The First Komi-Zyrian Universal Dependencies Treebanks

Niko Partanen[1], Rogier Blokland[2], KyungTae Lim[3], Thierry Poibeau[3], Michael Rießler[4]
niko.partanen@kotus.fi, rogier.blokland@moderna.uu.se,
kyungtae.lim@ens.fr, thierry.poibeau@ens.fr,
michael.riessler@uni-bielefeld.de

[1]Institute for the Languages of Finland
[2]University of Uppsala
[3]LATTICE (CNRS & ENS / PSL & Université Sorbonne nouvelle / USPC)
[4]University of Bielefeld

Abstract

Two Komi-Zyrian treebanks were included in the Universal Dependencies 2.2 release. This article contextualizes the treebanks, discusses the process through which they were created, and outlines the future plans and timeline for the next improvements. Special attention is paid to the possibilities of using UD in the documentation and description of endangered languages.

1 Introduction

Komi-Zyrian is a Uralic language spoken in the north-eastern corner of the European part of Russia. Smaller Komi settlements can also be found elsewhere in northern Russia, from the Kola Peninsula to Western Siberia. The language has approximately 160,000 speakers and, although not moribund, is still threatened by the local majority language, Russian. There is a long history of research on Komi, but contemporary descriptions and computational resources could be greatly improved. Over the last few years some larger documentation projects have been carried out on Komi. These projects have focused on the most endangered spoken varieties, while at the same time, new written resources for Standard Komi have became available.

This paper discusses the creation of two Komi treebanks, one containing written and another spoken data. Both the treebanks and the scripts used to create them are included in this paper as supplementary materials, and the treebanks are part of the Universal Dependencies 2.2 release (Nivre et al., 2018). The treebanks are called **Lattice** and **IKDP**, due to the fact that most of the work on them has been carried out at the LATTICE-CNRS laboratory in Paris, and the work has been done collaboratively with the IKDP-2[1] project, which is a continuation of earlier work that produced a language documentation corpus of Komi called IKDP (Blokland et al., 2009-2018). A comprehensive descriptive grammar of Komi with a focus on syntax is currently being written by members of the team. The present treebanks are intended to support the grammatical description.

The authors' recent research at LATTICE laboratory has focused on dependency parsing of low-resource languages, using Komi and North Saami as examples (Lim et al., 2018). The Lattice treebank was initially created for use in testing dependency parsers, and the IKDP treebank was created at a later date with the aim of also including spoken language data.

2 Language Documentation

Language documentation refers to a linguistic practice aiming at the provision of long-lasting and accountable records of speech events, usually carried out in the context of endangered languages and with the goal of understanding spoken communication beyond mere structural grammar. Himmelmann (1998) was the first to define "Documentary Linguistics" as separate from "Descriptive Linguistics", although with considerable overlap between the two. He also pays special attention to the interface between research outputs and primary data, ideally including audio and video recordings (Himmelmann, 2006). This has generally been the approach in the present work too, so that the spoken language UD corpus is directly connected to the documentary multimedia corpus

[1]https://langdoc.github.io/IKDP-2

through matching sentence IDs. This allows the treebank sentences to be connected to rich non-linguistic metadata. Additionally, the coded time-alignment in the original utterances provides information about turn-taking and overlapping at the millisecond level. The documentary corpus refers to the materials collected and processed within the language documentation activities, which are usually fieldwork-based and aim to represent various genres and speech practices, all of which are often under a threat of disappearance.

In language documentation, traditional annotation methods have mainly consisted of so-called interlinear glossing.[2] This is normally done manually or semi-manually, i.e. with little or no use of natural language processing tools (cf. Gerstenberger et al., 2016). With the available Komi data in our project, however, we wanted to apply an annotation method that would connect our work more closely to established corpus linguistics and NLP. Universal Dependencies appeared to be a very attractive annotation scheme as it aims at cross-linguistic comparability and already contains several Uralic languages. Komi-Zyrian is currently the sixth Uralic language to be included in the project.

