

# Developing technologies for the documentation and description of the low-resource Uralic languages Zyrian Komi and North Saami

Niko Partanen<sup>1</sup>, Thierry Poibeau<sup>2</sup>, Michael Rießler<sup>3</sup>

<sup>1</sup>University of Helsinki – niko.partanen@helsinki.fi

<sup>2</sup>CNRS & ENS / PSL & Université Sorbonne nouvelle, Paris – thierry.poibeau@ens.fr

<sup>3</sup>University of Eastern Finland, Joensuu – michael.riessler@uef.fi

## Abstract

The Uralic languages are spoken in northern Eurasia, and most of them (with the exception of Finnish, Hungarian and Estonian) are non-national endangered languages with varying official support and resources. Language technology can play a major role in better documenting and describing endangered languages and in making the related workflows faster and more efficient. However, applying contemporary methods effectively in this context remains a challenge. In our own projects, we have developed language technologies focusing on low-resource scenarios, specifically for the two Uralic languages Zyrian Komi and North Saami. In addition to providing an overview of this work, we detail what we see as the remaining obstacles and main challenges for our work. Although we focus on individual languages, our experiments translate to the wider situation of endangered languages in Northern Eurasia and beyond.

**Keywords:** Zyrian Komi (kpv), North Saami (sme), documentary linguistics, language technology, dependency parsing, OCR

## Дженьдӧдӧм текст

Урал кывъяс паськалӧмаӧсь Войвыв Евразияын. На пысьс унджыкыс (финн, эст да венгр кындзи) ӧнія кадӧ вошан выйынӧсь. Кыв технологияяс вермасны документируйтны кывъяссӧ, а сідзкӧ отсаласны видзны найӧс дзикӧдз вошӧмысь. Дерт, кокнимӧдасны и такӧд йитӧдын вӧчан мукӧд уджсӧ. Но колӧ пасйыны, мый ӧнія кадся методъясӧн вӧдитчӧны сьӧкыда на. Асланым проектын ми лӧсьӧдӧм коми-зыран да войвыв саами урал кывъяслы технологияяс. Ӧтувья серпас петкӧдлӧмысь кындзи танӧ ми видлалам и сійӧ мытшӧдъяссӧ, кодъяскӧд паныдасим удж нуӧдӧгӧн. Кӧть ми сёрнитам торъя кывъяс йпыльс, миян опыт петкӧдлӧ Войвыв Евразияса уна кывлысь серпассӧ.

## 1. Introduction

The Uralic languages form a family of languages spoken by approximately 25 million people, predominantly in north-eastern Europe and western Siberia. With the exception of Finnish, Hungarian and Estonian, all Uralic languages are endangered. In this context, language technologies can play a major role in better documenting and describing these languages. Developing language resources and technologies ensures that knowledge about their specificities will be retained and thereby enables us to help in preserving and teaching them, with information technology bringing major benefits to this end.

Originally, language technologies for Uralic languages have been developed only for written language variants (especially by the research groups Giellatekno and Divvun at the University of Tromsø). Our approach – informed by both computational linguistics and (fieldwork-based) documentary linguistics – also includes spoken language data. We have been developing technologies for Komi and Saami based on the most recent advances in natural language processing, and have applied them to a context where annotated corpus data and other resources are lacking.

In this paper, we give an overview of our most recent research (see more specifically Lim et al. (2018b); Partanen et al. (2018b) and Partanen et al. (2018a)) and we detail what are, in our view, the main challenges for language technologies in low-resource language documentation contexts in general. Our publications are focused on two specific languages but we think that a large variety of small languages can be approached in a similar way, especially when

raw digitized texts are already available but other kinds of resources, specifically annotated data, are lacking. In fact, this is true for most Uralic languages as well as for several Northern Eurasian languages from other families.

As the lack of annotated data is often emphasized as an obstacle, specific attention must be paid in this context to solutions that enable the rapid increase of annotated data. For most Uralic languages, the problem is not the lack of data, as material has been collected and published in most of these languages since at least the beginning of the 20th century (some of the textual materials are even old enough to be in the Public Domain, making it legally possible to create entirely open and easy reusable datasets). The problem is the lack of annotation for this data.

