Dutkansearvvi dieđalaš áigečála

vol 9 ◆ no 1 ◆ 2025

Tutkâmseervi tieđâlâš äigičaalâ

Tu′tǩǩeemsie′br tiõđlaž äi′ğğpââ′jjlõstt

Dutkansearvi

Dutkansearvvi dieđalaš áigečála
vol 9 ◆ no 1 ◆ 2025

Special issue
Indigenous knowledge and languages in interaction –
Amazonian and Arctic approaches

Guest editors
Gessiane Lobato Picanço
Justino Sarmento Rezende Tuyuka Dʉpó
Pirjo Kristiina Virtanen

Publisher
Sámi Language and Culture Research Association

ISSN 2489-7930

Language technology for the Uralic languages in Amazonian contexts

Jack Rueter
University of Helsinki

Niko Partanen
University of Helsinki

Abstract

This article explores the application of language technology developed for Uralic languages in the revitalization, invigoration and research of Indigenous Amazonian languages. It highlights the potential and challenges of digital tool development for linguistically and culturally rich languages with minimal resources. Uralic language technology derives from the facilitation of Saami languages and over 150 other languages world-wide, including Indigenous Amazonian languages, such as Apurinã and Sakurabiat, in the GiellaLT infrastructure. This language-independent framework has been established, emphasizing reusability of technology and code to address the virtual lack of digital resources for minority languages, which helps in the enhancement of language documentation, education, and revitalization efforts. The authors present collaborative work between European and Brazilian researchers, emphasizing the importance of co-design with native speaker communities to ensure tools, such as spellcheckers, morphological analyzers, and keyboards, reflect actual language use. It is noted that despite environmental and cultural differences between Northern Eurasia and the Amazon, both regions share common challenges, such as underused language archives and endangered language vitality. The research underscores the importance of cross-regional cooperation and open-source, modular infrastructures for advancing language technology in minority and endangered language contexts worldwide.

Keywords:
Language technology, Uralic languages, Amazonian languages, language revitalization, Universal Dependencies

Introduction

Language technology is what we think of when we discuss the creation and implementation of computer tools that facilitate the use of language in the ever-expanding digital dimensions of the modern world. The size and prominence of a language in society does not necessarily go hand in hand with its presence in digital space. Hence, even if a language is spoken by hundreds of thousands or even millions, it might be virtually absent on the Internet and might not have any publications available in databases. Even when a language is actively used and present, actual numbers apply, i.e., the smaller the number of language users, the fewer the number of people there are developing language-specific tools or contributing to media in and for that same language. What happens if there actually are people who want to publish online, keep a blog or journal, or they simply want to write in a non-majority language.

A non-majority-language journal or blog might be faced with additional challenges. Just to name three, the editors might have to manually proofread every piece, the majority-language society might require that all publications have a majority-language translation, and there might be interest among non-speakers to read the basic message of a media they would otherwise not comprehend. Hence, we provide the gist of a story related by the head of Giellatekno ‘language technology (in Northern Saami)’, Trond Trosterud, in one of his many workshops:

“Once there was a Northern Saami weekly where the editors spent over 50 percent of their time proofreading what they had written, because there was no automated way of doing it. The head of the Giellatekno said that if the newspaper could provide quantities of text in digital format, Giellatekno would provide the newspaper with a spell checker. The spell checker was made and has meant that, now with more time on their hands, the newspaper can be published at least twice a week.”

Giellatekno, which together with Divvun form today's GiellaLT, was originally established at what is now the Norwegian Arctic University in Tromsoe, Norway in 2001, specializing in Saami language technology and working with open source and rule-based solutions. Soon the Saami Parliament in Norway was interested in an organization specializing in the application of research outcomes to practical tool development -- entré Divvun, which in Northern Saami means 'correction'. Today, Giellatekno and Divvun together are known as the umbrella organization GiellaLT, where research and tool development continues, not only for Saami languages but for over 160 languages around the world.

