CORLI

a linguistic consortium for corpus, language and interaction

CORLI and HUMA-NUM

- CORLI = Corpus, Languages, and Interaction
  - a French consortium of Huma-Num involved in linguistic research and teaching
- Huma-Num = Humanités Numériques (Digital Humanities)
  - A French governmental organisation that provides resources and tools for digital humanities and social sciences

Organisation of CORLI

- CORLI is a network of laboratories in France
    - Steering committee representing 22 different laboratories
    - The goal is to represent all corpus linguistic research
- Coordination by Franck Neveu (Institute of French Linguistics)
- Created in 2016
    - 5 years experience as
        - Corpus Ecrits
        - IRCOM

Goals

- CORLI is a self-organised network stemming from the linguistic community
  - The project is compiled by the previous steering committee
  - The committee has changed according to the changes and priorities in linguistics and digital technologies
- The project and goal of CORLI are evaluated and validated by the scientific committee of Huma-Num

Organization: General actions

  - Concerns all domains of corpus linguistics

- Describing resources
  - Help people describe their own data
- Evaluation of resources
  - Scientific evaluation
  - Availability of data (technical and rights)
- Technical courses and information
- Finalization of corpora

Technical courses and information

- Corpus annotation tools
  - ELAN, CLAN, SPASS
- Corpus exploration tools
  - TXM (textometric exploration, R queries, interface with CQP)
  - Iramuteq (interface from texts to many R libraries)
- Video and sound recording
  - Getting the best quality video and sound
  - Formats for corpus use
- Metadata
  - Course about coding metadata
  - Using software as provided by repositories or other tools

Finalization of corpora

- Call for small amounts of financial support for finalization of corpora
  - The corpus should be already advanced
  - The corpus should complement the already existing corpora
  - The corpus has to be deposited in an official repository (CLARIN centres)
  - The corpus should be at least open access for research
- 25 submissions to the 2017 call: 13 accepted for a global budget of 40 000 €.

Organization: Workgroups

  - Target a subfield or a specific goal
- Corpus deposition and evaluation
  - Guidelines and technical courses about corpus deposition (how to do it, formats, metadata)
- Legal information
  - Guidelines about legal procedures in France (mostly applicable in Europe next year)
- Multilingual / plurilingual
- Exploration and formats
- Multimodality and new form of communication

Workgroup: Exploration and formats

- Exploration
  - Tools for analysing and processing corpora
  - What tools exist, which formats do they use, how data can be prepared
- Formats
  - Metadata
    - Description of a minimal set of metadata for the analysis of oral language corpora
      - Guidelines for coding the metadata
    - Work planned with the same objectives for written data
  - TEI for Oral language
    - Coding all major transcription formats using the TEI structure (CLAN, ELAN, Transcriber, Praat)
  - Tools for metadata and format conversion

Development of tools for shared formats

  - Development is shared with Ortolang: http://ct3.ortolang.fr
- Conversion tool
  - Application formats: Clan, Elan, Praat, Transcriber -> TEI (and back)
  - No data lost in conversion to TEI and back to the same format
  - Minimal data lost in conversion between application formats
- Developing a tool for easy metadata editing
  - Tool is web-based - run in standalone page (no server) or local application
  - Tool is based on XML models that can described and produce any XML format
    - Models allow to write user help and full description of the format to be coded by the user
  - Edit only specific nodes and leave other data unchanged

Workgroup: Multimodality and new forms of communication

- Development of cutting-edge practices
  - human interaction
    - gesture, visual languages, co-verbal communication
  - computer-mediated communication and social media corpora
- New specific tools and formats (or the extension of previous tools and formats)
- Visualisation of data
- Information about corpus creation and tools

Integration into CLARIN

- Centres
  - Two C-Centres (Cocoon, SLDR)
  - One planed B-Centre (Ortolang)
    - Ortolang metadata already harvested by VLO
  - Corli as a K-Centre (might included participation of other Huma-Num consortia: Cahiers, Ethnology)
- Metadata
  - Metadata inner format (CMDI, Olac) is handled by the repositories or specific tools
    - Not by the user who should not be bothered by technical specificities
    - Olac is widely used in France, integration to CLARIN will be done using automatic conversion of existing data
  - General politics about formats is to favour easy conversion between formats
    - Metadata has to be in an official organized format
- All data should be deposited in Clarin Centers (or candidate centres)
- Integration as France as an observer implies participation of CORLI members in CLARIN
- All data should at least be open access for research
  - Most people want CC-BY-NC licence
  - Usually this is the most restrictive licence effectively used