Work with Komi complements well the developments associated with the emergence of new Uralic treebanks in 2017, with new repositories created for North Saami[3] and Erzya (Rueter and Tyers, 2018). Another noteworthy trend is that there are several treebanks currently being created for endangered languages in situations similar to that of Komi. As far as we have been able to ascertain, these are, at least: Dargwa spoken in the Caucasus (Kozhukhar, 2017), Pnar[4] spoken in South-East Asia and Shipibo-Konibo[5] spoken in Peru. The description of the last treebank mentioned does not indicate the use of language documentation materials, but as the language is very small, the context is comparable. To our knowledge, the IKDP treebank discussed here is the first treebank included in the UD release that is directly based on language documentation material. It is too early to say whether there will be more similar treebanks in the future and within what timeframe, but having more materials like these included in UD would fit into the original ideas of the multifunctional language documentation enterprise very well.

3 Methodology

The initial analysis of Komi plain text was created using Giellatekno's[6] open infrastructure (Moshagen et al., 2014), which is currently at a rather mature level for Komi. The syntactic analysis component demands the most further work, which in turn can be guided by the work on treebanks. Similar rule-based architectures have already been used for other treebanks as well. The Northern Saami and Erzya corpora, for example, seem to have been created using a similar approach. Some work has been conducted with integrating these NLP tools into workflows commonly used in language documentation (Gerstenberger et al., 2017a,b, 2016). Since these languages often lack larger annotated resources, the use of infrastructures other than rule-based ones has not been common or possible, but these workflows have been implemented in a modular fashion that would make enable the integration of other tools when they become available or reach needed accuracy.

It has been demonstrated that it is possible to convert annotations from Giellatekno's annotation scheme into the UD scheme (Sheyanova and Tyers, 2017), and this has also worked well in our case, although the exact procedure will continue to be refined while the token count of the corpus grows, which will ultimately also reveal rarer and not-yet-analysed morphosyntactic features. After starting with manually editing CoNLL-U files, the UD Annotatrix tool (Tyers et al., 2018) was adopted in January 2018, which marked the midpoint in the project's timeline. This greatly improved the annotation speed and consistency.

The treebank creation thus consisted of the following steps:

1. Sending Komi sentences to the Giellatekno morphosyntactic analyser (consisting of an FST component for morphological categories and a syntactic component using Constraint Grammar)

[2] Cf., e.g., the Leipzig Glossing Rules https://www.eva.mpg.de/lingua/resources/glossing-rules.php

[3] https://github.com/UniversalDependencies/UD_North_Sami-Giella

[4] https://github.com/UniversalDependencies/UD_Pnar-PTB

[5] https://github.com/UniversalDependencies/UD_Shipibo_Konibo-PUCP

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

2. Resolving the remaining ambiguity manually

3. Adding the missing syntactic relations manually to the UD Annotatrix

4. Automatically converting the analyzer's XPOS-tags into UPOS-tags and converting morphological feature tags into their UD counterparts

5. Manual correction and verification

The current workflow involves a rather large amount of manual work. We are interested in testing various approaches to morphological and syntactic analysis so that different (rule-based, statistic-based and hybrid) parsers can eventually replace the manual work. Some tests have already been carried out with the dependency parser used by the Lattice team in the CoNLL-U Shared Task 2017 (Lim and Poibeau, 2017) and a follow-up project (Partanen et al., 2018).

The treebank processing pipeline has been tied to several scripts and existing tools. The primary analysisis done within the Giellatekno toolkit (building on FST Morphology and Constraint Grammar), where tokenization, morphological analysis and rule-based disambiguation are tied to the script 'kpvdep'. The script returns a vislcg3 file that contains all ambiguities left after the analysis. Once the ambiguities are resolved manually, the vislcg3 file can be imported into the UD Annotatrix tool. As a final step, the Giellatekno POS-tags and morphological features are converted to follow the UD standard with a Python script, originally written by Francis Tyers[7]. A modified version of the script with the conversion pattern file is stored in **not-to-release** folder in the dev-branch of the Lattice treebank, which is the location where all development scripts of both treebanks will be maintained.

