Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

On searchable Mordvin corpora at the Language Bank of Finland, EMERALD

Jack Rueter
University of Helsinki, Finland

jack.rueter@helsinki.fi

Abstract

This paper provides a brief background to the development of searchable Erzya and Moksha corpora on the Fin-CLARIN/Language Bank of Finland Korp server[1] with a special emphasis on Erzya-Moksha Electronic Resources[2] And Language Diversity (EMERALD) [cf. Rueter 2020a] and the nature and structure of the Electronic Resource Moksha-Erzya (ERME[3])[4] and ERME version 2[5].

It points to important players in the development of searchable corpora for Mordvin languages as of 2024. It briefly introduces coexisting corpus materials featuring the Erzya and Moksha languages at the Language Bank of Finland [cf. Rueter and Partanen 2019]. It also provides an illustration of the metadata attributed to each piece in the ERME corpora as well as a description of the morphosyntactic annotation adopted in ERME.

Finally, it makes suggestions for future enhancement and extensions of ERME as well as its implications for other research corpora [cf. Rueter 2023].

keywords

Erzya, Moksha, monolingual corpus, parallel corpus, morphosyntactic annotation, metadata attribution

INTRODUCTION

The Mordvin languages, Erzya and Moksha, belong to the Finno-Ugric branch of the Uralic language family. These two languages are spoken by approximately half a million people, who have traditionally lived in the Volga Basin of what is now known as the Russian Federation [cf. Sarv 2002; Rueter 2013]. Documentation of these two languages rests on the work of native speakers and foreigners, alike. We are still building traditions of language documentation for languages of the World, and the following presents development in this endeavor at the Language Bank of Finland.

I EMERALD

EMERALD and the ERME corpora provide an answer to a lack of openly searchable documentation of the Erzya and Moksha languages. A generation of researchers has been born who cite publishers instead of actual authors. Hence, they ignore dialectal backgrounds of writers when dealing with the literary language. When looking for variation, they fail to distinguish original language materials from translations, seasoned language users from

[1] https://korp.csc.fi (forthcoming movement to https://www.kielipankki.fi/korp)

[2] Associated with a grant from the Kaisi and Kaino Heikkilä Foundation 2016 (Erzya-Moksha Electronic Resources And Language Diversity - Fieldwork and Early Literary Texts = EMERALD-FELT), initiating the introduction of Heikki Paasonen's eight-volume Mordwinische Volksdictung to the Language Bank of Finland Korp Server.

[3] In Erzya èr′me [эрьме] is ‘wealth’.

[4] ERME 2017: urn:nbn:fi:lb-201407306 (licensing is shown in the urn files)

[5] ERME version 2, 2023: urn:nbn:fi:lb-2023021601

unmonitored newcomers. Researchers sought no alignment between fieldwork research and literary language development in terms of spatial-temporal dimensions.

EMERALD and ERME is intended to provide people searching Erzya and Moksha corpora with better access to what they are investigating, and it gives them open information on source metadata. This includes six types of information:

(1) bibliographic information for citations;
(2) to-the-page and to-the-sentence citation information;
(3) page, word and character counts;
(4) timestamps and geographical identifiers for temporal and spatial location of the author;
(5) visualization of morphosyntactic structure, and
(6) visualization of geographical positioning of the author and literary piece.

To-the-page/sentence citation information, means that the researchers fortunate enough to have access to physical or digital source materials are able to locate a given sentence by simply following reference to the text provided for each sentence – no cryptic reference to a closed search engine identifier. Counting characters, words and pages might provide us with a better construal of individual pieces and their representation of a given language. Morphosyntactic structure is illustrated in Universal Dependencies[6] treebank style and is made possible through rule-based analysis and disambiguation [cf. Rueter et al. 2020].

Short comings of the corpus are that license limits the context to a "necessary one", which in the present situation, means the paragraph is the largest context, as opposed to, let us say, 80 keystrokes left and right of a given concordance point.

II Background

Work has been conducted in Mordvin studies since 1705 [see Witsen], and the two Mordvin literary languages Erzya and Moksha have developed tremendously over the past two centuries. Despite the fact that the outside world has frequently classified these two language forms as supradialects of the same language, early grammarians have noted that fluency in one language does not necessarily imply fluency in the other [cf. Ornatov″ 1838[7]; Gabelentz 1839]. The collection of vernacular texts begins in the late 1800s, which in addition to work published by Evsev′ev (1892) and Paasonen (1891, 1941), is augmented by the popularization of Erzya and Moksha vernaculars, most prominently in the 1920s and 1930s [cf. Foley, 2007]. Even though collections of words, texts and grammars have been published since the beginning of the 18th century, it is not until the 2010s that the first searchable corpora have been made openly available to the research and language communities, and the general public.

The concepts of availability and accessibility stemmed from two developments. On the one hand, there was the construction of the University of Helsinki Language Corpus Server (UHLCS) beginning in the early 1990s [cf. Broeder et al. 2000; Suihkonen 2000; Suihkonen 2003; Koskenniemi et al. 2007], which has since transferred the Language Bank of Finland and is now returning to the University of Helsinki. On the other, there was the Language Programme of the Kone Foundation, whose outcomes have resonated in Uralic language description,

[6] https://universaldependencies.org/
[7] https://rueter.github.io/emerald/historical-mordvin-grammars/docs/ornatov-pavel_mordovskaja-grammatika_1838.html

documentation and preservation no less than 2012–2020 [Rueter 2014; Hakkarainen 2017; Jauhiainen et al. 2021; Hakkarainen et al. 2023[8]].

2.1 Mordvin Corpora at the Language Bank of Finland

Soon monolingual corpora with multiple languages appeared. These included the outcomes of the National Library of Finland «Kindred Language» Pilot (2012–2015)[9], the initial digitization of endangered language materials from the 1920s and 1930s, readers, non-central news and enlightenment media available with licensing through Fenno-Ugrica[10], and the «Suki» project (Finno-Ugric Languages and the Internet)[11], which scraped the net for Uralic language texts resulting in Wanca[12], and the first ERME corpus, and a demo corpus for research and fieldwork texts from the Finno-Ugrian Society publication series: SUS[13]. Subsequently, came morpho-syntactically annotated corpora: Parallel Bible Verses for Uralic Studies (PaBiVUS[14]), and the Universal Dependencies Uralic UD-2.10[15], corpora of manually annotated corpora, featuring Uralic languages. In March of 2023, ERME version 2 appeared with over two million tokens of Erzya and 800 thousand tokens of Moksha. Autum 2023, saw the induction of more parallel corpora into the Language Bank of Finland workflow: a children's book about war published first in 1938: Uspenskij[16], a school reader for natural sciences 1939 and 1940: Tetûrev[17], a childrens book portraying the ideal citizen, translated in the 1950s and 1960s: Morozov[18], and a book providing historical information on Finland written in the 1990s: Finland Yesterday and Today[19].