It must be noted that Zyrian Komi and North Saami are relatively well described and linguistically understood languages with speaker numbers ranging from about 30,000 (North Saami) to 160,000 (Zyrian Komi). Both languages also receive official support, have well-established written norms and are regularly used in media and education, even in vocational and higher academic education. Retrieving textual data for corpus building and (written) language technology is also comparably easy. In fact, extensive written corpora have been created for both languages already: the SIKOR North Saami free corpus with over 30M tokens by Divvun/Giellatekno and the Корпус Коми языка (“Corpus of the Komi language”) with over 50M tokens by the Centre for Innovative Language Technology in Syktyvkar, Komi Republic. Whereas these two corpora are tagged using rule-based NLP, various new approaches in language technology

are also evolving. A recent evaluation of currently available language technology for Finnish (Pirinen, 2019), for instance, showed that the Turku Neural Parser Pipeline clearly outperforms the older rule-based systems for Finnish, and when enough annotated data becomes available, the same can also be expected for other Uralic languages.

## 2. Summary of the research done so far

We have mainly developed our research along three different directions. One is *the integration of various language technologies in order to get more efficient NLP workflows for fieldwork-based language documentation* (Gerstenberger et al., 2017a; Gerstenberger et al., 2017b). The achievements in corpus building sketched in the section above concern only written language; fieldwork data representing spoken language has not been included in relevant projects earlier. A central approach for us has been to find ways to use language technology so that language documenters – collecting new data in the field or working with legacy data in archives – can work faster and more reliably with their language data. Ideally this should be done in connection with the language documentation activities that would normally take place anyway, and using the tools the researchers are already familiar with. This is also related to the ability to preserve and reuse the same material later on, although there are needs for improvement in all levels, from data management and archiving to final publication and re-use in research.

Compared to traditional NLP workflows for written texts, ours must integrate speech technologies, for speech transcription or signal analysis. As for written texts, processing workflows may go beyond pure NLP, so as to integrate document analysis and OCR when it comes to corpus integration of earlier documents (Blokland et al., 2019).

We have also conducted several case studies about *dependency parsing in these low-resource scenarios* (Lim et al., 2018b; Partanen et al., 2018b). Dependency parsing is now a relatively mature technology, mainly based on advanced machine learning techniques that require large amounts of annotated data to get accurate results. This is a major issue for low resource scenarios, but recent techniques based on multilingual models and language transfer have made it possible to get working results even in extremely low resource scenarios. With Uralic languages the most obvious approaches for multilingual systems would consider closely related languages and contemporary contact languages, and our experiments have covered both. The results are of course far from perfect but our aim in the long run is of course to use these methods for language documentation. Automatic annotations need to be revised and corrected, but they are useful to kick-start the annotation process and they also make it possible to considerably increase the size of the data produced (which, in turn, makes it possible to train better parsers that will require less manual correction).

The third portion of our work has focused more into *concrete resource creation, which is illustrated by two Zyrian Komi treebanks* (Partanen et al., 2018b) and a large spoken language corpus (Blokland et al., 2020). This shows that technical advances work hand in hand with the production of resources and help maintain and document en-

dangered languages. Our work aligns closely with observations others have made in relation to this field, namely that even a small amount of annotated data still brings at the moment clear improvements into any multilingual scenarios (Meechan-Maddon and Nivre, 2019). As our own datasets have grown, we are replicating and extending our earlier experiments, with the goal of reaching a workable solution for our continuous language documentation work. Although we have been successful in integrating language technology into language documentation workflows (Gerstenberger et al., 2017a), there are still numerous open questions about how the whole infrastructure should be set up so that resources and applications would be most beneficial for both field linguists and computational linguists. Some of these open questions are discussed next.

### 2.1. Persistent archiving of language documentation corpora

In the last 20 years a large number of language documentation projects have been conducted all over the world and provided vast digital resources on endangered and previously undocumented languages. There are, however, numerous problems in the actual use of these materials, especially in more computerized workflows. Language documentation projects produce complex multimedia collections and associated metadata. Lots of attention has been paid to open and shared formats in language documentation (Seyfeddinipur et al., 2019). Still, it is often a major challenge to maintain long-term consistency in such collections. As an outcome, language documentation corpora may be unsystematic in ways that make reusing them difficult. This relates closely to the fact that work on individual languages often continues for years, even decades: for this reason, work practices within a language documentation project have to be thoroughly documented, so that even changing and entirely new teams can connect and continue previous work.

A solution in our own projects is to define the intended data structures in a machine readable format, and to build a set of tests that continuously validate that both structure and content are within expected definitions (Partanen, 2019a).