4 Data Sources and Design Principles

Most of the work on the Komi language is currently being done by collaborators of FU-Lab[8] in Syktyvkar, the capital of the Komi Republic in Russia. The work of FU-Lab, led by Marina Fedina, has been particularly exceptional, as it has resulted in a significant number of Komi-language books being digitalized, made available online[9] and converted into a linguistic corpus.[10] The corpus is currently 40 million words large, and the long-term goal is to digitalize all books and other printed texts ever published in Komi-Zyrian. The number of publications is approximately 4,500 books, plus tens of thousands of newspaper and journal issues. A significant portion of the latter are available in the Public Domain as part of the Fenno-Ugrica project of the National Library of Finland[11]. We have exclusively chosen to use openly available data for the Lattice treebank in order to ensure as broad and simple reuse as possible. The forthcoming releases will include more genres of text, such as newspaper texts and longer sections of Wikipedia articles.

All sentences in the Lattice treebank are presented in the contemporary orthography, even when they were originally published using various earlier Komi writing systems. The proportion of texts originally written in the Molodcov alphabet will rise dramatically in the next releases, as this is probably the most commonly used orthography in the upcoming texts. Storing several orthographic variants may be necessary. Conversion between systems has been carried out using FU-Lab's Molodcov converter[12]. The data originates from scanned books through text recognition, currently with loss of page coordinates. This connects to the question of how to retrieve arbitrary information from different sources that can be connected to the sentence IDs: metadata, page positions, page images, time codes and audio segments.

We considered it very important to also include spoken language in the treebank, ideally eventually covering all dialects. During the last years, one of the largest research projects investigating spoken Komi has been the IKDP project, led by Rogier Blokland in 20142016, which resulted in a large transcribed spoken language corpus (Blokland et al., 2009-2018). The IKDP treebank contains dialectal texts taken from this corpus, and since written Komi does not follow the exact same principles employed in the transcriptions, it seems problematic to mix these materials together. The orthographic conventions

[7] https://github.com/ftyers/ud-scripts/blob/master/conllu-feats.py

[8] http://fu-lab.ru

[9] http://komikyv.org

[10] http://komicorpora.ru

[11] https://fennougrica.kansalliskirjasto.fi

[12] http://fu-lab.ru/convertermolodcov

of the spoken treebank are basically similar to those used in the recent Komi dialect dictionary (Beznosikova et al., 2012), with only relatively subtle differences.What it comes to spoken features, corrections are kept and marked with the relation reparandum, but features such as pauses are not separately marked. The user can access the original archived audio, which enables a more detailed analysis of spoken phenomena if desired. In their typographic simplicity, the transcribed texts are reminiscent of some of the dialect texts published previously in various printed text collections (without the original audio recordings). The context of the spoken data here is therefore not only a faithful representation of the spoken signal, which could include also more exact phonetic transcriptions, but also the larger landscape of spoken language resources which we would like to integrate into our NLP ecosystem.

Furthermore, because local Komi speech and research communities are often conscious of orthographic norms, we wanted to draw a clear boundary between written and spoken representations. Additionally, the spoken language treebank contains a large number of Russian phrases due to code-switching, which makes it to some degree a multilingual treebank. In the IKDP treebank, Russian items are currently marked with a language tag in the misc-field, but verification that Russian annotations are consistent with monolingual Russian treebanks is a topic that requires further attention.

The sentences represent running texts and narratives, and, to a great extent, they link together into continuous larger text units. There are deviations from this in situations where individual examples have been selected in order to include instances of each dependency relation in the treebank. This was done particularly in the early stages of the treebanks when it was important to gain more understanding of how different syntactic relations are tagged consistently in UD. In the upcoming releases, occurrences of each morphosyntactic phenomena present in Komi may also be hand-picked from corpora to ensure that they occur in the treebanks, the need for which is discussed next.