The people at GiellaLT have taken on the challenge of addressing an outside interest in what is actually written in Saami-language publications. They have developed a rule-based translation system for online Saami-language newspapers that allows the interested but non-fluent foreigners a peek into the news media. This means that the writers can spend their time more on what they actually want to do and are good at. At the same time, it must be noted that GiellaLT whole-heartedly develops this translation tool on a small language to large language scenario, but NOT the other way around -- speakers of majority languages can cope with below optimal level texts, but introducing less than perfect machine-translated texts for a minority language would only pollute the environment. Since GiellaLT stresses the concept of reusable code, this system has been enabled in a way that can be readily applied to languages being described in the GiellaLT infrastructure. This means direct support for small languages with limited resources. Native speakers, of course, have a head start in this kind of development; they have a beneficial knowledge of their

own native or heritage language, which they can continue to enhance, but they can also develop additional skills and expertise, which they might use regardless of the language.

Development of language technology for a minority language is not only a way of establishing that language in the digital age. It also means the establishment of new domains for the language. By promoting language speakers as language professionals and facilitators, we are adding esteem to the cultural aspects of the language as well. Language facilitation can also be sped by sharing language-independent infrastructures, such that ready solutions for any number of descriptive, implementation or development issues can be followed, e.g., analyzers for linguists, spell checkers and grammar checkers for writers, enhancement for text-to-speech apps, and translation. Since no one is actually going to become a millionaire describing a small language, we need to provide open-source golden corpora for the chance contributors to the development and description of these languages. To this end, we suggest among others work in the Universal Dependencies (UD) project, where language resources intended for the development of language technology can also be displayed as a scientific citing venue for language research and archives. Universal Dependencies is a project that contains annotated text materials in many different languages, trying to use the same annotation scheme consistently. This makes the language materials comparable in a way that has not been possible before. UD might also serve as a platform for the curation of language understanding, and the annotated sentences can be used and displayed in different environments, i.e. in dictionaries or other applications.

These approaches are largely independent of individual languages. Thereby, although the relevance of language technology for the Uralic languages outside the field of Uralic studies may not be obvious at first glance, there is surprisingly much common ground. Uralic languages are endangered and they have a rich tradition of fieldwork-based data collection. They do have existing language resources, but they are not uniform in their written tradition or transcription conventions, and may need extensive work to be integrated into contemporary language technology platforms. Uralic languages are not alone in this situation, and from this point of view there are many similarities with languages all over the globe. We at the University of Helsinki have had systematic collaboration with our partners in Brazil, and this has given us space to exchange experiences and learn from each other. Our experience from this line of work has been very positive, and in this article, we discuss various points of view where we have found that Uralic studies and work on Amazonian languages can contribute to the bilateral enrichment of knowledge compiled by researchers and language communities on both sides. Although our approach highlights language technology, we are convinced that there are many other aspects where relations such as these can be very important if not vital to the individual language communities.

The environments of the Amazon and Northern Eurasia are quite diverse, and it may be a challenge to align the natural habitats of Amazonian rainforest and Northern Eurasian evergreen forest, steppe and tundra. Thus, there are fewer natural phenomena that can be identified among flora, fauna and fungi by the two sets of speech communities living in these areas, where even the weather is different – snow and ice versus rain every day or all day. The diversity in the environmental differences provides the foundation for diversity in traditional livelihoods, cultures and modes of subsistence. Both regions, however, have numerous endangered languages with a rich history of language documentation and the existence of large, usually underused, archive collections. Language shift is often seen as advancing rapidly, and the loss of

language weakens the cultural relevance of traditional societies. Losing one's linguistic homeland is associated with extensive and simultaneous changes in one's society and integration with surrounding societies. By bringing together researchers who have focused on different environments in their careers, we can accumulate their understanding of nuances in all of our work which is not obvious in a scenario where we confine ourselves to what is familiar to us.

Our work has primarily focused on language technology and applying it to the Indigenous and endangered languages. In the modern world, language technology has multiple applications, some which are very visible in daily life. Keyboards, both on computers and mobile devices, are a good example. This also illustrates how the tools of language technology need to be developed in collaboration with their users: the keyboard has to meet the needs of the language community, and their guidance and collaboration must be included in the project from the beginning. Whether something is needed and wanted should always be the first questions when initiating new work. Working with language means working with the cultural heritage of Indigenous peoples, which directly references responsibilities and implications (see Development of language technology, paragraph 5, above).