The corpora, at present, may be interpreted as representative of the two Mordvin literary languages, Moksha and Erzya, but a possible third language form – Shoksha[20] may find its way into the data set in the future.

The language materials can be divided into three types. There are historical wordlists, monolingual corpora and parallel corpora. Some have automated morpho-syntactic annotation, others do not.

2.1.1 Word lists

Although the first word lists for the Mordvin languages date back to1705 [Witsen], the Language Bank of Finland provides access to the Erzya and Moksha Mordvin Word List Corpus (UHLCS)[21]. The corpus consists of two lists: Erzya (23,500 words), Moksha (300 words) from the Bishop Damaskin collection commissioned by Catherine the Great, 1785 [see Estill, 1999; Estill, 2004; Feoktistov, 1971].

---

[8] https://www.kansalliskirjasto.fi/fi/blogi/fenno-ugrica-10-vuotta-vahemmistokielten-asemaa-tukemassa-ja-niihin-kohdistuvaa-tutkimusta
[9] https://www.kansalliskirjasto.fi/fi/projektit/sukukielten-digitointiprojekti
[10] urn:nbn:fi:lb-2014073056
[11] http://suki.ling.helsinki.fi/wanca/
[12] urn:nbn:fi:lb-2019052401
[13] urn:nbn:fi:lb-2016092001
[14] PaBiVUS: urn:nbn:fi:lb-2020021121
[15] UD v2.10: urn:nbn:fi:lb-2022061001 (see also forthcoming UD v2.13 http://urn.fi/urn:nbn:fi:lb-2024031207)
[16] Uspenskij: urn:nbn:fi:lb-2023042426
[17] Tetûrev: urn:nbn:fi:lb-2023042421
[18] Morozov: urn:nbn:fi:lb-2023082102
[19] Finland, Yesterday and Today: urn:nbn:fi:lb-2023041801
[20] Shoksha has, through time, been aligned with both the adjacent Moksha and the geographically distant but, presumably, genetically closer Erzya, (cf. Olga Erina & Jack Rueter. (2018, February 5). Šokša Kolčoznikin' val'gij 1932–1933. https://zenodo.org/badge/latestdoi/120288368), (https://fennougrica.kansalliskirjasto.fi/browse?value=fi %3DŠokša)
[21] Historical Mordvin word lists: urn:nbn:fi:lb-2014032611

2.1.2 Text corpora

The text corpora for Mordvin research is gradually becoming annotated, but there are still many other hurdles to overcome before they can equally contribute to the concept behind EMERALD. While the Fenno-Ugrica materials provide bibliographical reference to individual pieces of source literature for all sentences, the ideal situation would see page reference and morphological analysis. The Wanca materials, unfortunately, lack both bibliographical reference and morpho-syntactic annotation. This, however, has to do with the vast proportion of genre variety in Wanca and the primitive state of automated morpho-syntactic analysis for many of the minority Uralic languages when it was published. New versions of PaBiVUS, Uralic UD, Finno-Ugrian Society research corpora and smaller parallel corpora are gradually being adapted to more extensive annotation.

III CORPORA ANNOTATION À LA EMERALD

The description provided in every part of the ERME corpora is intended to improve the research community's comprehension of the languages. This involves painstaking preparation of all TEI-compatible[22] XML heads and sentence elements in the materials.

Each head comes with metadata covering bibliographic information both separately and as single entities. This means that author, title, publisher, etc. data are provided in the international library conversion for Cyrillic to Latin script (ISO-9) as individual attributes, and a conglomerate bibliography, see Figure 1, below. Further attributes provide information on the number of pages, words and characters as well as spatial and temporal delimiters for locating the author in time and space. The timestamps with explicit expression of when the work was completed and latitude-longitude markers for designating the author's place of birth will hopefully provide for the alignment of literary language development with previous and forthcoming stages of language documentation, e.g., fieldwork (1880–), publications in various media (1821–). Additionally, information is also given, where possible, on proofreaders and editors of both the original print and the electronic version for finer granularity of each individual corpus.

```
<text id="Ěrâmonʹ pinkst" author="Altyškin, Viktor" genre="story"
bibliog="Altyškin, Viktor. 1986: Ěrâmonʹ pinkst. Mordovskoj knižno
j izdatelʹstvasʹ. In: Ěrâmonʹ pinkst. pp. 1–48. – Saransk." publis
her="Mordovskoj knižnoj izdatelʹstvasʹ" publication_place="Saransk
" publication="Ěrâmonʹ pinkst" publication_year="1986" no_of_pages
="48" word_count="11,255" character_count="67,954" corrector="_" e
_corrector="Motalina, Tatʹâna 2000; Rueter, Jack 2023" page_range=
"1–48" datefrom="19860101" dateto="19861231" iso_lang="myv" timefr
om="000000" timeto="235959" _geo_author_origin="|MR_MREM;RU;54.333
056;45.586389|">
```

Figure 1. The head element in an ERME piece

In Figure 1., there is an id(entifier), the name of the piece, followed by author, genre, bibliog(raphic data), publisher, publication (especially necessary when piece is a part of a larger publication), publication year, number of pages, word count, character count, corrector, e-corrector (corrector of digital version), page range, datefrom, dateto, timefrom, timeto (essentially timestamps), ISO-639 language code and geo author origin (latitude and longitude of author's place of birth).

[22] Text Encoding Initiative: https://tei-c.org/

Each sentence bears its own additional set of metadata which satisfy three demands: location in the piece; possible translation(s), see Figure 2, below, and automated morphosyntactic annotation, Figure 3.

```
<sentence id="Кирдажт:chap1:para1:sent3:pg0" pgno="0" text="Удоман
ь пачк нузякссто зярыяксть аволдась кедьсэнзэ, бажась панемс, но т
онат тандадыльть а куватьс." text_eng="In his sleep he lazily wave
d at them with his hand. He wanted to chase them away but they cou
ld not scared away for long." text_fin="Puoliunessa hän huitaisi l
aiskasti kädellään ja yritti ajaa ne pois, mutteivät ne pelästynee
t pitkään.">
```

