Corpus Linguistics
Programme

- The concept of CoLABnet
- Doing Corpus Linguistics
- Teaching Corpus Linguistics
- The Corpus Linguistics CoLAB Project
- Concrete Measures
- Discussion and Planning
The CoLABnet concept

- Multi-level and interdisciplinary cooperation
  - Teaching, Training, Research
  - Linguistics, History, Sociology, Computer Science, Mathematics, Literature
The CoLABnet concept

- Integrating teaching and research
  - Following the idea of Humboldt
  - Maintaining eligibility for Erasmus+
  - Adding third-party funding
The CoLABnet concept

• Paving the way for
  • Bigger projects
  • Joint programmes
  • Better collaboration
Why Corpus Linguistics?

- Why doing corpora?
  - No good corpora for ABMCS
- Why doing corpus linguistics?
  - Corpus linguistics becomes more and more important in various linguistic subdisciplines and in computer science
  - Corpus Linguistics as genuine linguistic method
  - Well developed at HU
General corpora

- Hrvatski nacionalni korpus: http://www.hnk.ffzg.hr
- Hrvatska jezična riznica: http://riznica.ihjj.hr/index.hr.html
- Korpus savremenog srpskog jezika:
  http://www.korpus.matf.bg.ac.rs/korpus
- Korpus bosanskih tekstova:
  http://www.tekstlab.uio.no/Bosnian/Corpus.html
- SrpKorp RS: http://korpus.ffuis.edu.ba
Web corpora

- SrWaC: https://reldi.spur.uzh.ch/blog/serbian-web-corpus/
- HrWaC: https://reldi.spur.uzh.ch/blog/croatian-web-corpus/
Parallel corpora

- ParaSol: http://parasolcorpus.org/Ursynow
- Korpus savremenog srpskog jezika: http://www.korpus.matf.bg.ac.rs/korpus
Specialised corpora

- CHILDES Croatian:

- CHILDES Serbian:

- + many not accessible corpora
Spoken corpora

- Croatian Spoken Language Corpus (TalkBank):
  http://talkbank.org/access/CABank/Croatian.html

- Corpus of narratives (DIWNA):
  http://poincare.matf.bg.ac.rs/~andjelkaz/diwna/
Competencies at HU

- Institute for German and Linguistics:
  - Chair for Corpus Linguistics (Prof. Anke Lüdeling)
  - 10+ researchers
  - Leading in the development of several corpus applications
- Institute for Slavic Studies
  - Prof. with strong corpus linguistic focus (Prof. Roland Meyer)
  - 3+ researchers
- Faculty-owned server
Teaching Corpus Linguistics

- Important competencies for students
- Interdisciplinary juncture of several methods and linguistic fields
- One of the planned main topics of SL@HU
- Relatively easy integration of teaching and research through scalability of projects
Teaching Corpus Linguistics

- No specific programme at HU
- No regular courses at the Slavic department
- No (?) regular courses at partner universities
Corpus Ling as CoLABproject

- Dimensions:
  - Teaching – Research – Outcome

- Domains of Action:
  - Tool creation, Corpus design, Material collection, Corpus compilation, Analysis

- Target fields:
  - Linguistics, Sociology, Mathematics, Computer Science, Anthopology, Digital Humanities
• Tool creation:
  • Linguistics, Mathematics, Informatics
  • T: Scripting
  • R: Tools
  • O: Tools for ABCMS
  • Collaboration with Zurich and Belgrad (Math)
• Corpus design:
  • Linguistics, Sociology
  • T: Sociolinguistics
  • R: Corpus design
  • O: Templates
  • Ongoing project funded by HU
• Material sampling:
  • Linguistics, Anthropology, Sociology
  • T: Fieldwork, Collection, Processing
  • R: ...
  • O: Material
  • Use of existing spoken material (from dialectology, anthropology, sociolinguistics etc.)
Corpus compilation:
- Linguistics, Computer Science, Mathematics
- T: Database, Scripting
- R: Processing
- O: Corpus
• Corpus analysis:
  • Linguistics, Sociology, Mathematics, Computer Science, Anthopology
  • T: Statistics, Interpretation, Tools
  • R: ...
  • O: Papers, Theses
Mobility formats

- Workshops for students and staff (train the trainer) on corpus creation and corpus linguistics by HU scholars
- Workshops on corpus linguistics didactics, development of similar or joint courses and harmonized curricula in Berlin
- Workshops on corpus design, corpus creation in Berlin
- Training for staff at HU
- Individual research/Study stay
Concrete measures

- Establish frame project with specific aim (spoken corpus, written corpus, dialect corpus, special corpus, scope, size, ...) and smart design
- Create a common virtual workplace with pre-configured tools and pipelines (Ecco - Project funded by HU)
- Organize mobilities
- Next step: Workshop in 2018 in Berlin