The collaboration described in this article was initiated within a project coordinated by professors Pirjo Kristiina Virtanen, Sidney Facundes and Thiago Cardoso Mota. This has included contact and collaboration between researchers at the University of Helsinki, Finland, Federal University of Pará (UFPA), Belém, and Federal University of Amazonas, Manaus, Brazil. Both introduce knowledge exchange, regular travel and stays of the participants in these universities, but also contact maintained in the meanwhile, has been very important in developing various collaborations within the framework described here.

Among the current collaborators, Jack Rueter, has worked with the description of indigenous languages of the Americas since 2014 (Rueter et al. 2021, 2023). While the first descriptions are limited to basic phenomena of Salishan, Sahaptin languages of the Pacific Northwest, his greatest progress has been in collaboration with Brazilian researchers of Arawakan Apurinã and Tupian Sakurabiat/Mekéns. It is in work on Apurinã that collaboration is establishing a workflow where researchers and native speakers together document illustrative use of the language and share this information for finite-state description by Rueter. The collaboration produces many outcomes that enhance the studies of the target languages.

The finite-state description indicated above is used in analysers of Apurinã for the researchers and the possible introduction of a spell-checking instrument dependent upon an Apurinã authoritative organ and an extensive set of open-source keyboards for the Apurinã language. The analyser for researchers is also used in the annotation of texts published in the Universal Dependencies project, but it can also be flipped for use as a generator. The generator can produce word forms for more extensive collaboration between fieldworkers and native speakers interested in documentation of the limitations of regular morphology, i.e., the generator only produces what it is told to produce. If it produces "regular forms" that are not acceptable, the native speaker researcher is able to identify limitations to necessary/possible generation. This, of course, also requires multiple voices in evaluation. Evaluation will need plenty of additional work.

Uralic language technology as an extension of Saami language technology

Language technology, when understood from a broad perspective, encompasses a

wide array of tools and technologies that can be used to process linguistic materials. This implies the need for an infrastructure that might be made available for the study and documentation of several languages at once. Despite the fact that some of the technologies, such as those involved in the construction of rule-based morphological analysers, require relatively extensive language specific development in order to be applied to a new language, there are many aspects of an ideal infrastructure that can be reused. This, in itself, introduces scientific research questions beyond the descriptive and typological ones inherent in linguistic fieldwork – how to build a language-independent infrastructure suitable for language research, revitalization and maintenance.

The construction of an infrastructure for language research makes up a notable portion of any language-research project. In order for the infrastructure to be shared in the study and documentation of several languages, it must not be limited to the structures of an individual language, instead, it should be designed as a language-independent infrastructure with extensive modularity and more than one language-research team to drive it. There must also be certain principles agreed upon by the teams that include adherence to reusability, language independence as well as collaboration with language curating institutions and long-term maintenance and archiving. With a language-independent infrastructure of this nature in place, new language-research teams can follow the lead of teams already working, contribute to diversity and concentrate on their own research. Teams working with Uralic languages in many countries are well aware of the open-source, Saami language-technology infrastructure «GiellaLT» based at the Norwegian Arctic University in Tromsø, Norway. One way to access these tools is a Python package developed by our collaborator (Hämäläinen 2019). At the same time, distinct infrastructures have also been developed to maintain lexical information (Alnajjar et al. 2020). The goal within Uralic language studies has been described as digital documentation of Uralic Languages with open-source tools and modern NLP methods (Hämäläinen et al. 2023). Language documentation has been taken into consideration from early on when developing these methodologies, with various ways to integrate these tools into ELAN files and other tools commonly used in the field (Gerstenberger et al. 2017; Jouste et al. 2022). Collaboration with our Amazonian colleagues has also drawn our attention to the need to address SIL Fieldworks based workflows as well. At least integrating and using data stored in this format in NLP tools would be an important step forward. For many languages the largest collected lexicons are stored in this software. The situation is somewhat different with the Uralic languages, where historical lexical collections from the early 20th century are often the largest and most extensive type of resource, and these have often already been published as dictionaries.