Figure 2. ERME sentence head

In Figure 2., the id(entifier) attribute bears the value: ‘name of piece’ + ‘chapter number’ + ‘paragraph number’ + ‘sentence number’ + ‘page number’. A separate page number attribute is also given along with subsequent attributes for text (in the original), text_eng (possible English translation), and text_fin (possible Finnish translation). The id attribute provides us with the precise location of the sentence in the piece, i.e., page number, chapter, paragraph, and sentence identification, which will be helpful to the fortunate with access to the physical or virtual literary pieces.

```
<sentence id="Кирдажт:chap1:para1:sent3:pg0" pgno="0" text="Удомань пачк нузякссто зярыяксть аволдась кедьсэнзэ, бажась панемс, но тонат тандадыльть а куватьс." text_
eng="In his sleep he lazily waved at them with his hand. He wanted to chase them away but they could not scared away for long." text_fin="Puoliunessa hän huitaisi lai
skasti kädellään ja yritti ajaa ne pois, mutteivät n
e pelästyneet pitkään.">
Удомань 1       удома   NOUN    N       Case=Gen|Definite=Ind|Number=Plur,Sing  2       dep     _       CGdephead=#1~&gt;2|CGdeprel=@&gt;P|GTtags=SP,Gen,Indef
пачк    2       пачк    ADP     Po      AdpType=Post    5       case    _       CGdephead=#2~&gt;5|CGdeprel=@ADVL&gt;|GTtags=Po
нузякссто       3       нузякс  NOUN    N       Case=Ela|Definite=Ind|Number=Plur,Sing  5       obl     _       CGdephead=#3~&gt;5|CGdeprel=@ADVL&gt;|GTtags=SP,Ela,In
def
зярыяксть       4       зярыяксть       ADV     Adv     NumType=Mult    5       advmod  _       CGdephead=#4~&gt;5|CGdeprel=@ADVL&gt;|GTtags=Iter
аволдась        5       аволдамс        VERB    V       Mood=Ind|Number[subj]=Sing|Person[subj]=3|Tense=Past    0       root    _       CGdephead=#5~&gt;0|CGdeprel=@F
MV|GTtags=Ind,Prt1,ScSg3
кедьсэнзэ       6       кедь    NOUN    N       Case=Ine|Number=Plur,Sing|Number[psor]=Sing|Person[psor]=3      5       obl     _       CGdephead=#6~&gt;5|CGdeprel=@&
lt;ADVL|GTtags=SP,Ine,PxSg3|SpaceAfter=No
,       7       ,       PUNCT   CLB     _       6       punct   _       CGdephead=#7~&gt;6|CGdeprel=@X|GTtags=CLB
бажась  8       бажамс  VERB    V       Mood=Ind|Number[subj]=Sing|Person[subj]=3|Tense=Past    0       root    _       CGdephead=#8~&gt;5|CGdeprel=@FMV|GTtags=Ind,Pr
t1,ScSg3
панемс  9       панемс  VERB    V       VerbForm=Inf    9       dep     _       CGdephead=#9~&gt;9|CGdeprel=@FS~&lt;FMAINV|GTtags=Inf|SpaceAfter=No
,       10      ,       PUNCT   CLB     _       11      punct   _       CGdephead=#10~&gt;11|CGdeprel=@X|GTtags=CLB
но      11      но      CCONJ   CC      _       8       dep     _       CGdephead=#11~&gt;8|CGdeprel=@CVP|GTtags=
тонат   12      тона    DET     Det     Case=Nom|Definite=Ind|Number=Plur|PronType=Dem  0       dep     _       CGdephead=#12~&gt;0|CGdeprel=@X|GTtags=Dem,Pl,Nom,Inde
f
тандадыльть     13      тандадомс       VERB    V       Aspect=Hab|Mood=Ind|Number[subj]=Plur|Person[subj]=3|Tense=Past 0       root    _       CGdephead=#13~&gt;5|CG
deprel=@FMV|GTtags=Ind,Prt2,ScPl3
а       14      а       PART    Pcle    Polarity=Neg    14      dep     _       CGdephead=#14~&gt;14|CGdeprel=@NEG&gt;|GTtags=Neg
куватьс 15      куватьс ADV     Adv     Case=Ill        14      advmod  _       CGdephead=#15~&gt;14|CGdeprel=@&lt;ADVL|GTtags=Ill|SpaceAfter=No
.       16      .       PUNCT   CLB     _       5       punct   _       CGdephead=#16~&gt;5|CGdeprel=@X|GTtags=CLB
:
```

Figure 3. ERME sentence CoNLL-U-type structure

In Figure 3., we can observe one single disparity between UD CoNLL-U[23]. On the Korp server tradition has placed the word form before the token counter, i.e., columns 1 and 2 have swapped places. Otherwise, column 3 is lemma, 4 UD part of speech, 5 external part of speech (Giella part of speech), 6 alphabetized morphological features, 7 dependency head, 8 dependency, 9 extended dependencies, 10 miscelanious.

The CoNLL-U-type annotation reveals the state of automated rule-based morphosyntactic annotation. As can be seen in the dependency encoding ‹dep›, this is an ever-developing dimension of Mordvin language description. Rule-based morphosyntactic analyzers are being developed for the Erzya and Moksha languages on the Giella infrastructure, where they are automatically rendered reusable as spell checkers and the motor for morphologically savvy dictionaries [Rueter et al. 2020][24]. Multiple use of the rule-based description extends to the Universal Dependencies projects [see Rueter and Tyers, 2018; Zeman et al. 2023-11], and the

[23] CoNLL-U: https://universaldependencies.org/docs/format.html
[24] https://github.com/giellalt/lang-myv, https://github.com/giellalt/lang-mdf

shallow-transfer machine translation projects at Apertium [cf Rueter and Hämäläinen 2020; Rueter 2020a].

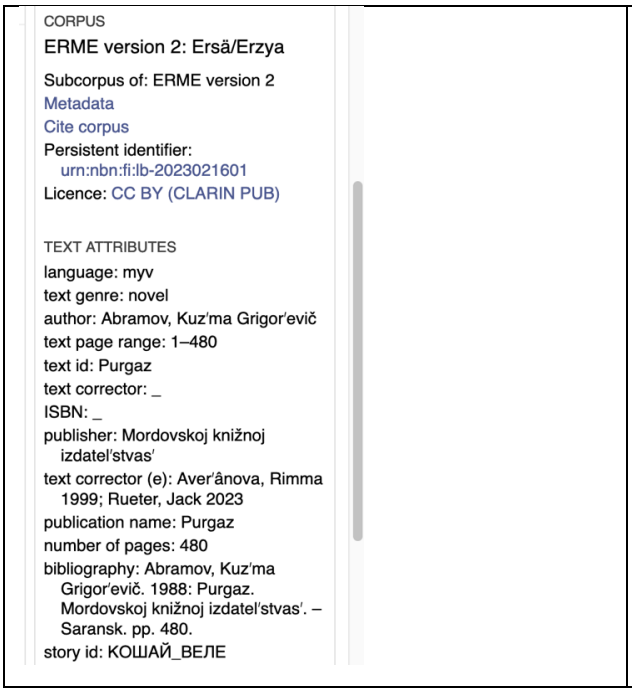

Figure 4a. Korp right margin, top.

