Towards a tool catalogue: some ideas

Dieter Van Uytvanck

Centre Meeting 2024

12 June 2024

CLARIN

Overview (1)

- “Where can I find all CLARIN tools/services”?
- Some attempts (eg https://www.clarin.eu/content/tools) to have lists of nice examples, but non-exhaustive

Visit the Switchboard →

Tools Per Centre

ASV Leipzig

ASV Toolbox is a modular collection of tools for the exploration of written language data. The topics covered contain language detection, POS-tagging, base form reduction, named entity recognition, and terminology extraction. On a more abstract level, the algorithms deal with various kinds of word similarity, using pattern based and statistical approaches. The collection can be used to work on large real world data sets as well as for studying the underlying algorithms.

Visit ASV Toolbox →

ARCHE

Featured example: Viennese Lexicographic Editor

Visit tool portal →

Overview (2)

- Many centres and consortia did their homework and provided tool metadata:
  - In their repositories (example)
  - In the Language Resource Switchboard (example)
  - In the ~~EOSC and/or~~ SSHOC marketplace
  - In some (national) infrastructures like e.g.
    - https://tools.clariah.nl/
    - The WebLicht tool metadata

Overview (3)

- Besides that, some manually curated lists exist, eg under the Language Resource Families:
  - Corpus Query Tools
  - Normalisation
  - Named Entity Recognition
  - Part-of-Speech Tagging and Lemmatisation
  - Tools for Sentiment Analysis

Ideas and wishes

- Have a single source of information to search through all CLARIN tools
- Use the metadata of those who did their homework
- Provide an easy way for others and newcomers to do their homework
- Re-use the software that we have developed already
- Make this a practical exercise, not a theoretical/political one
  - Use what we have, right now!

Ad-hoc concept (1)

- **Using the VLO as catalogue** for tool records, relying on:
  - Metadata from the repositories (Resource Type = “software, webservice”)
    - See VLO link for examples (n = 1239)
    - Many centres are in, no false positives, relevant entries
    - Not only DSpace repositories
  - Advantages:
    - Comes with de-duplication
      - e.g. double entries repository <> resource family
    - Comes with linkchecking
    - Comes with issue reporting

Current VLO results for Resource Type = “software, webservice”

- Assamese NLP Resources (1)
- Center of Estonian Language Resources (57)
- CLARIN:EL Catalogue (47)
- CLARIN.SI data & tools (122)
- CLARIN-DK-UCPH Repository (1)
- Clarin IS (112)
- CLARIN-LT (12)
- Clarino - Textlab (4)
- CLARIN-PL (119)
- ILC4CLARIN : ILC Data & Tools (15)
- ILC4CLARIN : OPEN Data & Tools (2)
- Language resources and tools of AiLab IMCS UL (6)
- LINDAT / CLARIAH-CZ Data & Tools (119)
- LRT + Open Submissions Data & Tools (172)
- ORTOLANG Repository (37)
- PORTULAN CLARIN (117)
- SADiLaR (81)
- SADiLaR Resource Catalogue (41)
- SADiLaR Resource Index (78)
- Språkbanken (1)
- Språkbanken NB (5)

Ad-hoc concept (2)

- **Getting the other sources into the VLO**
  - SSH Open Marketplace (n = 190)
    - has the CLARIN tools resource families in JSON
    - a bit more tedious to perform the conversion
    - but well-curated
  - CLARIAH-NL tools by Maarten van Gompel (n = 114)
    - Acts as an aggregator for all tools from NL, based on CodeMeta
    - Very thoroughly designed and implemented
    - Rich metadata (where available from the source) in JSON
  - Language Resource Switchboard entries, converted to CMDI (n = 66)
    - Fairly straightforward conversion (very basic JSON to CMDI)
  - Weblicht web services (n = 51)
    - Available as technical CMDI (=machine consumption)
    - Could use some small additions to make it useful for humans (eg via XSLT)

/etc

- Which CMDI Profile?
  - Probably the DSpace one
- Many tools (e.g. ELAN, Praat) are already in multiple sources (eg SSHOM, LRT inventory and the Finnish language bank)
  - Missing tools can be added via any of the supported “sources”
  - Try to come up with a flowchart for newcomers
- Sheet with some more details and links, open for comments

Possible steps forward

- “Today”: add the VLO link to the Tools web page
- “Over the summer”: first JSON to CMDI conversion:
  - Switchboard JSON (easy to parse/process)
  - SSH Open Marketplace
  - CLARIAH-NL tools