Towards Interoperablity of Parliamentary data and tools

ParlaCLARIN panel, LREC 2018

Jan Odijk 2018-05-07 Mizakaya, Japan

Background

- CLARIN-NL and CLARIAH projects
  - WIP (War in Parliament): http://portal.clarin.nl/node/1923
  - CLARIN Travelling Campus: Talk of Europe
    - Data Curation:
      - Europarl as Linked Open Data in RDF format with metadata made explicit
      - https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:62227
      - CMDI metadata, but not on a harvestable place
    - 3 `Creative Camps' (1 week, working on an application / service around the data)
- (with Dutch Language Institute) Constant stream of parliamentary data for lexicographic/linguistic research and for language mdelling en MT (in the context of the ELRC project)

Role for CLARIN

- Current and recent workshops
  - Focus on inventory, visibility, findability, accessibility
- CLARIN should do more, in particular *Interoperability*
  - Not easy because there are many different uses and user groups:
    - Research: linguists, computational linguists, speech technologist, historians, political scientists
    - Society: journalists, political analysts, politicians, general public

Role for CLARIN

- Formats in current workshop:
- `XML’: each is different
- `TEI’: many subtypes and even a modified TEI

Row Labels	Sum of count
TXM	1
'a reusable format'	1
CONLL-U	1
Linked Data (RDF)	1
WPL (vertical)	1
unspecified annotation format	1
CWB	2
unspecified metadata format	2
XML	6
TEI	6
unspecified	16
Grand Total	**38**

Interoperability

- Centrally
  - Determine (limited set of) preferred formats
    - Exchange formats and `live' formats
    - Should ideally be easily extensible with new annotations and data (audio, video)
  - Create converters between preferred formats
- Have national consortia create
  - Converters from legacy formats
  - Applications/ services that operate on the preferred formats
    - Targeted at the intended users (HSS researchers but also NLP researchers)
  - New data that comply with the preferred formats
- Will only be successful if CLARIN can offer the best tool /application set (but that should be feasible with > 20 member countries)

Thanks for your attention!