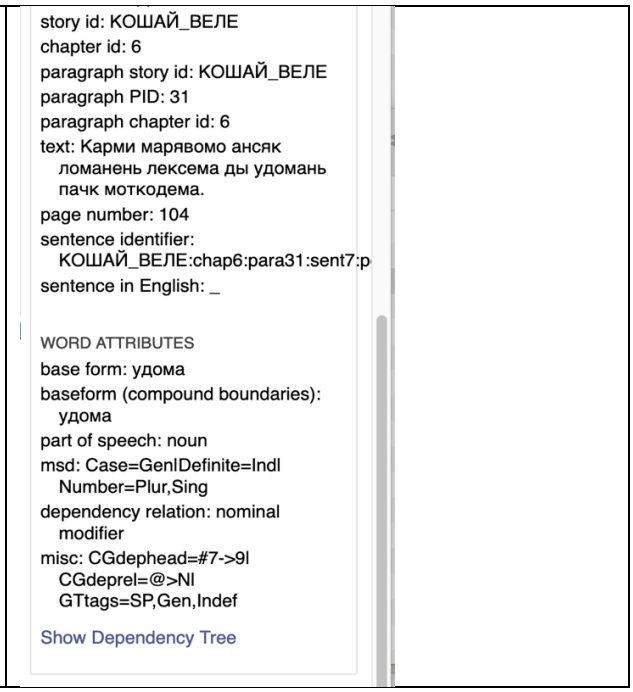

Figure 4b. Korp right margin, bottom.

Figures 4a and 4b, above, show how this metadata appears in the right margin of the Korp interface. In addition to the metadata described earlier, including CoNLL-U, there is a final button for selecting a visualization of the dependency tree – please, try it.

3.1 Monolingual corpora

The Electronic Resource for Moksha and Erzya (ERME) version 2 corpora consist of original-language texts in the two Mordvin languages, Erzya (over 2,041,000 tokens), and Moksha (over 855,000 tokens). These corpora are rich in metadata and automated morpho-syntactic annotation, as described above. A second, but smaller pair of original language corpora are found in the Uralic UD v2.10 corpora with manually annotated morpho-syntactic analyses for each sentence: Erzya (over 17,000 tokens), Moksha (over 3,000 tokens). Additionally, there are the large corpora without morpho-syntactic annotation: Fenno-Ugrica: Erzya (nearly 900,000 tokens), Moksha (nearly 618,000 tokens), and Wanca: Erzya (over 355,500 tokens), Moksha (over 257,500 tokens). The latter two corpora sets contain both original language and translated texts, a dimension that must be dealt with in EMERALD annotation in the near future. Both the Fenno-Ugrica and Wanca have extensive representation of other Uralic languoids, Fenno-Ugrica – 10, Wanca – 29.

3.2 Parallel corpora

Implicit parallel corpora research in Mordvin studies dates back to an 'Attempt at a Grammar of Mordvin (read: Erzya)'[25] by Herr Conrad von Der Gabelentz 1839 where he made a meticulous study of the morphological structure as attested in the Erzya-language Gospel from 1821[26]. At the Language Bank of Finland, part of the same Gospel texts and their modern

[25] Herr Connan von Der Gabelentz: https://rueter.github.io/emerald/historical-mordvin-grammars/historical-mordvin-grammars/docs/gabelentz_hcvonder-versuch-einer-mordwinischen-grammatik-1838-39.html
[26] The Erzian New Testament from 1827 including the Gospel of 1821: https://rueter.github.io/emerald/myv_new-testament_1821-1827/myv_bible-new-testament_1821-1827.html

equivalents are made available in the Parallel Biblical Versus for Uralic Studies (PaBiVUS) corpus. This corpus actually contains the 27 books of the New Testament in 10 languages, three of which – Erzya, Moksha and Finnish – have been automatically annotated, but in future versions this annotation will be improved upon and extended to the other languages, i.e., Karelian, Khanty, Komi, Mansi, Olonets-Karelian, Permian Komi, Russian, Udmurt, Veps.[27] PaBiVUS was introduced in 2020 with morpho-syntactic annotations for a few of the languages. Since then, more parallel corpora with morpho-syntactic annotations have been introduced to the Language Bank of Finland's Korp server. The ones presently available as this article goes to press are children's book 'Four Battles' by Uspenskij[28] and a Christmas Gospel, which has text-to-speech enhancement. There are also another 3 parallel corpora that ought to be in the Korp search engine by summer 2024: 'Environment and Natural Science for 3rd Grade, Part 1'[29], 'Finland, Yesterday and Today'[30], 'Pavlik Morozov'[31].

IV FUTURE WORK

In the future, enhancement of the automated annotation process will be continued on the rule-based morphosyntactic analyzer and disambiguation rules constructed on the Giella infrastructure. Extended versions will be published for PaBiVUS, Uralic UD v2.13, and ERME with extensions to include pieces from the Erzya and Moksha journals (1929–2000s), Sâtko and Mokša, Erzya and Moksha respectively. Lexical work will come into the picture for improvement in translation [cf. Hämäläinen et al. 2021]. The next implementation of Korp will also have a Russian-language interface, something that will help Erzya and Moksha native speakers in their use of the search engine.

V ACKNOWLEDGEMENTS

I would like to thank all people and institutions that have helped to make the development of these corpora possible. This include sponsors Kone Foundation, Kaisi ja Kaino Heikkilä Rahasto Finno-Ugrian Society, Employers the University of Helsinki (FinCLARIN/Language Bank of Finland), University of Turku (DigiLang), Open infrastructures Giella (Giellatekno & Divvun), Apertium, the authors of all these wonderful language materials, my coauthors and coworkers and the individuals both native and not who have improved my working knowledge of these ever-important languages. Most of all, I am indebted to my family, who allows me the privilege to carry on with this kind of developmental research.

References

Broeder, D., Suihkonen, P. M., & Wittenburg, P. (2000). Developing a standard for meta-descriptions of multimedia language resources. teoksessa *Proceedings of the Linguistic Exploration Workshop on Web-Based Language, Documentation And Description* University of Pennsylvania. http://morph.ldc.upenn.edu/exploration/expl2000/papers/

Erina, Olga & Rueter, Jack. (2018, February 5). Šokša Kolxoznikin′ val′gij 1932–1933. (Proofread Shoksha-language materials from the collection of .pdf copies for issues of the Kolxoz'nikin' val'gij newspaper 1932 (7), 1933 (16). (https://fennougrica.kansalliskirjasto.fi/browse?value=fi %3DŠokša) https://zenodo.org/badge/latestdoi/120288368)

ERME = Rueter, J., & Yerina, O. (2014). *Erzya and Moksha Extended Corpora*. Bibliographic metadata, no lemmatization or morphosyntactic analysis. Online: urn:nbn:fi:lb-201407306

[27] Neither of the Mari literary languages are in the 2020 release of PaBiVUS, but the next release, possibly autumn 2024 will see both Hill Mari and Meadow & Eastern Mari. Saami language materials still require negotiation. At such time, it would also be feasible to add Estonian, Hungarian and perhaps some adjacent languages as well.

