Views

- **Make corpus views easy to understand**
  - Address targeted groups and take into consideration established practices in individual disciplines.
- **Connect custom views to general-purpose platforms**
  - This will allow a wider audience to query the corpus with standard corpus linguistic methods.
- **Use standard transcribing conventions**
  - Following standards such as ISO/TEI guidelines for transcriptions of speech will facilitate interoperability among different corpus tools.



Example: Map Task Corpus of Heritage Bosnian/Croatian/Montenegrin/Serbian

- **Provided on a customised corpus platform**
  - Tailored for teachers and students of BCMS in diaspora: **maptask.slav.uzh.ch**.
  - Provides features such as simple querying, annotating transcripts and sharing annotations.
- **Integrated into CLARIN.SI concordancers noSketch Engine and KonText**
  - Enables exploring the corpus with common corpus linguistic methods for a broader audience.



A European Research Infrastructure

CLARIN ERIC was established in 2012 and received ESFRI Landmark status in 2016

www.clarin.eu

clarin@clarin.eu
@CLARIN.ERIC
github.com/clarin-eric

zhaw Angewandte Linguistik

CLARIN-CH

University of Zurich[UZH]