5 Some Questions Arising From Komi-Zyrian

As the majority of languages in UD are larger Indo-European languages, the project does not yet include many examples of languages with very complex case systems. For example, Komi has two values of nominal case that were not yet included in the earlier documentation, namely the **approximative** and the **egressive**. One issue arising when comparing current treebanks is the cross-comparability of the case labels applied. Komi has two cases that express a path of some sort, traditionally called prolative and transitive in Komi and Uralic linguistics. These would match closely with a case label already in the UD documentation, perlative, found in Warlpiri, but the fact that there are two very similar cases already makes the labeling problematic. Differences in case labeling are related to further linguistic analyses that are possible with the corpora, as well as to parsing accuracy in multilingual scenarios. In the present treebanks, the traditional labels for Komi cases are used.

Another theoretical question arising from Komi concerns the way different cases can be combined, resulting in "double case marking". For example, it is entirely possible to use several spatial case markers linearly combined in one and the same inflected noun form, and, although this is somewhat rare, examples can be easily found even for more marginal combinations. For example, the case suffixes for elative and terminative can combine to mark subtle changes in focus: vengrija-ic̣-ed̡ Hungary-ELA-TER ‘all the way from Hungary’ (see e.g. (Bartens, 2003, 53). This raises the question of how to best annotate this in UD. Of course each combination could be labeled as a new case, which is also sometimes seen in the literature on Komi nominal case (Kuznetsov, 2012, p. 374), but this would greatly increase the number of case values that need to be documented, and most of them would be very marginal and specific to individual languages. Another solution would be to allow several case affixes to be added to one word form. However, this would only help when several cases are clearly combined and would not be useful when new spatial cases have emerged from postpositions, a phenomenon typical of Komi and Udmurt dialects.

Currently, a large portion of the cases in UD documentation are used only in Hungarian. Including more languages with large case systems, such as Uralic or Northeast Caucasian languages like Lezgian, would only increase the number of names for case values used mainly in individual languages. Eventually this also boils down

to the question of how comparable the cases in different languages actually are. Haspelmath has argued convincingly that case labels are valid only for particular languages (Haspelmath, 2009, 510), and the issue probably cannot be explicitly solved within UD either, but for the sake of usability of treebanks and their suitability for multilingual NLP applications, some harmonization would seem desirable. One alternative could be to create a higher layer of mapping that connects language-specific labels to broader shared categories. In this way, both Komi cases expressing a path could be connected to a concept of *movement along a path*, but the language-specific nuances would not be lost.

6 Conclusion and Further Work

The written and spoken treebanks have 1389 and 988 tokens, respectively. Due to their small size, they have not been split into test and development sets. Based on this experience, it already seems clear that providing annotations in this framework has several advantages compared to traditional methods used in language documentation projects. The main benefit is the comparability between different languages, and also straightforward licensing and distribution within UD framework.

It can be argued that tagging according the UD principles is necessarily a compromise, and that it may not express all particularities of individual languages. One possible way to solve this problem is to include further annotations in the misc-column. Another possible approach would be to provide different parts of the documentary corpus with varying degrees of annotations. In any case, based on our experience, we would strongly encourage endangered language documentation projects to take a small segment of their materials and add to it an additional layer of annotations in the Universal Dependencies framework. Language documentation data is usually stored in archives that require access requests. This is not very compatible with openly available treebanks. Still, it should be possible to collect small subsets of materials with the clear intention and permission for these recordings to be openly licensed, or to use texts old enough that they are copyright free.

New material is currently being brought into the Lattice treebank. The main genres obtained from Fenno-Ugrica collection are newspaper texts, nonfiction works and schoolbooks. Samples of these, along with some larger Wikipedia texts, will be included in the next UD release 2.3. The next phase of the IKDP treebank will include individual texts from the Komi recordings made by Eric Vászolyi in the 1950s and 1960s (Vászolyi-Vasse, 2003), which the present authors have acquired permission to re-publish electronically. These texts originate from a time and place of intensive language contact between Komi-Zyrian and Tundra Nenets, what makes them a particularly interesting target for further study.

