Northern European Journal of Language Technology, 2016, Vol. 4, Article 1, pp 1–9
DOI 10.3384/nejlt.2000-1533.1641

Foreword to the Special Issue on Uralic Languages

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

Trond Trosterud
HSL-fakultehta
UiT Norgga árktalaš universitehta
trond.trosterud@uit.no

Francis M. Tyers
HSL-fakultehta
UiT Norgga árktalaš universitehta
francis.tyers@uit.no

Veronika Vincze
MTA-SZTE
Szegedi Tudomány Egyetem
vinczev@inf.u-szeged.hu

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

Jack Rueter
Helsingin yliopisto
Nykykielten laitos
jack.rueter@helsinki.fi

March 7, 2017

Abstract

In this introduction we have tried to present concisely the history of language technology for Uralic languages up until today, and a bit of a desiderata from the point of view of why we organised this special issue. It is of course not possible to cover everything that has happened in a short introduction like this. We have attempted to cover the beginnings of the (Uralic) language-technology scene in 1980's as far as it's relevant to much of the current work, including the ones presented in this issue. We also go through the Uralic area by the main languages to survey on existing resources, to also form a systematic overview of what is missing. Finally we talk about some possible future directions on the pan-Uralic level of language technology management.

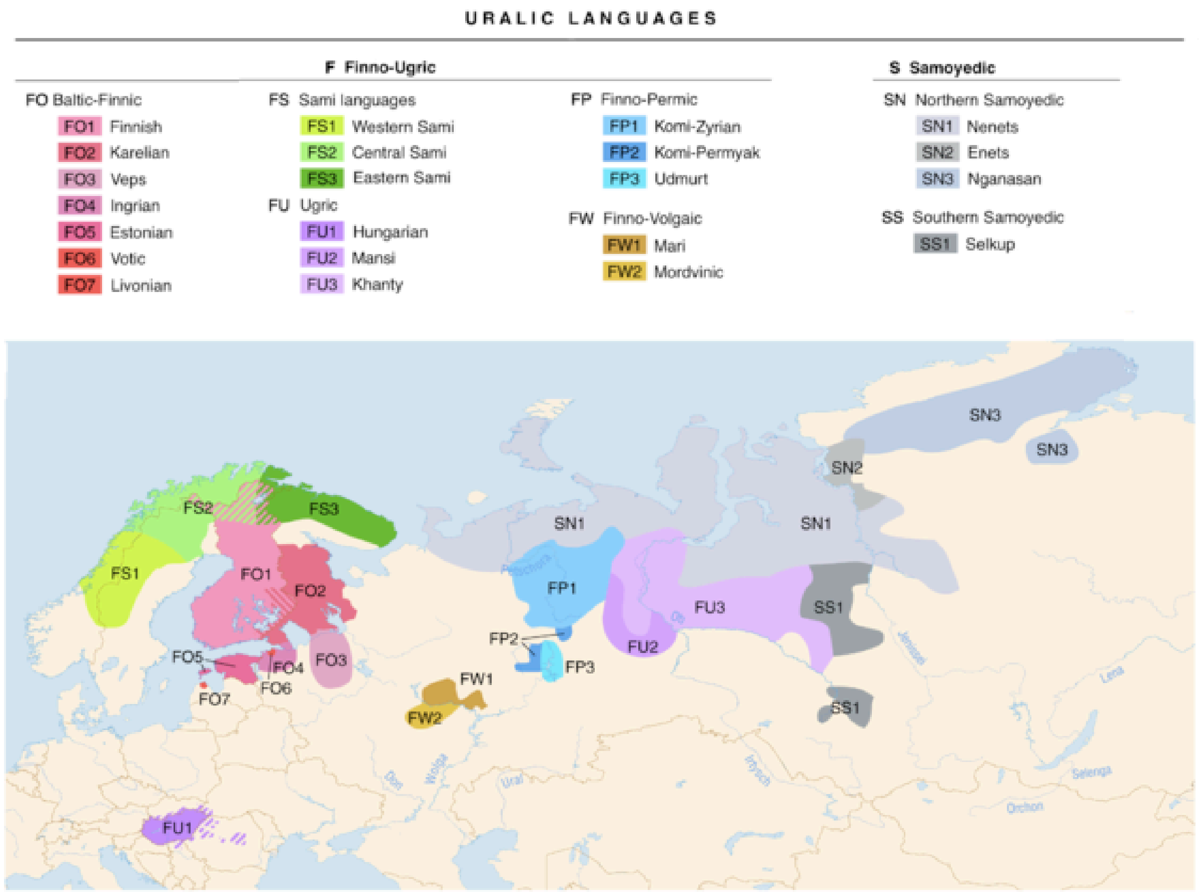

Figure 1: A map of the Uralic language area show approximate distribution of languages spoken by area. (Source: wikimedia commons)