[28] A children's book by Uspenskij published in several languages of the Soviet Union from 1938: http://urn.fi/urn:nbn:fi:lb-2023042426

[29] Tetûrev, Erzya (1939), Moksha (1940): '*Environment and Natural Science for 3rd Grade, Part 1*'. http://urn.fi/urn:nbn:fi:lb-2023042421

[30] Häkkinen (1997): '*Finland, Yesterday and Today*'. http://urn.fi/urn:nbn:fi:lb-2023041801

[31] Gubarev (1947 Russian): '*Pavlik Morozov*' http:// urn.fi/urn:nbn:fi:lb-2023082102

ERME2 = Rueter, J., & Erina, O. (2023). *Erzya and Moksha Extended Corpora version2*. Bibliographic metadata, and automatic morpho-syntactic analysis, with over 2 million tokens in Erzya and 800,000 tokens in Moksha. Online: urn:nbn:fi:lb-2023021601

Estill, Dennis (1999). Erzya and Moksha Mordvin Word List Corpus (UHLCS). Erzya (23,500 words), Moksha (300 words). Online: http://urn.fi/urn:nbn:fi:lb-2014032611

Estill, Dennis (2004). Diachronic change in Erzya word stress. Suomalais-Ugrilaisen Seuran Toimituksia 246. – Helsinki.

Evsev′ev, M. (1892). *Mordovskaâ svad'ba* (Мордовская свадьба) / [М. Е. Евсевьев]. - [Санкт-Петербург : Тип. С. Н. Худекова, 1892].

Feoktistov, A. P. (1971). *Russko-mordovskij slovar', Iz istorii otečestvennoj leksikografii.* Izdatel′stvo Nauka, Moskva.

Foley, K. (2007). Literacy and Education in the Early Soviet Union. Online version: https://web.archive.org/web/20120313183610/http://www.russia.by/russia.by/print.php?subaction=showfull\&id=1190296667\&archive=\&start_from=\&ucat=22\&

Finland = Luutonen, J., Rueter, J., Axelson, E. (2023). Finland, Yesterday and Today, parallel corpus, Korp Version in Erzya, Finnish, Komi-Zyrian, Meadow and Eastern Mari, Moksha, Russian, Udmurt. Online: urn:nbn:fi:lb-2023041801

Gabelentz, Herr Conon von der (1839). Versuch einer Mordwinischen Grammatik. *Zeitschrift für die Kunde des Morgenlandes*. II. 2–3. Göttingen: Druck und Verlag der Dieterlichschen Buchhandlung. 235–284, 383–419.

Hakkarainen, Jussi-Pekka; Puura, Ulriikka & Partanen, Niko (2023-06-06). Fenno-Ugrica 10 vuotta. Vähemmistökielten asemaa tukemassa ja niihin kohdistuvaa tutkimusta mahdollistamassa, osa yksi. (Blog) Online: https://www.kansalliskirjasto.fi/fi/blogi/fenno-ugrica-10-vuotta-vahemmistokielten-asemaa-tukemassa-ja-niihin-kohdistuvaa-tutkimusta

Hakkarainen, Jussi-Pekka (2017-01-24). Sukukielten digitointiprojekti: Jatkohankkeen loppuraportti 2014-2016. Online: https://urn.fi/URN:ISBN:978-951-51-2910-9

Hämäläinen, M., Alnajjar, K., Rueter, J., Lehtinen, M., & Partanen, N. (2021). An Online Tool Developed for Post-Editing the New Skolt Sami Dictionary. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference* (pp. 653-664). (Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference). Lexical Computing CZ s.r.o..

Jauhiainen, Tommi; Jauhiainen, Heidi & Lindén, Krister (2021). Suomalais-ugrilaiset kielet ja internet -projekti 2013–2019. In Mika Hämäläinen, Niko Partanen and Khalid Alnajjar (eds.), *Multilingual Facilitation.* – Helsinki. pp. 228–247. DOI: https://doi.org/10.31885/9789515150257

Koskenniemi, K., Suihkonen, P., & Vitie, E. (Toimittajat) (2007). *Multilingual Resource Collection of University of Helsinki Language Corpus Server: the organization of the UHLCS at the CSC*. Department of General Linguistics, University of Helsinki. http://www.ling.helsinki.fi/uhlcs/csc-corpora/csc-corpora-main.html

Morozov = Vitali Gubarev (1947) = Luutonen, J., Rueter, J., Axelson, E. (digital eds.) (2023). Pavlik Morozov, parallel corpus, Korp in Chuvash, Erzya, Finnish, Hill Mari, Hungarian, Khanty, Komi-Permyak, Komi-Zyrian, Mansi, Meadow and Eastern Mari, Russian, Tatar, Udmurt. Online: urn:nbn:fi:lb-2023082102

OMD = Očerki mordovskih dialektov [Essays of the Mordovian dialects]. Saransk: Mordov. kn. izd-vo Publ., Vol. 1. 1961. 396 p.; Vol. 2. 1963. 448 p.; Vol. 3. 1963. 276 p.; Vol. 4. 1966. 382 p.; Vol. 5. 1968. 399 p. (In Russian)

Paasonen, H. (1891). *Proben der mordwinischen Volkslitteratur*. Gesammelt von H. Paasonen. Erster Band. Erzyanischer Theil. Finnisch-Ugrischer Gesellschaft. Suomalais-Ugrilaisen Seuran aikakauskirja. Journal de la Sociêtê finno-ougrienne. IX. Helsingissä. Suomalaisen Kirjallisuuden Seuran kirjapainossa.

Paasonen, H. (1941). *Mordwinische volksdichtung* Gesammelt von H. Paasonen, herausgegeben un übersetzt von Paavo Ravila. *III. BAND.* Suomalais-ugrilaisen Seuran Toimituksia LXXXIV. Helsinki: Suomalais-Ugrilainen Seura.

Ornatov″, P. (Ортановъ, Павелъ) (1838). *Мордовская грамматика* /составленная на нарѣчій мордвы мокши Тамбовской семинаріи профессоромъ, магистромъ, Павломъ Орнатовымъо. Москва: Въ Сунодальной типографіи, 1838.-106 с.

