The Swedish–Finnish Korp collaboration – past, present and future

Johan Roxendal and Jyrki Niemi

Språkbanken and FIN-CLARIN

Interaction design in the context of CLARIN
Nordic CLARIN Network workshop
Gothenburg, 3–4 May 2017

Corpus interface for the Language Bank of Finland

- Developing software of your own **may** make it more likely that you get the features you wish . . .
  - . . . **if** you have enough time, money, people and skills
- FIN-CLARIN did not have the resources, so we investigated available software
  - Corpus Workbench (CWB), Manatee/Bonito, Poliqarp, . . .
- We first considered Glossa (of U Oslo) as a CWB user interface
- Then Korp came around, and we adopted it in summer 2012

Språk BANKEN FIN-CLARIN

Assessing Korp for the Language Bank of Finland

- Advantages
  - Free and open-source
  - Actively developed, not only by a single person
  - Modern, user-friendly look and feel
  - Uses CWB, which works
  - Separate Web service (backend) and user interface (frontend)
  - Existing contacts between FIN-CLARIN and Språkbanken
- Disadvantages
  - Lacked features we would like to have (and still lacks some)
    - E.g., authentication; collocates with different statistical measures
  - Some features are specific to Swedish corpora or annotations
  - CWB is oldish and has its limitations
    - E.g., handling XML and nested structures

Språk BANKEN FIN-CLARIN

What is in it for Språkbanken?

- Advantages
  - Getting users is always good
  - Bug reports
  - Feature requests
- Disadvantages
  - It takes time to prepare a project for distribution
  - Requires writing distribution documentation
    – but that is also a good thing

Språk BANKEN FIN-CLARIN

Korp in the Language Bank of Finland

- Configured for the corpora in Finland
- Slightly tailored to our corpus annotations
- Some annotations are mapped to those expected by Korp
- A few added features
  - Specific to the Finnish site
    - Standardized URNs, links to metadata, applying access rights
    - CLARIN licence categories for restricted corpora
  - More general
    - Shibboleth authentication
    - KWIC download (export)
    - Name classification picture (to be rewritten)
    - Search only in sentences containing given words or lemmas
    - Other minor features
- We use the Git version control system to keep track of changes
- Currently lags about 1,5 years behind Språkbanken's Korp
- In practice, a fork of Språkbanken's Korp

Språk BANKEN FIN-CLARIN

Problems with forking

- Upstream (origin) = Språkbanken, downstream = FIN-CLARIN
- Even though Git helps, it takes time to merge the features added in FIN-CLARIN into a new release of Korp
  $\Rightarrow$ Easy to fall behind the upstream
- FIN-CLARIN features should be offered to Språkbanken
  - If merged upstream, merging downstream the next time should be easier
- What about features not merged upstream?
  - Maybe convert them to some kind of plugins if possible
    - Would require architectural support
  - Otherwise merging the features to new upstream releases may require source code changes
    $\Rightarrow$ We lose track of what code (commits) belongs to what feature
- To accept or not to accept a pull request?

Språk BANKEN FIN-CLARIN

Korp collaboration in the past and present

- Testing and reporting bugs, via email and Trello
- Requesting for information on (corpus) configuration
- Requesting features, via email and Trello
- Face-to-face meetings: Korp workshop and Korpathon
- A method for sharing FIN-CLARIN code with Språkbanken using pull requests
  - Not used in practice after testing

Språk BANKEN FIN-CLARIN

Goals and wishes for future Korp collaboration

- More communication via email, Trello, GitHub, possibly video conferencing
- Routinely discuss issues before starting to work on them in order to avoid duplicate work
- Inform well in advance of new features requiring changes to corpora
- Share FIN-CLARIN's code with Språkbanken via GitHub's pull requests
- Adopt faster new features (Språkbanken's new Korp releases) in the Language Bank of Finland
- Feedback on feature requests
- Regression tests for new features

Språk BANKEN FIN-CLARIN

Thank you!

- Questions?

Facilitating collaboration in software projects

- What may make others use your software
  - Open-source
  - Easy to install and configure with good instructions
    - For corpus tools, document requirements of corpus format
  - Preferably portable
  - Good browser and device support (for a Web application)
  - Designed with localization in mind
  - Features specific to a site, language, corpus or corpus annotation scheme separated from the common part
    - For example, should not assume a certain annotation scheme
- How to support others tailoring your software
  - Code well documented or self-documenting
  - Relatively mainstream programming languages, tools, frameworks
  - A plugin architecture of some kind
- How to make it easier for others to contribute
  - Coding style guidelines to keep style consistent
  - Clean interfaces for adding new features
- Social factors
  - It may sometimes help to meet face-to-face

Språk BANKEN FIN-CLARIN