INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Mylly - The Mill: A new platform for processing speech and text corpora easily and efficiently

Mietta Lennes[1], *Jussi Piitulainen*[1], *Martin Matthiesen*[2]

[1]University of Helsinki, Finland
[2]CSC - IT Center for Science, Finland

mietta.lennes@helsinki.fi, jussi.piitulainen@helsinki.fi, martin.matthiesen@csc.fi

Abstract

Speech and language researchers need to manage and analyze increasing quantities of material. Various tools are available for various stages of the work, but they often require the researcher to use different interfaces and to convert the output from each tool into suitable input for the next one.

The Language Bank of Finland (Kielipankki) is developing an on-line platform called Mylly for processing speech and language data in a graphical user interface that integrates different tools into a single workflow. Mylly provides tools and computational resources for processing material and for the inspecting the results. The tools plugged into Mylly include a parser, morphological analyzers, generic finite-state technology, and a speech recognizer. Users can upload data and download any intermediate results in the tool chain. Mylly runs on CSC's Taito cluster and is an instance of the Chipster platform. Access rights to Mylly are given for academic use.

The Language Bank of Finland is a collection of corpora, tools and other services maintained by FIN-CLARIN, a consortium of Finnish universities and research organizations coordinated by the University of Helsinki. The technological infrastructure for the Language Bank of Finland is provided by CSC – IT Center for Science.

Index Terms: speech annotation, speech analysis, automatic speech recognition, speech corpora, workflow

1. Introduction

The Language Bank of Finland (https://www.kielipankki.fi) provides many useful tools for researchers who use language materials. These tools include, e. g. automatic parsers of text and the automatic speech recognizer AaltoASR [1].

Since many technological tools are work in progress, they do not always include a user interface that would be easily accessible for researchers and students working in fields of Humanities or Social Sciences. The tools under development are usually command-line applications that require the user to be familiar with Unix commands and their parameters.

Chipster (http://chipster.csc.fi, [2]) is a graphical interface for analysis workflows. Chipster has been previously available for researchers in, e.g., bioinformatics. Mylly is a fresh instance of the Chipster technology supplemented with various tools that are available in the Language Bank of Finland. The Chipster interface is currently provided as a Java WebStart application.

2. Workflow

Mylly provides a Chipster interface for applying Kielipankki tools. An example view of the Mylly interface is shown in Figure 1. The current session is shown as a graph that links each file to the files from which it was made. Mylly also provides a log on the creation history of each file. The contents of certain kinds of files can be inspected directly in the GUI, others can be opened in a local browser, and all of them can be downloaded.

Users log in to their own Mylly sessions. They can upload their own files, open a saved session, or use the tools in Mylly to acquire data. Input files and a tool can then be selected, parameters set, and the tool launched to run on the server. The result files appear as new nodes in the GUI as soon as they are ready.

We have implemented a mechanism that sends a job from Mylly to the batch processing queues on the server system. This way resources are allocated to the job as they become available in the cluster, instead of competing with interactive users. For instance, the speech recognition tools are set to run this way.

CSC is working to replace Java with HTML5, which will make the interface easier to access in a regular web browser.

3. Current tools

A number of morphological analyzers and other processing tools from the HFST (Helsinki Finite-State Technology [3]) project are currently available via Mylly. The extensive OMorFi [4] lexical transducers for Finnish as well as more general finite-state tools for building and inspecting finite-state transducers are also available.

The AaltoASR software package [1] can be used either for recognizing speech or for aligning manually created transcripts to audio. The quality of the result is dependent on whether the speaker and the speaking style are a good match with the acoustic speaker model and the language model that are being applied by the automatic recognizer. In AaltoASR, it is possible to perform the recognition process in stages. The user may manually check a part of the transcript and use the corrected material for training a new, improved recognition model, which can then be applied in a new recognition attempt. This semi-automatic method can be useful especially in transcribing larger corpora.

Using Mylly, the user is no longer required to run tools from the command line, since the same tools can be applied by selecting the corresponding options from the menu within Mylly. In Mylly, the entire workflow becomes easier to manage. The command line tools are still available to those who prefer a more flexible environment.