1 Introduction

This special issue in Uralic language technology in NEJLT was conceived in the first international workshop on Uralic languages. At the workshop we recognised the need for journal publications tying up the progress of the field that has maybe not so much been dealt in such a centralised manner. We issued an open call for papers inviting publications of Uralic computational linguistics recent developments. We received three high quality submissions, which are centrally relevant to the field. In this introduction we describe the state-of-the art in Uralic language technology and how these three publications tie into the moving field of Uralic language technology.

2 The state of the art for Uralic language technology

Not all of the Uralic languages have the same level of support either within language technology nor in support of basic linguistic resources and research infrastructures in place. The national languages: Finnish, Estonian and Hungarian have the most resources and historical development. Besides that, also most Saami languages are strongly represented by the work done by Divvun and Giellatekno at UiT The Arctic University of Norway, and all in all more than half of the Uralic languages have at least some degree of resources and practical software available. This overview has been written based on input from the research communities in the relevant area and in particular the researchers mentioned in the author list of this introductory article.

Finnish language technology has its roots in the University of Helsinki early 1980's. Perhaps the most famous and cited work for the era is Kimmo Koskenniemi's doctoral dissertation [1] on two-level morphology, which remains to be the key technology behind most of the contemporary language technology in Uralistics. Another academic innovation in rule-based processing of natural languages that is central to the Uralic language technology came from professor Fred Karlsson in the 1990's: the Constraint Grammar [2] is likewise a core component of most of the uralic language technology today. Another innovation in Finnish academia that is widely used and cited comes from the statistical natural language processing part of the field that has grown and surpassed rule-based approach in many parts in the 2000's, at the Helsinki University of Technology Morfessor [3] was developed, a program that can recognise morphs or morph-like segments in text without human-written grammar, and is central to speech technology work for several Uralic languages as well as many machine translation approaches.

The first wave of Finnish morphologies was developed in the 1980's, including Koskenniemi's two-level implementation of Finnish. The so-called `twol-fi`, coupled with Karlsson's constraint grammar rules for Finnish was one of the most popular and widest in academic use. In the 1990's, a number of companies were founded around Finnish language technology resources, an extensive detail of these has been written by Arppe.[1] Other resources made at the time include commercial spelling and grammar checkers by Lingsoft and a dependency parser `fi-fdg`. A lot of these resources are lost or unusable for most users due to commercial licencing issues and for that reason the focus in the universities was from the early 2000's shifted towards building free and open-source software versions of the systems [4]. The free resources originally made at the University of Helsinki include a morphological parser OMorFi [5], the original constraint grammar of Finnish open-sourced by the author Fred Karlsson, free text-to-speech systems (*suopuhe* and *simple4all* for Finnish), the free statistical POS tagger `finnpos`, as well as lexical resources such as FinnTreeBank and FinnWordNet. At the University of Turku, the free systems and corpora include a dependency treebank, with statistical parser to complement the corpus, a Finnish parse bank and a Finnish propbank.[2] Outside academia there are free/open-source resources such as spelling and grammar checker tools for Finnish by *voikko*, and free machine translation systems using *Apertium*. For machine translation there is also a free controlled language translation system implemented in Haskell by the grammatical framework project. [6]

The current state of the Finnish language technology resources was captured by the European white paper [7], and the ongoing within pan-European CLARIN project does a good work in providing catalogues of Finnish resources, currently maintained in Kielipankki.[3]

[1] `https://kitwiki.csc.fi/twiki/bin/view/FiLT/ArppeEn`

[2] `http://bionlp.utu.fi`

[3] `http://kielipankki.csc.fi`

Hungarian is the most widely spoken non-Indo-European language in Europe. It is the official language of Hungary, where ca. 98% of the population of 10 million claims Hungarian as their native language, and it is one of the 24 official languages of the European Union. It is also spoken by Hungarian communities in the seven neighbouring countries, the largest one being an approximately 1.5 million community in Romania. Additionally, immigrant communities use it worldwide, primarily in the United States, Canada and Israel. With its 13 million speakers, Hungarian is 14th on the list of the most populous European languages [4]. Abroad, Hungarian is an official language in the Serbian region of Vojvodina, as well as in three municipalities in Slovenia. Hungarian is officially recognised as a minority or regional language in Austria, Croatia, Romania, Serbia, the Ukraine, and Slovakia.

