Report on the Second International Workshop on Computational Linguistics for Uralic Languages

Tommi A. Pirinen, Eszter Simon, Francis M. Tyers, Veronika Vincze

The Second International Workshop on Computational Linguistics for Uralic Languages (SIWCLUL) was held in Szeged in January 2016. The goals of the conference series include increased co-operation between the researchers, universities and research centres working on Uralic languages. The event gathered a number of participants from all over Eurasia, including Finland, Hungary, Estonia, Ireland, Germany, Austria and Norway among others. The conference also marked a start of an *Association for Computational Linguistics' Special Interest Group for Uralic Languages* (ACL SIGUR).
Keywords: *Finno-Ugric Languages and Linguistics, Computational linguistics*

1 Introduction

The Second International Workshop on Computational Linguistics for Uralic Languages was held in Szeged, Hungary, on 20 January 2016. The objective of the workshop was to bring together researchers working on computational approaches to working with the following languages: Finnish, Hungarian, Estonian, Voru, Setu, the Sámi languages, Komi (Zyrian, Permyak), Mordvin (Erzya, Moksha), Mari (Hill, Meadow), Udmurt, Nenets (Tundra, Forest), Enets (Tundra, Forest), Nganasan, Selkup, Mansi, Khanty, Veps, Karelian, Ingrian (Izhorian), Votic, Livonian, Ludic, Kven and other related languages.

The first edition of the workshop was held in Tromsø, Norway, in January 2015. This series of workshops is a new attempt to gather researchers of Uralic computational linguistics together, to ensure that they work towards common goals with a minimal amount of overlapping and redundant work. To that effect, the conference series has also formed a new special interest group under the guidance of the Association for Computational Linguistics (ACL).

Two organisers of the first event were also organisers of the second one, thus guaranteeing the continuity between the parts of the series. Local organisers were researchers from the University of Szeged and from the Research Institute for Linguistics of the Hungarian Academy of Sciences.

Original, substantial and unpublished papers were solicited that describe work-in-progress systems, frameworks, standards and evaluation schemes. Additionally, demos and tutorials were also invited which present systems and standards that pursue the goal of interoperability and unification of different projects, applications and research groups. The topics in which papers were expected are: parsers, analysers and processing pipelines of Uralic languages; lexical databases, electronic dictionaries; finished end-user applications

aimed at Uralic languages, such as spelling or grammar checkers, machine translation or speech processing; evaluation methods and gold standards, tagged corpora, treebanks; reports on language-independent or unsupervised methods as applied to Uralic languages; surveys and review articles on subjects related to computational linguistics for one or more Uralic languages; any work that aims at combining efforts and reducing duplication of work; and proposals concerning how to elicit activity from the language community, agitation campaigns, games with a purpose. To maximise the possibility of reproducibility, replication and reuse, submissions that present free/open-source language resources and make use of free/open-source software were particularly encouraged.

One of the aims of this gathering is to avoid unnecessary duplicated work in the field of Uralistics by establishing connections and interoperability standards between researchers and research groups working at different sites. It is now recognised as a serious problem that there is a lack of gold standards and evaluation metrics covering all Uralic languages including those with national support, thus any work towards better resources in these fields were greatly appreciated.

There were 10 accepted papers, 4 of which were presented as oral presentations in two sessions, while the others were poster presentations and/or interactive demonstrations. Additionally, two tutorials were included in a separate session. The topics and languages discussed in the workshop were wide and varied. This year the conference featured a state-of-the-art introduction to Estonian language technology resources. As one of the aims of the workshop series is to promote interoperability between the related Uralic languages, multiple presentations and two tutorials were held to highlight best common practices in the fields of computational linguistics intersecting software engineering.

The workshop gathered 28 scholars from 8 countries including Hungary, Estonia and Finland, where the national language belongs to the Uralic family, and countries such as Russia and Norway, where several Uralic languages are spoken as minority languages.

After a short opening, the first presentation was given by the invited lecturer, András Kornai. In a poster boaster session, each participant whose paper was accepted for poster presentation or demonstration had a few minutes to introduce his/her poster's topic. This was followed by the poster and demo session, and two sessions for oral presentations. In the afternoon, there was another poster and demo session, followed by two tutorials, while the event was closed with a SIGUR meeting and some closing remarks.

Below we report on the presentations and posters under thematic schemes: While Section 2 gives a brief overview of the presentations and discussions under the topic of best common practices, in Section 3 we introduce language-specific resources presented in the workshop. In Section 4 we describe our efforts to form a special interest group and possible related activities. Section 5 presents a sort of a desiderata for future revisions of the conference and pan-Uralic co-operation.

2 Best Common Practices in Uralic Computational Linguistics

The invited talk was given by András Kornai. He is full professor at the Budapest University of Technology, senior scientific advisor at the Computer and Automation Research Institute of the Hungarian Academy of Sciences, and the leader of the mathematical linguistics

research group in the Research Institute for Linguistics of the Hungarian Academy of Sciences. In his talk, entitled *Computational linguistics of borderline vital languages in the Uralic family*, he applied the methodology of Kornai (2013) to the Uralic family with the specific goal of *triage*, to help the community decide where the effort is best placed. As in battlefield triage, where the relatively lightly wounded and the very heavily wounded are treated last, he suggested to direct the very limited resources of the computational linguistics community towards the middle class of borderline languages where neither vital nor still/heritage status can be established.

Thierry Poibeau and Svetlana Toldova had a poster which presented some preliminary experiments concerning the automatic processing of Finno-Ugric languages. They presented symbolic methods as well as machine learning ones. Given the lack of corpora for some languages, they found that finite state transducers may sometimes be the best approach, even if machine learning techniques are supposed to outperform symbolic methods.