One tool developed for «GiellaLT», which is so common in the documentation of languages and whose importance is seldom considered, is the keyboard. The idea of this tool is that a keyboard be set up for each individual language with one file that describes all layouts for that language, i.e., the layout file should describe the requirements for Android, Windows, MacOS, iPhone, iPad, Chrome and other instances. The motivation for one keyboard for each language lies in the fact that even now, in the Windows operating system, the keyboard tells the computer what the input language is. So, by making an Apurinã keyboard, we are automatically enabling use of an Apurinã spell checker. The challenges of such an undertaking are numerous. First, the standard layout for the majority language of a given country (Brazil) should be noted, i.e., both linguists and language users will find it easier to

begin using the keyboard. Second, the keyboard should provide for all characters used in the modern language and historical documentation, and the strokes required for producing characters and punctuation should be mnemonic from the language users' perspective. Finally, in order to produce a successful keyboard a professional user should participate in the development. In work with the Apurinã keyboard, for example, there were six people involved – two working directly in the infrastructure and four providing extensive feedback regarding key positions in various layouts and the smoothness of download and updating,

Introspection of Amazonian and Uralic language research traditions, archives and actors is a way of providing further impetus to the extension of a shared research infrastructure. By introducing new players with different approaches to similar data sets, research and language revitalization histories, «Language technology in the Amazonian/Uralic context» workshop, conducted in September and October 2023, played an important role in introducing points of mutual benefits for diverse research teams. Awareness of openly available language tools and methodologies will hopefully provide an understanding of where development is needed and can be continued.

Understanding the importance and challenge of larger sets of tools, methods and the modularity of a shared research infrastructure brings the Amazonian/Uralic teams back to the needs and practicalities of individual language research. Whereas morphological analysers are initially constructed for linguists, these same analysers with normative adaptation can be used as components in spell checkers and computer-assisted language learning tools. The analyser, for example, inherently contains the lexicon of the language, which, in principle, can be directly linked to corpus analysis and dictionary creation. The unrecognized forms have to be explicitly explained in the construction of the analyser, and if something cannot be described, this would suggest that something is not totally understood yet.

At times, work with phenomena not completely understood can be enhanced through collaborative or parallel work. Collaboration provides unexpected insight from other research traditions. It shows us the overlap between lexicography, phonetics, morphology, syntax, language learning, even etymology, translation, etc. In a word, collaboration instills an overview of what information is necessary or auxiliary for different aspects of language documentation. It also introduces new open-source venues where many of these overlapping segments might be joined for a better comprehension of the world's languages. One such venue is the Universal Dependencies project inaugurated January 15, 2015.

The Universal Dependencies project produces biannual releases of curated, annotated corpora intended for improved language-technological training and testing grounds. This same project has since become recognized for its potential in the study of language typology. During the past decade, many Uralic languages have been included in the Universal Dependencies project. Lately, the same advancement has been observed in languages of Amazonia and South America more widely. The goal of the Universal Dependencies project is to provide similarly annotated corpora for a large range of languages, with the same annotation scheme and underlying data structure. This makes the Universal Dependencies treebanks a great source for comparative research, and these treebanks have seen increasing use within recent years. Besides this they also form comparable and uniform structures for the development of language technology and tools of natural language processing.

One practical result of this collaboration has been the release of the first Apurinã

Universal Dependencies treebank. The treebank contains fully annotated sentences from the Apurinã language. When they were being created, numerous linguistic questions also had to be addressed. Apurinã was the first Arawakan language in the project, and this alone necessitated that some of the features in this language family be adequately thought over and addressed. It is very significant for the Universal Dependencies project that the languages represent as many language families as possible. And for each language, different genres and styles should ideally also be present.