One possibly useful addition to the treebank could be English glosses in the misc-field, since many linguists are used to working with data from endangered languages in a format like this. The English gloss could contain a contextual translation of the lemma, for example, which would make the sentences in the treebank much more accessible to different linguistic audiences.

In terms of size, the target is to reach 5,000 tokens in both treebanks during 2018, and to increase this to 20,000 in the first half of 2019. Our long-term goal is to create a resource that would contribute to research on Komi and provide better resources for Natural Language Processing of this language, which has yet to receive sufficient attention in computational linguistic research.

7 Acknowledgements

We want to thank the reviewers for their useful comments. This work has been developed in the framework of the LAKME project funded by a grant from Paris Sciences et Lettres (IDEX PSL reference ANR-10-IDEX-0001-02). Thierry Poibeau is partially supported by a RGNF-CNRS (grant between the LATTICE-CNRS Laboratory and the Russian State University for the Humanities in Moscow). Kyungtae Lim is partially supported by the ERA-NET Atlantis project. Niko Partanen's work has been carried out at the LATTICE laboratory, and besides Partanen, both Rogier Blokland and Michael Rießler collaborate within the project Language Documentation meets Language Technology: the Next Step in the Description of Komi, funded by the Kone Foundation. Thanks to Jack Rueter for numerous discussions on Komi and Erzya, and to Alexandra Kellner for proofreading the paper.

References

Raija Bartens. 2003. Kahden kaasuspäätteen jonoista suomalais-ugrilaisissa kielissä. In Bakró-Nagy Marianne and Károly Rédei, editors, *Ünnepi kötet Honti László tiszteletére*, pages 46–54. MTA, Budapest.

L.M. Beznosikova, E.A. Ajbabina, N.K. Zaboeva, and R.I. Kosnyreva. 2012. *Komi sërnisikas kyvčukör. Slovar dialektov komi âzyka: v 2-h tomah*. Institut âzyka, literatury i istorii Komi naunogo centra Uralskogo otdeleniâ Rossijskoj akademii nauk, Syktyvkar.

Rogier Blokland, Marina Fedina, Niko Partanen, and Michael Rießler. 2009-2018. IKDP. In *The Language Archive (TLA): Donated Corpora*. Max Planck Institute for Psycholinguistics, Nijmegen.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 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. ACL.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2017b. Instant annotations: Applying NLP methods to the annotation of spoken language documentation corpora. In *Proceedings of the 3rd International Workshop on Computational Linguistics for Uralic languages*, pages 25–36. ACL.

Martin Haspelmath. 2009. Terminology of case. In Andrew Spencer and Andrej L. Malchukov, editors, *The Oxford handbook of case*, pages 505–517. OUP, Oxford.

Nikolaus Himmelmann. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Ulrike Mosel, and Nikolaus Himmelmann, editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter, Berlin.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Alexandra Kozhukhar. 2017. Universal dependencies for Dargwa Mehweb. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 92–99. ACL.

Nikolay Kuznetsov. 2012. Matrix of cognitive domains for Komi local cases. *Journal of Estonian and Finno-Ugric Linguistics*, 3(1):373–394.

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018. 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*. ELRA.

KyungTae Lim and Thierry Poibeau. 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. ACL.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 71–77. ELRA.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić,

Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 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.

Niko Partanen, KyungTae Lim, Michael Rießler, and Thierry Poibeau. 2018. Dependency parsing of code-switching data with cross-lingual feature representations. In *Proceedings of the 4th International Workshop on Computational Linguistics for Uralic languages*, pages 1–17. ACL.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *Proceedings of the 4th International Workshop on Computational Linguistics for Uralic languages*, pages 106–118. ACL.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75. ACL.

Francis M Tyers, Mariya Sheyanova, and Jonathan North Washington. 2018. UD Annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17. ACL.

Eric Vászolyi-Vasse. 2003. *Syrjaenica: Narratives, folklore and folk poetry from eight dialects of the Komi language. Vol. 1, Upper Izhma, Lower Ob, Kanin Peninsula, Upper Jusva, Middle Inva, Udora.* Savariae, Szombathely.