4. Future prospects

The set of tools that is available via Mylly continues to increase. In the future, Mylly will provide simple tools to extract text from ordinary document formats in tokenized and segmented forms that are suited for further processing in Mylly. Tools will be provided for querying the corpora available through the Korp interface (https://korp.csc.fi). Statistical analysis and visual representation tools that can be applied on standard kinds of data

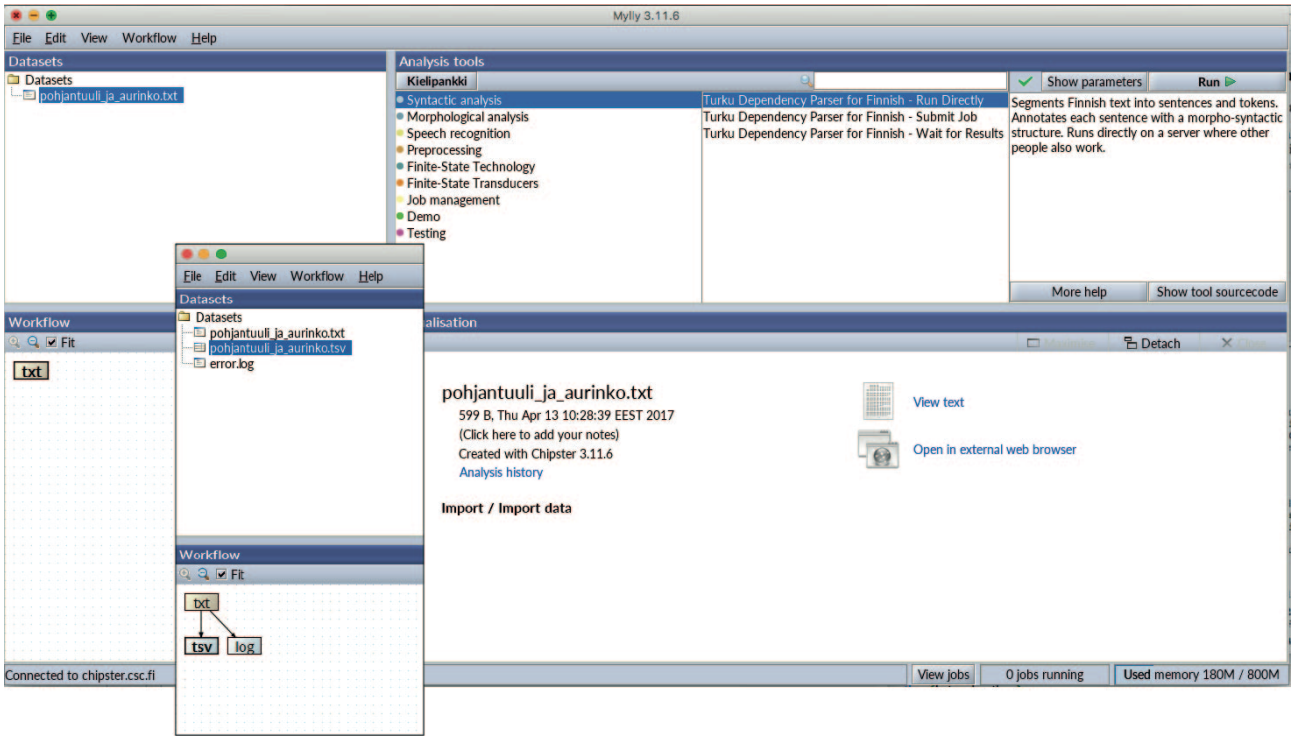

Figure 1: *A view of the Mylly interface. In this example, a text file was imported (see the top left panel) and processed with the Turku Dependency Parser (see the menus on the top right panel). The complete workflow is visualized on the bottom left panel. The resulting files are shown in the partial screenshot over the screenshot of the initial stage.*

will also be available. We expect Mylly to become a viable tool in teaching and research.

5. Conclusions

Tools with web based interfaces are often relatively easy to use but they may not be easily adapted to the particular demands of a specific research problem. On the other hand, command-line tools or "raw" scripts that run in a Linux-based High Performance Computing (HPC) environment are very flexible but they have a high learning curve. Mylly, the Chipster instance of the Language Bank of Finland, fills the gap between these alternatives.

Mylly enables researchers to manage, save and share complete research workflows that can include many intermediate stages and apply a number of different tools on speech and language material and data. Mylly helps users keep track of their methods and makes the research process easier to describe and to replicate. The instructions for using Mylly are updated at the website of the Language Bank of Finland (https://www.kielipankki.fi/support/mylly/).

6. References

[1] "AaltoASR - Aalto University Automatic Speech Recognition System," [Software] Available in the Language Bank of Finland, 2017, http://urn.fi/urn:nbn:fi:lb-2014091904.

[2] M. A. Kallio, J. T. Tuimala, T. Hupponen, P. Klemelä, M. Gentile, I. Scheinin, M. Koski, J. Käki, and E. I. Korpelainen, "Chipster: user-friendly analysis software for microarray and other high-throughput data," *BMC Genomics*, vol. 12, no. 507, 2011.

[3] "HFST - Helsinki Finite-State Transducer Technology," [Software] Available in the Language Bank of Finland, http://urn.fi/urn:nbn:fi:lb-20140730183.

[4] "OMorFi - Open Morphology for Finnish," [Software] Available in the Language Bank of Finland, http://urn.fi/urn:nbn:fi:lb-20140730127.