Partanen, N., Horváth, C., Kellner, A., Bradley, J., Janurik, B., & Rueter, J.. *Rinnakkaiskorpus L. Uspenskin kirjasta "Neljä taistelua"; Korp-versio* [korpus]. Kielipankki. Saatavilla http://urn.fi/urn:nbn:fi:lb-2023042426

PaBiVUS = Rueter, J., & Axelson, E. (2022). Parallel Bible Verses for Uralic Studies, Korp. In Dvina-Karelian, Finnish, Komi-Permyak Online: urn:nbn:fi:lb-2020021121

Rueter, Jack (2023). On searchable Mordvin corpora at the Language Bank of Finland. In Hämäläinen, Mika; Öhman, Emily; Alnajjar, Khalid (eds.) *Lightning Proceedings of HLP4DH and IWCLUL 2023.* Lightning Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities & 8th International Workshop on Computational Linguistics for Uralic Languages NLP4DH & IWCLUL 2023. – Helsinki. pp 36–42. Online: https://zenodo.org/records/10214495

Rueter, J. (2020a). Корпус национальных мордовских языков: принципы разработки и перспективы функционирования/ действия. In *ФИННО-УГОРСКИЕ НАРОДЫ В КОНТЕКСТЕ ФОРМИРОВАНИЯ ОБЩЕРОССИЙСКОЙ ГРАЖДАНСКОЙ ИДЕНТИЧНОСТИ И МЕНЯЮЩЕЙСЯ ОКРУЖАЮЩЕЙ СРЕДЫ* (pp. 118-127). Издательский центр Историко-социологического института. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjHnNn2nJbuAhXMl4sKHQfBDqIQFjADegQIARAC&url=https%3A%2F%2Fwww.mrsu.ru%2Fru%2Fgetfile.php%3FID%3D115710&usg=AOvVaw2XqWSU-zFaq-sPZkpzI9-t

Rueter, J. M. (2020b). Linguistic Distance between Erzya and Moksha. Dependent Morphology. In Е. Ф. Клементьева, Т. И. Мочалова, & И. Н. Рябов (Eds.), *ФИННО-УГОРСКИЕ ЯЗЫКИ В СОВРЕМЕННОМ МИРЕ: ФУНКЦИОНИРОВАНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ: Материалы Всероссийской научно-практической конференции, посвященной 95-летию заслуженного деятеля науки РФ, доктора филологических наук, профессора Цыганкина Дмитрия Васильевича* (pp. 90-110). МГУ им. Н. П. Огарёва.

Rueter, J. (2014). The Livonian-Estonian-Latvian Dictionary as a threshold to the era of language technological applications. *Journal of Estonian and Finno-Ugric Linguistics*, *5*(1), 251–259. https://doi.org/10.12697/jeful.2014.5.1.14
Rueter, J. (2013). The Erzya Language, Where is it spoken? *Études finno-ougriennes*, *45*. https://doi.org/10.4000/efo.1829
Rueter, J., & Hämäläinen, M. (2020). Prerequisites For Shallow-Transfer Machine Translation Of Mordvin Languages: Language Documentation With A Purpose. In *Материалы Международного образовательного салона* (pp. 18-29). Ижевск: Институт компьютерных исследований.
Rueter, J., Hämäläinen, M., & Partanen, N. (2020). Open-Source Morphology for Endangered Mordvinic Languages. In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS) (pp. 94–100). The Association for Computational Linguistics. Online: https://aclanthology.org/2020.nlposs-1.13/
Rueter, J., & Partanen, N. (2019). *On New Text Corpora For Minority Languages On The Helsinki korp.csc.fi Server*. 32–36. Paper presented at Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы, Ufa, Russian Federation.
Rueter, J. M., & Tyers, F. M. (2018). *Towards an open-source universal-dependency treebank for Erzya*. Julkaisun esittämispaikka: International Workshop for Computational Linguistics of Uralic Languages, Helsinki, Suomi.
Sarv, Heno (2002). *Indigenous Europeans East of Moscow.* Population and Migration Patterns of the Largest Finno-Ugrian Peoples in Russia from the 18th to the 19th Centuries. Dissertation Geographicae Universitatis Tartuensis, 17. Tartu.
Suihkonen, P. M. (2003). Metadata descriptions for combining information on multimodal data located at the University of Helsinki Language Corpus Server. teoksessa S. Darányi (Toimittaja), *Proceedings of the "Higher Order Morphologies' Observer 2003" Conference on Information Society: Cultural Heritage and Folklore Text Analysis* Budapest University of Technology and Economics, Budapest, Hungary. http://bada.hb.se/handle/2320/2715
Suihkonen, P. M. (2000). On Meta-Descriptions for Cross-Linguistic Electronic Linguistic Data. teoksessa *LREC 2000* The Institute for Language and Speech Processing (ILSP) and The National Technical University of Athens, Greece. http://www.mpi.nl/IMDI/documents/documents.html#LREC_2000
Tetûrev = Rueter, J., Erina, O., & Axelson, E. (2023). Tetûrev: Environment and Natural Science for 3rd grade, part one, Korp. Parallel corpora in Erzya and Moksha. Online: urn:nbn:fi:lb-2023042421
Uspenskij = Rueter, J., Kellner, A., Janurik, B., Partanen, N., Horváth, C., & Bradly, J. (2023). Parallel Corpus of the book "Four Battles", written by L. Uspenskij; Korp version in Erzya, Hill Mari, Komi-Permyak, Komi-Zyrian, Mansi, Meadow and Eastern Mari, Moksha, Russian, Udmurt. Online: urn:nbn:fi:lb-2023042426