In 2022 and 2023, language technology courses were taught in direct collaboration between the Universities of Helsinki and UFPA, in Belém. The content of the courses included familiarization with different tools of Natural Language Processing, starting with the rule-based analysers the research group in Helsinki has been working with for a long time in the context of Uralic languages. The lectures also introduced methodologies that can be used to process image-, audio- and video-based materials. This included especially scanned transcribed documents with audio- and video-aligned transcriptions. First, the course was conducted online, and later in Belém. Experiences were very good and we plan to continue this initiative. Different presentations and seminars, both in Helsinki and Belém, complement these longer teaching initiatives in an excellent manner.

Amazonian languages in their own contexts

There are some key differences between the languages spoken in Amazonia and the Uralic languages. First of all, the Uralic languages all belong to one single, well-established language family, whereas in Amazonia there is a vast number of language families and language isolates. This means that the differences between languages are inevitably larger, and the possibilities of reusing materials from one language to the next are less obvious. Learning numerous Amazonian languages takes necessarily more effort than learning various Uralic languages, as it is not possible to build upon similar grammatical structures that are found throughout the Uralic languages and easily reused in a fairly comparable manner in these languages.

The Indigenous languages of Amazonia that we have been working with through this collaboration are, in general, smaller than many Uralic languages, and the areas where they are spoken are smaller as well. The language families that we have been introduced to are Arawakan, Tupian and Jê. The language diversity in Amazonia is all in all much greater than it is in Northern Eurasia. Northern Eurasia attests to at least eight language families. Amazonia, in contrast, is the home of at least ten language families and three language isolates. Usually, Amazonia is reported as having 15 to 20 language families, while in general the typological similarities in the language families of the Northern Eurasia reduce the diversity in this region even further.

Some aspects of dissimilarity between Uralic languages and those of the Amazon actually lie in traditions of their documentation. While both research traditions might well recognize the same phenomena, they might not apply the same terminology. The concept of consonant gradation is familiar virtually to anyone studying Balto-Finnic or Saamic languages, so the presence of consonant variation at word boundaries might immediately be seen as evidence of consonant gradation, whereas the phenomena might actually be a matter of allophonic variation. The terminology used for describing object and subject marking on verbs in Uralic studies is by tradition “objective” and “subjective” conjugation, while description of conjugation marking in languages of the Amazon might refer to “subject-object”

conjugation. Distinctions between languages might even be observed in the virtual absence of counting systems. Linguistic diversity can best be studied through collaboration between specialists working with diverse languages. One example of this can be seen in the open-source Universal Dependencies project where both languages of Amazonia and Uralic languages are receiving more and more attention.

The Universal Dependencies (UD) project is making a concerted effort to find relations for describing all languages of the world. At the same time the UD project provides for the presentation of a new type of open-source text corpora. This is an important aspect in minority language studies whose corpora are not limited to one type (see Rueter & Partanen, 2019). In UD, annotators learn a new awareness, and yet they might become confused when dealing with languages from vastly different environments. Whereas the Uralic languages often have negative auxiliaries, which conjugate for person or indicate tense, mood or aspect, there is often a temptation to call the Apurinã word of negation, which does not conjugate or show tense, aspect or mood an auxiliary when, in fact, the word of negation might better be described as a particle of negation. Collaboration in this kind of project, although initially complicated, can prove to help co-researchers find new solutions for describing the phenomena of each others' language of study.

Another similarity is that while within the majority of the Uralic language speech area the main contact language is Russian, in the majority of the Amazonian context it is Portuguese. The similarity is that a larger, Indo-European language plays the role of majority language. From a wider perspective, of course, this is a simplification, but, in principle, it has many implications. For example, most of the bilingual dictionaries for these languages use the same language as the target language. Similarly a large part of the existing grammatical resources are in that individual target language.

Shared similarities in linguistic data

One central similarity in both the Amazonian and Uralic regions is that in the end the linguistic data types are fairly comparable. Text collections and dictionaries are similar sources, and contain comparable elements, even when they are dealt with in different research traditions. Local research traditions may be different, for example, in the question of what kind of annotations have been preferred, or what the role of historical linguistics has played for the wider orientation of the field. For Uralic studies, most of the traditional research has indeed focused on the relations between various Uralic languages, and etymological research needs that are very closely tied to the meticulously thorough collection of lexicon. At the same time, questions such as language contact have started to be asked just in the last decades. There may be marked differences in how the research tradition of the field has been set in the very multilingual context of Amazonia and within Uralic linguistics. At the same time, when the data itself has been collected, there are certain aspects that are very universal, or at least appear shared here.