Hungarian belongs to the Uralic language family, it is one of the Ugric languages of its Finno-Ugric branch. After the break-up of Ugric unity, which is believed to have lasted from 3000 to 2000 BCE [8], Hungarian drifted far from her Ugric sister languages (Khanty and Mansi) both geographically and linguistically.

Hungarian is the longest documented language of the Uralic language family, its recorded history spans over more than 800 years. The first three centuries of Old Hungarian are documented by fragments: nearly all documents were written in Latin with sporadic Hungarian words. The first surviving coherent Hungarian text is the Funeral Sermon and Prayer from around 1192-1195.

Modern Hungarian is written using an expanded Latin alphabet. In addition to the standard letters of the Latin alphabet, Hungarian uses several modified Latin characters to represent the additional vowel sounds of the language. These include letters with acute accents (á,é,í,ó,ú), diaereses (ö and ü) and their long counterparts (ő and ű).

Like all Uralic languages, Hungarian is a highly agglutinating language, which means that it expresses elements in a single word with affixes; mainly suffixes, but also some prefixes and one circumfix. In Hungarian, most suffixes harmonise with the stem they are attached to, which means that most suffixes exist in two or three alternative forms differing in the suffix vowel, and the selection of the suffix alternate is determined by the stem vowels. In Hungarian, a lot of the prepositional meanings found in English are expressed by cases. The number of cases is up to 18, depending on definition.

According to the META-NET White Paper Series [7], language technology support for the Hungarian language is fragmentary for speech processing, machine translation and text analysis, while it is rated as moderate for speech and text resources. The standard pre-processing steps (tokenisation, morphological analysis, shallow parsing, etc.) are completed for Hungarian, but treating the more difficult semantics requires further research.

Speech recognition and machine translation of Hungarian are studied at several universities and workplaces, but free tools and data are currently not available. It is a typical phenomenon at the Hungarian NLP market that the number of free databases and open-source programs is quite low. However, there exist some exceptions: the `hun*` preprocessing tools[5] are freely available, in addition to several corpora and databases that contain morphological, syntactic and/or semantic annotation[6], most notably, the Szeged

[4] `https://en.wikipedia.org/wiki/Languages\_of\_Europe`

[5] `http://mokk.bme.hu/en/eszkozok/`

[6] `http://rgai.inf.u-szeged.hu/index.php?lang=en\&page=resources`

Treebank [9], which is the largest manually annotated treebank for Hungarian.

Over the past decade, a number of important electronic language resources (dictionaries, corpora, lexical databases) as well as processing resources (spell checkers, morphological analysers, etc.) have been developed. Activities, however, have not been synchronized, and not uncommonly similar resources have been developed in parallel at different locations. As a consequence, there are two morphological analysers for Hungarian: `Hunmorph` [10] and Humor [11]. Different formalisms have been used in these, which are either incompatible or difficult to convert from; there is also a lack of documentation and in many cases copyright issues are unclear. Nevertheless, in recent years the international trends of standardisation and uniformisation of existing resources have reached Hungary as well. Several projects started off with the objective of integration and interoperability. For instance, the morphological and dependency annotation of the Szeged Treebank has just been converted to the Universal Dependencies format. Universal Dependencies is an international project aiming to create a standardised tagset for morphological and dependency annotation applicable to as many languages as possible [12], one of which is Hungarian. Besides, the morphological tagset developed for Old Hungarian is also being converted to the UD format.

There are several spell checkers for Hungarian, the most widely used one is the open source `Hunspell` tool[7], which has been integrated into several software packages such as OpenOffice, Mozilla Firefox, Mozilla Thunderbird and Google Chrome.