Witsen, N. (1705). *Noord en Oost Tartarye, Ofte Bondig Ontwerp Van eenig dier Landen en Volken Welke voormaels bekent zijn geweest. Beneffens verscheide tot noch toe onbekende, en meest nooit voorheen beschreve Tartersche en Nabuurige Gewesten, Landstreeken, Steden, Rivieren, en Plaetzen, in de Noorder en Oosterlykste Gedeelten Van Asia En Europa Verdeelt in twee Stukken, Met der zelviger Land-kaerten: mitsgaders, onderscheide Afbeeldingen van Steden, Drachten, enz. Zedert naeuwkeurig onderzoek van veele Jaren, door eigen ondervondinge ontworpen, beschreven, geteekent, en in 't licht gegeven, Door Nicolaes Witsen.* (First print: Amsterdam, 1692; Second print: Amsterdam, 1705. Reprint in 1785.) 't Amsterdam By François Halma, Boekverkooper op de Nieuwendyk.
Zeman, Daniel; Nivre, Joakim ; Abrams, Mitchell ; Ackermann, Elia ; Aepli, Noëmi ; Aghaei, Hamid ; Agić, Željko ; Ahmadi, Amir ; Ahrenberg, Lars ; Ajede, Chika Kennedy ; Akkurt, Salih Furkan ; Aleksandravičiūtė, Gabrielė ; Alfina, Ika ; Algom, Avner ; Alnajjar, Khalid ; Alzetta, Chiara ; Andersen, Erik ; Antonsen, Lene ; Aoyama, Tatsuya ; Aplonova, Katya ; Aquino, Angelina ; Aragon, Carolina ; Aranes, Glyd ; Aranzabe, Maria Jesus ; Arıcan, Bilge Nas ; Arnardóttir, Þórunn ; Arutie, Gashaw ; Arwidarasti, Jessica Naraiswari ; Asahara, Masayuki ; Ásgeirsdóttir, Katla ; Aslan, Deniz Baran ; Asmazoğlu, Cengiz ; Ateyah, Luma ; Atmaca, Furkan ; Attia, Mohammed ; Atutxa, Aitziber ; Augustinus, Liesbeth ; Avelãs, Mariana ; Badmaeva, Elena ; Balasubramani, Keerthana ; Ballesteros, Miguel ; Banerjee, Esha ; Bank, Sebastian ; Barbu Mititelu, Verginica ; Barkarson, Starkaður ; Basile, Rodolfo ; Basmov, Victoria ; Batchelor, Colin ; Bauer, John ; Bedir, Seyyit Talha ; Behzad, Shabnam ; Belieni, Juan ; Bengoetxea, Kepa ; Benli, İbrahim ; Ben Moshe, Yifat ; Berk, Gözde ; Bhat, Riyaz Ahmad ; Biagetti, Erica ; Bick, Eckhard ; Bielinskienė, Agnė ; Bjarnadóttir, Kristín ; Blokland, Rogier ; Bobicev, Victoria ; Boizou, Loïc ; Borges Völker, Emanuel ; Börstell, Carl ; Bosco, Cristina ; Bouma, Gosse ; Bowman, Sam ; Boyd, Adriane ; Braggaar, Anouck ; Branco, António ; Brokaitė, Kristina ; Burchardt, Aljoscha ; Campos, Marisa ; Candito, Marie ; Caron, Bernard ; Caron, Gauthier ; Carvalheiro, Catarina ; Carvalho, Rita ; Cassidy, Lauren ; Castro, Maria Clara ; Castro, Sérgio ; Cavalcanti, Tatiana ; Cebiroğlu Eryiğit, Gülşen ; Cecchini, Flavio Massimiliano ; Celano, Giuseppe G. A. ; Čéplö, Slavomír ; Cesur, Neslihan ; Cetin, Savas ; Çetinoğlu, Özlem ; Chalub, Fabricio ; Chamila, Liyanage ; Chauhan, Shweta ; Chi, Ethan ; Chika, Taishi ; Cho, Yongseok ; Choi, Jinho ; Chun, Jayeol ; Chung, Juyeon ; Cignarella, Alessandra T. ; Cinková, Silvie ; Collomb, Aurélie ; Çöltekin, Çağrı ; Connor, Miriam ; Corbetta, Claudia ; Corbetta, Daniela ; Costa, Francisco ; Courtin, Marine ; Crabbé, Benoît ; Cristescu, Mihaela ; Cvetkoski, Vladimir ; Dale, Ingerid Løyning ; Daniel, Philemon ; Davidson, Elizabeth ; de Alencar, Leonel Figueiredo ; Dehouck, Mathieu ; de Laurentiis, Martina ; de Marneffe, Marie-Catherine ; de Paiva, Valeria ; Derin, Mehmet Oguz ; de Souza, Elvis ; Diaz de Ilarraza, Arantza ; Dickerson, Carly ; Dinakaramani, Arawinda ; Di Nuovo, Elisa ; Dione, Bamba ; Dirix, Peter ; Dobrovoljc, Kaja ; Doyle, Adrian ; Dozat, Timothy ; Droganova, Kira ; Duran, Magali Sanches ; Dwivedi, Puneet ; Ebert, Christian ; Eckhoff, Hanne ; Eguchi, Masaki ; Eiche, Sandra ; Eli, Marhaba ; Elkahky, Ali ; Ephrem, Binyam ; Erina, Olga ; Erjavec, Tomaž ; Essaidi, Farah ; Etienne, Aline ; Evelyn, Wograine ; Facundes, Sidney ; Farkas, Richárd ; Favero, Federica ; Ferdaousi, Jannatul ; Fernanda, Marília ; Fernandez Alcalde, Hector ; Fethi, Amal ; Foster, Jennifer ; Fransen, Theodorus ; Freitas, Cláudia ; Fujita, Kazunori ; Gajdošová, Katarína ; Galbraith, Daniel ; Gamba, Federica ; Garcia, Marcos ; Gärdenfors, Moa ; Gerardi, Fabrício Ferraz ; Gerdes, Kim ; Gessler, Luke ; Ginter, Filip ; Godoy, Gustavo ; Goenaga, Iakes ; Gojenola, Koldo ; Gökırmak, Memduh ; Goldberg, Yoav ; Gómez Guinovart, Xavier ; González Saavedra, Berta ; Griciūtė, Bernadeta ; Grioni, Matias ; Grobol, Loïc ; Grūzītis, Normunds ; Guillaume, Bruno ; Guiller, Kirian ; Guillot-Barbance, Céline ; Güngör, Tunga ; Habash, Nizar ; Hafsteinsson, Hinrik ; Hajič, Jan ; Hajič jr., Jan ; Hämäläinen, Mika ; Hà Mỹ, Linh ; Han, Na-Rae ; Hanifmuti, Muhammad Yudistira ; Harada, Takahiro ; Hardwick, Sam ; Harris, Kim ; Haug, Dag ; Heinecke, Johannes ; Hellwig,