There is a great similarity in how much of the existing collected material remains unpublished, and the organization and curation of the data collected during the last few decades continuously demands very extensive resources from the researchers and students working with them. The best practices in the field of language documentation are well understood and there have been decades of discussion addressing desirable workflows, but there is still a very acute need to automate these processes and establish conventions for handling this sort of data as part of the daily

data collection and research workflows. This challenge is no different, regardless of whether the linguistic fieldworker is pondering these questions on the bank of the Rivers Volga or Purus.

The need to represent the same texts in different transcription systems and orthographies, depending on audience and intended use, is also relatively similar. Individual researchers have used diverse transcription systems with different ideas about phonology, and also the orthographies may address different issues at varying levels of accuracy. Linguists may need some information in the transcriptions, and the language users may need something else. The way we would like to frame this problem is that we do not need to choose one system, but ideally transcription and orthography, or some different transcription levels, can be automatically derived from one another. An additional useful part here is that this demands very thorough analyses of the strengths of different solutions, which may further provide orientation in their use.

Some of the technologies investigated in this collaboration are still being adopted both in Uralic and Amazonian studies. For example, we are still waiting for consistent and high-quality results in speech recognition for endangered Uralic and Amazonian languages. There have been individual positive reports (Partanen et al. 2020), but applying these tools in practice has not yet been done. In an ideal scenario, we would be able to use existing transcriptions and their audio to fine tune a speech recognition model for a specific endangered language and a given corpus of recordings. We are nearly at the point where we can do speech recognition for the segments where a local majority language is spoken. Nonetheless, to advance this from individual experiments and tests, it would be necessary to systematically evaluate how well this functions currently and how much work is involved in the correction of the output.

At the moment the text recognition of rare and complex scripts has developed relatively far, such that both printed and handwritten texts can be extracted fairly easily. This easily leads to more effective reuse and publishing of archival materials (Partanen et al. 2022).

Conclusion

Most importantly, Uralic-Amazonian linguistic collaboration described in this study has potential to benefit the speakers of endangered languages by advancing the level of language technology support for these languages. The use of languages in different domains of society is crucial, and the digital environments are becoming commonplace everywhere. The development of language technology allows the use of indigenous-language keyboards on various devices, among these computers and mobile phones. When these devices are used, spell checking and dictionaries are among the tools that are crucial in ensuring that digital communication is effortless and well-functioning. Naturally, there is the question whether the community sees this development as desirable or necessary, but it's important to provide the possibility.

Linguistic work over decades, if not centuries, has resulted in large amounts of materials in indigenous languages that are not currently available to the communities. The methods described here allow digitizing and processing more effectively with many different types of data, which may be of crucial importance when scarce resources in endangered languages can be made better available to the communities from which they originated. Collaboration between the language community members, linguists and natural language processing researchers is continuously of utmost importance, and the language community would ideally guide the direction and priorities of the development.

There are also continuous new developments that have to be taken into account. We've recently demonstrated that even some smaller Uralic languages can be very efficiently processed with Large Language Models (Partanen 2024). Hämäläinen et al. (2024) have also pointed out the need to take these technologies into account, even in the context of endangered languages. At the same time, the questions of ethics and responsibility, as well as data ownership, become all the more important and must be considered carefully and from different perspectives. Recent studies have also shown very promising results in using LLMs in glossing endangered languages (Ginn et al. 2024), which would be very useful in the context of language documentation both in Northern Eurasia and Amazonia. The future remains promising but needs extensive collaboration and understanding of our shared issues and questions. ♦

References

Alnajjar, Khalid, Mika Hämäläinen, Jack Rueter, & Niko Partanen. 2020. "Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement." In Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations, eds. Michal Ptaszynski, Bartosz Ziolko, 1-6. Barcelona: International Committee on Computational Linguistics (ICCL).