Due to the variable word order characteristic of Hungarian, we cannot rely on exploiting particular linear configurations alone when syntactic parsers are developed. On the other hand, morphological case markers and postpositions lend themselves to being used as cues for parsing. `HunTag`[8], a MaxEnt-based sequential tagger can be used as a shallow parser [13] or as a Named Entity recogniser [14] as well. As for deep syntactic analysis, both constituency and dependency parsers [15] have been recently developed, which make use of statistical methods and were trained on the manually annotated Szeged Treebank.

Recently, focus on development for NLP companies and research institutes lies on providing trend- and text-analysis tools which integrate natural-language processing tools to find the relevant information in unstructured text. For this purpose part-of-speech taggers, dependency parsers and named entity recognisers have been developed for Hungarian, which are mostly based on statistical learning algorithms. Another new direction in Hungarian NLP research is to process non-canonical texts, be it old fragments from earlier centuries or texts created by social media users, for which purpose the tools developed for standard texts so far have been adapted and modified accordingly.

The Estonian field of language technology is also active and well-developed, with resources such as `est-nltk` [16] as well as freely-available morphological analysers. There is a dependency treebank for Estonian, ArborEst which follows the VISLCG system for annotation, and also a version following the annotation guidelines of the Universal Dependencies project.

The languages with the most resources, after the major national ones, in the Uralic area, are the Saami languages. Several of the ten Saami languages have a literary history going back three centuries or more. The oldest of the six languages for which standardised orthographies are in use is from 1979 (the one for North Saami). Linguistically, the Saami

[7] `http://hunspell.sourceforge.net/`

[8] `https://github.com/recski/HunTag/`

languages have a repertoire of morphological categories resembling the Finnic languages, with less cases (6-11), with less or no noun phrase internal agreement. As for other Northern (Samoyed and Ugric) Uralic languages, the Saami languages show dual number for verbs.[17] Four of the six orthographies possess letters not found in Western European languages or (for Kildin Saami) Russian, this hampered the digitalisation of Saami literacy for at least a decade, and in some cases still does.

The available digital resources for Saami languages largely thanks to the language technologies managed by the Giellatekno and Divvun groups at UiT. Their repository includes free and open source tools and dictionaries. The resources include traditional finite-state morphological analysers, syntactic analysers written within the framework of constraint grammar [18], as well as a number of end-user facing tools derived from them, such as spell-checkers[19], e-learning programs[20], machine translation systems and electronic dictionaries[21], and written and annotated corpora[22], for North, South, Lule, Inari and partly also Skolt Saami.

Thanks to the open *Giella* infrastructure with Giellatekno research and Divvun tool development a number of minority Uralic languages as well as other circumpolar languages have developing projects. [23] In addition to the Saami languages, sizeable analyser development has been carried out for Kven, Livonian, Olonets-Karelian, Võro; Erzya, Moksha; Hill Mari, Meadow Mari; Komi-Zyrian, Udmurt; Nenets. Open-source bilingual dictionaries with Norwegian, Finnish, Hungarian and Russian as source or target language are also present for most of these languages. In addition to these languages, the Giella infrastructure has been developed for ten other Uralic languages as well, technological support for these language is at most fractional and scarce.

3 Future plans and desiderata

The future of Uralic language technology scene depends, we believe, in increased co-operation of the researchers and sites in topic of Uralic language technologies. Many existing text and lexicon resources are not digitally available for research and development, and the basic tools for grammatical analysis of most Uralic languages are far from being fully developed. For a majority of the Uralic languages the analysers are still good enough to be used for practical purposes, and the best path forwards seems to be a dialectic process of research and development, along the path already initiated. To further the cause, the authors of this special issue have organised an Association of Computational Linguistics (ACL) Special Interest Group (SIG) for Uralic Languages (ACL SIGUR). SIG is a loosely organised body of researchers interested in specific topic. Current endeavours of the SIG include yearly meetings in form of a work-shop, that may in future be extended to workshops or conferences. On the practical side of things, we are trying to co-ordinate a standard for analysis of Uralic Languages under the Universal Dependencies project. We also endorse researchers to work on treebanks and similar resources to increase visibility of the minority Uralic languages within the international research community. This ensures that Uralic languages will have a strong presence in the international annotation standard.

Since the resources in computational linguistics are not necessarily long-lived or stable, we have not included in this article a collection of links or pointers to them, instead we are organising these in an internet page [9] that is kept up-to-date with help of feedback from the community.