### 2.2. Text recognition

A large number of linguistic resources for endangered languages, representing transcribed spoken language, have been published in books or are stored as manuscripts. For many publications audio recordings underlying the transcripts are stored in various private or public archives. The usefulness of these data for future work with endangered languages is without question (Blokland et al., 2019). However, merging this analogue data into digitally-born corpora can be challenging. Finding the relevant audio files and digitizing them as well as converting rare and non-standard writing systems designed (e.g. variants of various phonetic alphabets typically used in the printed texts) into contemporary standard orthography can be such challenges. Much more problematic, however, is often the exact matching between the transcript and the original audio. A reason for this seems to be that the recorded speech may be fuzzy and subject to interpretation and the later published version has

gone through orthographic and stylistic editing without taking the original recording into account.

So far our own work with such legacy resources has focused on digitizing the relevant texts, building OCR models for rendering the different original scripts and integrating the resulting data into our corpus infrastructure. The alignment of the processed texts with the original audio, if available, is an upcoming task.

The tools for performing text recognition in itself have improved considerably during the last years. Only a few years ago alternatives to ABBYY FineReader were relatively few, even though the problems present with this commercial software are numerous (Partanen, 2017). Recently it has been possible to train very well performing OCR models with open source software, such as Tesseract, Ocropy and Calamari (Partanen and Rießler, 2019). In connection to this kind of work we have published Ground Truth datasets and OCR models (Partanen and Rießler, 2019; Partanen, 2019b). The best practices in sharing Ground Truth data need to be taken into account, for example, by following the conventions used by the National Library of Finland (Kettunen et al., 2018). However, after the texts have been retrieved from the documents, several issues remain in their successful transliteration and normalization. There has been very promising work on normalizing Finnish dialect texts as a character level machine translation task, with achieved word error rate in around 5% (Partanen et al., 2019). More work is acutely needed in whether such approaches are viable also with endangered languages and when less data is available. Same normalization problem, however, also in language documentation context.

### 2.3. Dependency parsing

Since 2017 a number of experiments have been carried out by our team with dependency parsing of low-resource languages. The system used in these experiments is the Multilingual BIST parser (Lim and Poibeau, 2017). The experiments were done with cross-lingual scenarios where data from related languages and contact languages were used alongside the minimal training data in the target language (Lim et al., 2018b; Lim et al., 2018a). The experiments were promising, and the LAS score on Northern Saami was 51.54 and for Zyrian Komi 56.66. This improved from the parsing results demonstrated for Northern Saami in CoNLL 2017 Shared Task (Lim et al., 2018b, p. 2233). Despite improvements the achieved performance was generally not high enough for practical applications. Additional experiments were done with code-switching data, in order to understand how well this kind of a multilingual system is able to parse data that contains both of the languages it was trained on (Partanen et al., 2018b). It has to be noted that in spoken data, such as typically included in spoken corpora of endangered languages, code-switching is the rule rather than the exception. Any work with dependency parsing of such data needs to consider this in order to be applicable in real-world tasks. Our results showed no major differences between the monolingual and multilingual test sets, which, however, remains open to further analysis as the test data was very small.

In order to create the foundation for more comprehensive

testing of different NLP methods, two Zyrian Komi treebanks (Partanen et al., 2018a) were created as part of the Universal Dependencies (UD) project (Nivre et al., 2018). We believe these treebanks will also become important resources for linguistic research, beyond computational linguistics. Increasing their size and coverage has been a continuous effort since the beginning and resulted in various updates and releases. Part of this work has also focused on comparing the UD treebanks across different other Uralic languages, in order to ensure that the cross-linguistic treebank data remain comparable in the future, especially when new treebanks are added and the annotation scheme is taken into use in a new language (Partanen and Rueter, 2019).

Recent work of Lim et al. (2020 accepted) presents good results on dependency parsing with the use of semisupervised learning. A small initial training treebank is appended with a larger amount of plain text, which is used to learn a meta structure that improved LAS scores even by 9.3 points. If such improvement could be seen also in scenarios we have tested in our previous papers, we would be approaching the point where the result could be useful for documentary linguistics working with fieldwork data. This is particularly interesting since mixing small manually tagged data sets with larger amounts of untagged text fits exactly the scenario described in this paper: both Komi and Saami have very large non-annotated corpora and relatively small manually created resources.

Since the latest UD release contains Karelian (closely related to Finnish), Skolt Saami (closely related to North Saami) and Permiak Komi (closely related to Zyrian Komi) treebanks, the possibilities of multilingual dependency parsing between new very closely related languages is increasingly becoming possible to investigate.

## 3. Conclusions

This paper presented our ongoing work using language technology for better linguistic documentation and description of Zyrian Komi and North Saami. However, language technology can also be applied in practical projects aiming at language revitalization and language maintenance. Therefore, building any kind of language technology for an endangered language is potentially of relevance for speakers and learners of endangered languages as well as for language planners. Building language technology while paying attention to Open Source technologies and datasets ensures that at least the results will be available for the community and potentially reusable in practical applications in the future.

