CONFERENCE REPORT

Language technology through citizen science
19 August 2016, Oulu, Finland

Péter Koczka

The workshop titled Language technology through citizen science was organized within the 12th Congress for Finno-Ugric Studies held in Oulu in 2016 (for more details see the conference report on the Congress in the present volume). Following the First International Workshop on Computational Linguistics for Uralic Languages held in January 2015 in Tromsø, the section had mostly familiar faces. Its aim was to bring together researchers working on computational approaches to working with the Finno-Ugric languages. All of them exhibit rich morphological structure, which makes processing them challenging for state-of-the-art computational linguistic approaches, many are endangered and the majority also suffer from a lack of resources. That is why the varying subjects and project presentations never failed to mention the importance of open sourcing tools and making other resources publicly available.

The series of ten presentations started with Jeremy Bradley from the Ludwig Maximilian University of Munich, who introduced the term "nerdview" to the audience, by which he meant that open source or open access material does not necessarily provide access to all audiences if the user interface is not easy to use. That is the reason behind his newly fashioned web site presented here, the Mari Web Project.

Next, Jussi-Pekka Hakkarainen from the National Library of Finland presented the Digitization Project of Kindred Languages, in which the main goal is to create and digitize materials in the Uralic languages as well as develop tools to support linguistic research and citizen science. Through the project, researchers will gain access to new corpora to which all users will have open access regardless of their place of residence. The project's objective is to make sure that the new corpora are made available for the open and interactive use of both the academic community and the language communities as a whole. Since the Uralic related texts consists of around 200 000 pages, "nichesourcing" was used, a specific type of crowdsourcing where tasks are distributed amongst a small crowd of scientists.

Tommi Jauhiainen and Heidi Jauhiainen from the University of Helsinki described a tool under development which aims to collect webpages written in Uralic languages (except for Finnish, Estonian and Hungarian) and publish the links on a portal for researchers. To achieve this, the Heritrix open source web crawler is being used with addition of a language identifier trained for 350 languages (of which 34 are Uralic) to filter out the needed webpages. The automated crawling and filtering system is also intended to build sentence, clause and word corpora.

The Aanaar Saami e-lexicography presentation by the members of the University of Tromsø (Trosterud et. al) shed light on the odd situation of a language with only approximately 450 speakers, yet with a rich lexicographic tradition but without any electronic dictionaries. To create the desired e-dictionaries (Aanaar Saami-Finnish and North Saami-Aanaar Saami), the creation of a North Saami-Aanaar Saami transfer lexicon was necessary, which was achieved by combining two dictionaries (North Saami-Finnish and Aanaar Saami-Finnish) and pivoting via Finnish. The e-dictionaries can serve

Finno-Ugric Languages and Linguistics Vol. 4. No. 1-2. (2015), 60-61. http://full.btk.ppke.hu
ISSN: 2063-8825

as helping tools for language learners, as facilities for lexicon research and practical lexicography. Notably, they plan to use it as the lexicographical foundation for machine translation programs for Aanaar Saami and for further language revitalisation work in written Aanaar Saami.

Kanner et al. introduced the possibilities offered by virtual wiki platforms in a way to create terminology and lexicography work. This is applied in two projects, the Bank of Finnish Terms in Arts and Sciences (BFT) and the lexicographical wiki platform. The Bank of Finnish Terminology in Arts and Sciences (BFT, tieteentermipankki.fi) is a database of Finnish scholarly and scientific terminology for all academic disciplines practiced in Finland. The BFT is maintained by limited crowdsourcing using wiki software, it consists of a multilingual extensive terminology and it is freely available to all researchers. Their future plans include opening the interface and its contents in English and, for example, in kindred languages of Finnish as well.

Tommi Pirinen from the Dublin City University explained his project, Omorfi. The tool is a freely available open source analyser with a lexical database that consists of all sorts of lexicographical information usable for large variety of computational linguistics and general applications requiring processing of Finnish word-forms in context. The data in the database is sourced from the Research Institute of Languages in Finland as the Nykysuomen sanalista, from open source project by language enthusiast engineers from the Joukahainen project and data from the massively crowd-sourced Wiktionary project. The Omorfi tool can be used for more than morphological analysis of Finnish, it could be the engine behind user-facing applications such as spellcheckers and other tools not solely designed for linguists.

Sven-Erik Soosaar from the Institute of the Estonian Language presented language technology tools developed for Tundra Nenets, which include spell checking and e-learning tools. Most of the aforementioned applications are hosted in the University of Tromsø's Oahpa! environment. The word analyser and generator are developed in twolc, but the dialectal variability should be taken into account, since spelling rules of Nenets are not strict and dialectal pronunciation is reflected in written language, which appears to be a serious obstacle along with the very limited amount of machine readable texts available in the language.

Jeremy Bradley described what benefits the computational linguistic tools and complex annotated corpora can offer for linguistic research. Presenting this, the usage of participial verbal forms in postpositional constructions, a well-documented feature of Mari and many of its neighbours, was chosen.

As a closing of the day, Rueter et al. introduced the development of a sandbox for open transducer technology. The objective is to raise the general public's awareness of such tools' existence, especially open-source ones, such as the Helsinki Finite-State Transducer (HFTS). However, being aware of certain tools does not mean they are available to everyone. Teachers can face administrative difficulties when trying to introduce something new, or they are simply not allowed to install applications in their environments. To solve this issue, the sandbox under development will be available on-line, accessible through any web browser. This way transducers can be created and used without installing the slightest bit of extra software.

Péter Koczka
Research Institute for Linguistics, Hungarian Academy of Sciences
mg.peter.koczka@gmail.com