4 Papers in this issue

The accepted papers are: Antonsen et al. *A North Saami to South Saami Machine Translation Prototype*, Gerstenberger et al. *Instant Annotations: One Example of Utilizing Language Technology in the Documentation of Endangered Uralic Languages* and Grönroos et al. *Low-Resource Active Learning of Morphological Segmentation.*

Antonsen et al. may serve as an example of rule-based machine translation system for closely related languages. This work may serve as an excellent introduction for developers of other Uralic language technology resources on the potential of the language technologies for Uralic languages.

The paper by Gerstenberger et al. discusses work in progress on digitising and annotating spoken language data for Pite Saami, Kola Saami and Izhva Komi. They show how computational tools, such as finite-state transducers can complement traditional annotation tools for spoken-language documentation.

The statistical morphological systems of low-resource languages is very central development for all the Uralic languages, most of which still lack full-fledged morphologies and large corpora, and even for more developed language technologies to complement the systems in terms of coverage, the statistical systems are needed for additional hypotheses of loan words, neologisms and such common out of vocabulary items that traditional knowledge based systems lack.

References

[1] Kimmo Koskenniemi. *Two-level morphology.* PhD thesis, 1983.

[2] Fred Karlsson. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 168–173. Association for Computational Linguistics, 1990.

[3] Mathias Creutz and Krista Lagus. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0.* Helsinki University of Technology, 2005.

[4] Kimmo Koskenniemi. How to build an open source morphological parser now. *Resourceful Language Technology*, page 86, 2008.

[5] Tommi A Pirinen. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. 28:381—393, 2015.

[6] Aarne Ranta. Grammatical framework. *Journal of Functional Programming*, 14(02):145–189, 2004.

[9] http://acl-sigur.github.io/matrix.html

[7] Eszter Simon, Piroska Lendvai, Géza Németh, Gábor Olaszy, and Klára Vicsi. The hungarian language in the digital age – a magyar nyelv a digitális korban, 2012.

[8] Katalin É. Kiss. Introduction. In Katalin É. Kiss, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 1–8. Oxford University Press, 2014.

[9] Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. The szeged treebank. In Václav et al. Matoušek, editor, *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, pages 123–131. Springer, 2005.

[10] Viktor Trón, Gyögy Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: Open source word analysis. In *Proceedings of the ACL Workshop on Software*, pages 77–85. Association for Computational Linguistics, 2005.

[11] Attila Novák. Milyen a jó humor? [what is good humor like?]. In *Proceedings of the 1st Hungarian Computational Linguistics Conference, SZTE, Szeged*, page 138–144.

[12] Joakim Nivre. Towards a universal grammar for natural language processing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, page 3–16. Springer.

[13] Gábor Recski and Dániel Varga. A hungarian np chunker. In *The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics*, pages 87–93. ELTE School of English and American Studies, 2009.

[14] Eszter Simon. *Approaches to Hungarian Named Entity Recognition.* PhD thesis, BME, 2013.

[15] Richárd Farkas, Veronika Vincze, and Helmut Schmid. Dependency parsing of hungarian: Baseline results and challenges. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 55–65.

[16] Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. Estnltk - nlp toolkit for estonian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[17] Pekka Sammallahti. *The Saami Languages. An Introduction.* Davvi Girji, 1998.

[18] Lene Antonsen, Trond Trosterud, and Linda Wiechetek. Reusing Grammatical Resources for New Languages. In *Proceedings of LREC-2010*, Valetta, Malta, 2010. ELRA.

[19] Lene. 2013 Antonsen. Čállinmeattáhusaid guorran. *Sámi dieđalaš áigečála*, (2):7—32, 2013.

[20] Antonsen Lene, Saara Huhmarniemi, and Trond Trosterud. Constraint grammar in dialogue systems. In *NEALT Proceedings Series*, volume 8, pages 31–21, 2009.

[21] Ryan Johnson, Lene Antonsen, and Trond Trosterud. Using finite state transducers for making efficient reading comprehension dictionaries. In NEALT Proceedings Series, editor, *Proceedings of the 19th Nordic Conference of Computational Linguistics*, volume 16, pages 59—71, 2013.

[22] Saara Huhmarniemi, Sjur Moshagen, and Trond Trosterud. Usage of xsl stylesheets for the annotation of the sámi language corpora. In *Proceedings of the Linguistic Annotation Workshop*, pages 45—48, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[23] Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. Open-source infrastructures for collaborative work on under-resourced languages. In *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, LREC, pages 71–77, Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era.