From the perspective of reusability, persistent archiving is very central to computational workflows in documentary linguistics, especially if multimedia data is included. Unfortunately, best practices around regularly and automatically updating the archived collections have yet to be established. Persistent identifiers are used, but conventions such as semantic versioning are still rare. Archived materials usually cannot be updated through an API, which makes it difficult to interact with the collections and to link their maintenance to more automatized workflows, like the ones we use in our projects.

## Acknowledgements

The research summarized in this paper has been funded by various organisations, among them the German Science Foundation, Kone Foundation, Paris Sciences et Lettres, RGNF-CNRS, and Volkswagen Foundation. Thierry Poibeau is supported by a Prairie (Paris Artificial Intelligence Research Institute) fellowship. Thanks to Vasily Chuprov for translating the abstract into Komi.

## 4. Bibliographical References

- Blokland, R., Partanen, N., Rießler, M., and Wilbur, J. (2019). Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA, February 26–27, 2019*, volume 2, pages 24–30. University of Colorado.
- Blokland, R., Fedina, M., Partanen, N., and Rießler, M. (2020). Spoken Komi Corpus. The Language Bank of Finland version.
- Gerstenberger, C., Partanen, N., and Rießler, M. (2017a). 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*, pages 57–66.
- Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2017b). Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology: Special Issue on Uralic Language Technology*.
- Kettunen, K., Kervinen, J., and Koistinen, M. (2018). Creating and using Ground Truth OCR sample data for Finnish historical newspapers and journals. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pages 162–169.
- Lim, K. and Poibeau, T. (2017). A system for multilingual dependency parsing based on bidirectional lstm feature representations. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 63–70.
- Lim, K., Partanen, N., and Poibeau, T. (2018a). Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues. *Traitement Automatique des Langues*, 59(3):67–91.
- Lim, K., Partanen, N., and Poibeau, T. (2018b). Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Lim, K., Yoon Lee, J., Carbonell, J., and Poibeau, T. (2020 accepted). Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Meechan-Maddon, A. and Nivre, J. (2019). How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120.
- Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bellato, S., Bengoetxea, K., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdensfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Mackentanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Mortensen, B., Moskalevskiy, B., Muischnek, K., Murawaki, Y., Müürisepp, K., Nainwani, P., Navarro Horňáček, J. I., Nedoluzhko, A., Nešpore-Běrzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúókun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Sadde, S., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulite, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Stella, A., Straka, M., Str-

- nadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niek-erk, D., van Noord, G., Varga, V., Vincze, V., Wallin, L., Washington, J. N., Williams, S., Wirén, M., Wolde-mariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Partanen, N. and Riebler, M. (2019). An OCR system for the Unified Northern Alphabet. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 77–89.
- Partanen, N. and Riebler, M. (2019). langdoc/unified-northern-alphabet-ocr: Unified Northern Alphabet OCR Ground Truth, March.
- Partanen, N. and Rueter, J. (2019). Survey of Uralic universal dependencies development. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 78–86.
- Partanen, N., Blokland, R., Lim, K., Poibeau, T., and Riebler, M. (2018a). The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.
- Partanen, N., Lim, K., Riebler, M., and Poibeau, T. (2018b). Dependency parsing of code-switching data with cross-lingual feature representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–17, jan.
- Partanen, N., Hämäläinen, M., and Alnajjar, K. (2019). Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146.
- Partanen, N. (2017). Challenges in OCR today: Report on experiences from INEL. In *Elektronnaja pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy*, pages 263–273.
- Partanen, N. (2019a). langdoc/elan-tests: Language documentation corpus validation scripts, December.
- Partanen, N. (2019b). nikopartanen/vyl-tujod-ocr: Vyl' Tujöd newspaper Ground Truth, May.
- Pirinen, T. A. (2019). Neural and rule-based Finnish NLP models—expectations, experiments and experiences. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 104–114.
- Seyfeddinipur, M., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B., Cruz, H., Drude, S., Epps, P. L., Ferreira, V., Galucio, A. V., Hellwig, B., Hinte, O., Holton, G., Jung, D., Buddeberg, I. K., Krifka, M., Kung, S., Monroig, M., Neba, A. N., Nordhoff, S., Pakendorf, B., von Prince, K., Rau, F., Rice, K., Riebler, M., Brenig, V. S., Thieberger, N., Trilsbeek, P., van der Voort, H., and Woodbury, T. (2019). Public access to research data in language documentation. *Language Documentation &*