Gerstenberger, Ciprian-Virgil, Niko Partanen & Michael Rießler. 2017. "Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region." In Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, eds. Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz, Association for Computational Linguistics, 57-66. Honolulu: Association for Computational Linguistics.

Hämäläinen, Mika. 2019. UralicNLP: An NLP library for Uralic languages. Journal of open source software 4(37): 1345.

Hämäläinen, Mika, Jack Rueter, Khalid Alnajjar & Niko Partanen. 2023. "Working Towards Digital Documentation of Uralic Languages with Open-Source Tools and Modern NLP Methods." In Proceedings of the Big Picture Workshop, eds. Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, Noah A. Smith, 18-27. Singapore: Association for Computational Linguistics.

Hämäläinen, Mika, Emily Öhman, So Miyagawa, Khalid Alnajjar, Yuri Bizzoni, Jack Rueter & Niko Partanen. 2024. "The Growing Importance of Humanities for NLP in the Era of LLMs." In Lightning Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, eds. Mika Hämäläinen, Emily Öhman, Khalid Alnajjar, 2-6. Helsinki: Association for Computational Linguistics.

Jouste, Marko, Jukka Mettovaara, Petter Morottaja & Niko Partanen. 2022. "Archive infrastructure and spoken language corpora for Saami languages in Finland." In Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-1, 2022, eds. Karl Berglund, Matti La Mela & Inge Zwart. CEUR-WS, Vol. 3232: 269-278. Aachen: RWTH Aachen University.

Partanen, Niko 2024. "Using Large Language Models to Transliterate Endangered Uralic Languages." In Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, eds. Mika Hämäläinen, Flammie Pirinen, Melany Macias, Mario Crespo Avila, 81-88. Helsinki: Association for Computational Linguistics.

Partanen, Niko, Rogier Blokland, Michael Rießler & Jack Rueter. 2022. "Transforming Archived Resources with Language Technology: From Manuscripts to Language Documentation." In Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference (DHNB 2022), Uppsala, Sweden, March 15-1, 2022, eds. Karl Berglund, Matti La

Mela & Inge Zwart: CEUR-WS, Vol. 3232: 370-380. Aachen: RWTH Aachen University.

Partanen, Niko, Mika Hämäläinen & Tiina Klooster. 2020. “Speech recognition for endangered and extinct Samoyedic languages.” In Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, eds. Minh Le Nguyen, Mai Chi Luong, Sanghoun Song, 523-533. Hanoi: Association for Computational Linguistics.

Ginn, Michael, Mans Hulden & Alexis Palmer. 2024. “Can we teach language models to gloss endangered languages?” In Findings of the Association for Computational Linguistics: EMNLP 2024, eds. Yaser Al-Onaizan, Mohit Bansal, Yun-Nung Chen, 5861–5876. Miami: Association for Computational Linguistics.

Rueter, Jack, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen & Niko Partanen. 2021. “Apurinã Universal Dependencies Treebank.” In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the America, eds. Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, Katharina Kann: 28-33. Association for Computational Linguistics.

Rueter, Jack, Mika Hämäläinen & Khalid Alnajjar 2023. “Modelling the Reduplicating Lushootseed Morphology with an FST and LSTM.” In Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), eds. Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer, Katharina Kann, 40-46. Toronto: Association for Computational Linguistics.

Rueter, Jack & Niko Partanen. 2019. “On new text corpora for minority languages on the Helsinki korp.csc.fi server.” In Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы, eds. Z. A. Sirazitdinov, Buskunbaeva, L. A., Išmuhametova, A. Š., Šamsutdinova, G., G., 32-36. Ufa: Baškirskaâ enciklopediâ.

Homepage for the journal and the association
www.dutkansearvi.fi

Contact
Dutkansearvi c/o
Alkuperäiskansatutkimus PL 24
(Unioninkatu 24)
00014 Helsingin yliopisto, Suomi/Finland

Association's membership fee
20 euros per year, students and pensioners 10 euros.
IBAN FI98 5723 0220 3848 66, BIC OKOYFIHH

Dutkansearvi
Saamen kielen ja kulttuurin tutkimusseura