Kristian Kankainen demonstrated his tool, Minority Translate, which streamlines the process of creating, editing and saving new articles in any language edition of Wikipedia, also the new language editions starting out in the Incubator. Wikipedia can be treated as a language resource in itself for the lesser resourced languages, as well as a source of several other language technology tools.

Johannes Dellert introduced a new method for inducing a language contact model from lexical data. Based on automatically gathered and manually annotated sets of etymologically related words, the method analyses possible paths of borrowing in terms of lexical flow. In an evaluation on a large lexical database comprising 1,016 concepts across 26 Uralic languages and 18 neighbouring languages, the method detected and correctly inferred the directionality of many instances of cross-family language contact.

Francis Tyers and Tommi Pirinen reported their experience with regard to interoperability of the Uralic languages' practices and tagging standards when used in the context of rule-based machine translation. The Uralic languages exhibit certain resemblances: many of them have similar case inventories, word order and non-finite clause forms. However, current rule-based grammatical resources take many different approaches to encoding this information. In their presentation, Tyers and Pirinen provided some guidelines and suggestions to facilitate future work in the direction of interoperability.

In the tutorials session, first Trond Trosterud presented the language resource repository from the University of Tromsø *Giellatekno*[1] group with best common practices in rule-based open source natural language processing resources. Afterwards Veronika Vincze and Francis Tyers presented the *Universal Dependencies*[2] annotation scheme, which is on track to become an international de facto standard for part-of-speech tagging and dependency annotations.

[1] http://giellatekno.uit.no

[2] http://universaldependencies.org/

3 Language-specific Resources

Jeremy Bradley had a poster presentation in which he introduced his efforts to create a web-based automatic transcription and transliteration software for Uralic and non-Uralic languages. For four literary standards – Meadow Mari, Hill Mari, Russian, and Tatar – an operational interface can be found at `transcribe.mari-language.com`. His poster detailed many of the fine aspects of writing systems used for (Meadow) Mari that he had to take into consideration when creating transcription mechanisms for that language.

Trond Trosterud presented their common research with his colleagues: Lene Antonsen, Marja-Liisa Olthuis and Erika Sarivaara. Their poster, entitled *Modelling the Inari Sámi morphophonology as a finite state transducer*, presented a set of morphophonological problems coming up when when they were working on a transducer for Inari Sámi, a language with a complex and not very well documented morphophonology. As they said: modelling the grammar as a finite state transducer gives more insight into the Inari Sámi morphophonology, and the resulting program will be the foundation of all future Inari Sámi language technology applications.

Tommi Pirinen and his colleagues, Antonio Toral and Raphael Rubino, reported on experiments with Finnish-English statistical machine translation. They jointly used rule-based and unsupervised approaches to segmentation. They found that in terms of automatic metrics, the best system is the one that combines both rule-based and unsupervised segmentations, while human evaluation shows that the outputs produced by a statistical machine translation system with rule-based segmentations are preferred over those of the system that uses unsupervised segmentations.

Axel Wisiorek and Zsófia Schön presented their poster on an Ob-Ugric database, a web-based framework for the storage and advanced retrieval of annotated corpora and corpus-based lexical databases of Khanty and Mansi dialects. The database building is a work in progress within the framework of the project titled *Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects (OUDB)*.

Peter Smit presented a generic model for automatic speech recognisiton applied to Northern Sámi as an example for a setup of lesser resourced languages. Since the lack of technology and applications may threaten the existence of these languages, it is important to study how to create speech recognizers with minimal effort and low resources.

Kadri Muischnek and her colleagues at the University of Tartu gave an overview of the state of the art of tools and resources for the syntactic analysis of Estonian. They presented a manually annotated dependency treebank containing 400,000 words. A morpho-syntactic disambiguator, a shallow parser and a dependency parser were also introduced, all of which are based on the Constraint Grammar formalism.

4 SIGUR

In the first workshop, there was a consensus that computational linguists working with Uralic languages should organise themselves in the form of an ACL-approved special interest group. The organisers then negotiated the founding of a group with the ACL secretary and the SIG officer. After affirmation from the ACL meeting, this second workshop's

business meeting was used as a co-ordinated founding meeting for the newly created SIG. The business meeting was attended by the participants who were also members of the SIG and decided upon the board and the founding papers of the new SIG. The details of the special interest group can also be found on the SIG website,[3] which also includes the public minutes of the meeting.

5 Future plans and desiderata

In the formal meeting it was decided that the workshop series should carry on and plans were set for the forthcoming course of action, including the next workshop potentially to be organised in St. Petersburg. One of the rationales behind wishing to hold the next workshop in Russia is to increase co-operation with researchers in Russia. As the majority of Uralic languages are situated in Russia, there is a vast amount of ongoing research and resources that are of interest to workshop-goers and researchers.

The newly formed special interest group will take an active role in co-ordinating computational linguistics for Uralic languages, including forming best current practices and sharing information and resources in a centralised place.

References

Kornai, A. (2013). "Digital Language Death". *PLoS ONE* 8.10.

Tommi A. Pirinen
Hamburger Zentrum für Sprach Korpora, Universität Hamburg
tommi.antero.pirinen@uni-hamburg.de

Eszter Simon
Research Institute for Linguistics, Hungarian Academy of Sciences
simon.eszter@nytud.mta.hu

Francis M. Tyers
University of Tromsø
francis.tyers@uit.no

Veronika Vincze
MTA-SZTE Research Group for Articial Intelligence
vinczev@inf.u-szeged.hu

[3] http://gtweb.uit.no/sigur