Oliver ; Hennig, Felix ; Hladká, Barbora ; Hlaváčová, Jaroslava ; Hociung, Florinel ; Hohle, Petter ; Huang, Yidi ; Huerta Mendez, Marivel ; Hwang, Jena ; Ikeda, Takumi ; Ingason, Anton Karl ; Ion, Radu ; Irimia, Elena ; Ishola, Olájídé ; Islamaj, Artan ; Ito, Kaoru ; Jagodzińska, Sandra ; Jannat, Siratun ; Jelínek, Tomáš ; Jha, Apoorva ; Jiang, Katharine ; Johannsen, Anders ; Jónsdóttir, Hildur ; Jørgensen, Fredrik ; Juutinen, Markus ; Kaşıkara, Hüner ; Kabaeva, Nadezhda ; Kahane, Sylvain ; Kanayama, Hiroshi ; Kanerva, Jenna ; Kara, Neslihan ; Karahóǧa, Ritván ; Kåsen, Andre ; Kayadelen, Tolga ; Kengatharaiyer, Sarveswaran ; Kettnerová, Václava ; Kharatyan, Lilit ; Kirchner, Jesse ; Klementieva, Elena ; Klyachko, Elena ; Kocharov, Petr ; Köhn, Arne ; Köksal, Abdullatif ; Kopacewicz, Kamil ; Korkiakangas, Timo ; Köse, Mehmet ; Koshevoy, Alexey ; Kotsyba, Natalia ; Kovalevskaitė, Jolanta ; Krek, Simon ; Krishnamurthy, Parameswari ; Kübler, Sandra ; Kuqi, Adrian ; Kuyrukçu, Oğuzhan ; Kuzgun, Aslı ; Kwak, Sookyoung ; Kyle, Kris ; Laan, Käbi ; Laippala, Veronika ; Lambertino, Lorenzo ; Lando, Tatiana ; Larasati, Septina Dian ; Lavrentiev, Alexei ; Lee, John ; Lê Hồng, Phương ; Lenci, Alessandro ; Lertpradit, Saran ; Leung, Herman ; Levina, Maria ; Levine, Lauren ; Li, Cheuk Ying ; Li, Josie ; Li, Keying ; Li, Yixuan ; Li, Yuan ; Lim, KyungTae ; Lima Padovani, Bruna ; Lin, Yi-Ju Jessica ; Lindén, Krister ; Liu, Yang Janet ; Ljubešić, Nikola ; Lobzhanidze, Irina ; Loginova, Olga ; Lopes, Lucelene ; Lusito, Stefano ; Luthfi, Andry ; Luukko, Mikko ; Lyashevskaya, Olga ; Lynn, Teresa ; Macketanz, Vivien ; Mahamdi, Menel ; Maillard, Jean ; Makarchuk, Ilya ; Makazhanov, Aibek ; Mandl, Michael ; Manning, Christopher ; Manurung, Ruli ; Marşan, Büşra ; Mărănduc, Cătălina ; Mareček, David ; Marheinecke, Katrin ; Markantonatou, Stella ; Martínez Alonso, Héctor ; Martín Rodríguez, Lorena ; Martins, André ; Martins, Cláudia ; Mašek, Jan ; Matsuda, Hiroshi ; Matsumoto, Yuji ; Mazzei, Alessandro ; McDonald, Ryan ; McGuinness, Sarah ; Mendonça, Gustavo ; Merzhevich, Tatiana ; Miekka, Niko ; Miller, Aaron ; Mischenkova, Karina ; Missilä, Anna ; Mititelu, Cătălin ; Mitrofan, Maria ; Miyao, Yusuke ; Mojiri Foroushani, AmirHossein ; Molnár, Judit ; Moloodi, Amirsaeid ; Montemagni, Simonetta ; More, Amir ; Moreno Romero, Laura ; Moretti, Giovanni ; Mori, Shinsuke ; Morioka, Tomohiko ; Moro, Shigeki ; Mortensen, Bjartur ; Moskalevskyi, Bohdan ; Muischnek, Kadri ; Munro, Robert ; Murawaki, Yugo ; Müürisep, Kaili ; Nainwani, Pinkey ; Nakhlé, Mariam ; Navarro Horñiacek, Juan Ignacio ; Nedoluzhko, Anna ; Nešpore-Bērzkalne, Gunta ; Nevaci, Manuela ; Nguyễn Thị, Lương ; Nguyễn Thị Minh, Huyền ; Nikaido, Yoshihiro ; Nikolaev, Vitaly ; Nitisaroj, Rattima ; Nourian, Alireza ; Nunes, Maria das Graças Volpe ; Nurmi, Hanna ; Ojala, Stina ; Ojha, Atul Kr. ; Óladóttir, Hulda ; Olúòkun, Adédayò ; Omura, Mai ; Onwuegbuzia, Emeka ; Ordan, Noam ; Osenova, Petya ; Östling, Robert ; Øvrelid, Lilja ; Özateş, Şaziye Betül ; Özçelik, Merve ; Özgür, Arzucan ; Öztürk Başaran, Balkız ; Paccosi, Teresa ; Palmero Aprosio, Alessio ; Panova, Anastasia ; Pardo, Thiago Alexandre Salgueiro ; Park, Hyunji Hayley ; Partanen, Niko ; Pascual, Elena ; Passarotti, Marco ; Patejuk, Agnieszka ; Paulino-Passos, Guilherme ; Pedonese, Giulia ; Peljak-Łapińska, Angelika ; Peng, Siyao ; Peng, Siyao Logan ; Pereira, Rita ; Pereira, Sílvia ; Perez, Cenel-Augusto ; Perkova, Natalia ; Perrier, Guy ; Petrov, Slav ; Petrova, Daria ; Peverelli, Andrea ; Phelan, Jason ; Pierre-Louis, Claudel ; Piitulainen, Jussi ; Pinter, Yuval ; Pinto, Clara ; Pintucci, Rodrigo ; Pirinen, Tommi A ; Pitler, Emily ; Plamada, Magdalena ; Plank, Barbara ; Poibeau, Thierry ; Ponomareva, Larisa ; Popel, Martin ; Pretkalniņa, Lauma ; Prévost, Sophie ; Prokopidis, Prokopis ; Przepiórkowski, Adam ; Pugh, Robert ; Puolakainen, Tiina ; Pyysalo, Sampo ; Qi, Peng ; Querido, Andreia ; Rääbis, Andriela ; Rademaker, Alexandre ; Rahoman, Mizanur ; Rama, Taraka ; Ramasamy, Loganathan ; Ramisch, Carlos ; Ramos, Joana ; Rashel, Fam ; Rasooli, Mohammad Sadegh ; Ravishankar, Vinit ; Real, Livy ; Rebeja, Petru ; Reddy, Siva ; Regnault, Mathilde ; Rehm, Georg ; Riabi, Arij ; Riabov, Ivan ; Rießler, Michael ; Rimkutė, Erika ; Rinaldi, Larissa ; Rituma, Laura ; Rizqiyah, Putri ; Rocha, Luisa ; Rögnvaldsson, Eiríkur ; Roksandic, Ivan ; Romanenko, Mykhailo ; Rosa, Rudolf ; Roșca, Valentin ; Rovati, Davide ; Rozonoyer, Ben ; Rudina, Olga ; Rueter, Jack ; Rúnarsson, Kristján ; Sadde, Shoval ; Safari, Pegah ; Sahala, Aleksi ; Saleh, Shadi ; Salomoni, Alessio ; Samardžić, Tanja ; Samson, Stephanie ; Sanguinetti, Manuela ; Sanıyar, Ezgi ; Särg, Dage ; Sartor, Marta ; Sasaki, Mitsuya ; Saulīte, Baiba ; Savary, Agata ; Sawanakunanon, Yanin ; Saxena, Shefali ; Scannell, Kevin ; Scarlata, Salvatore ; Schang, Emmanuel ; Schneider, Nathan ; Schuster, Sebastian ; Schwartz, Lane ; Seddah, Djamé ; Seeker, Wolfgang ; Seraji, Mojgan ; Shahzadi, Syeda ; Shen, Mo ; Shimada, Atsuko ; Shirasu, Hiroyuki ; Shishkina, Yana ; Shohibussirri, Muh ; Shvedova, Maria ; Siewert, Janine ; Sigurðsson, Einar Freyr ; Silva, João ; Silveira, Aline ; Silveira, Natalia ; Silveira, Sara ; Simi, Maria ; Simionescu, Radu ; Simkó, Katalin ; Šimková, Mária ; Símonarson, Haukur Barri ; Simov, Kiril ; Sitchinava, Dmitri ; Sither, Ted ; Skachedubova, Maria ; Smith, Aaron ; Soares-Bastos, Isabela ; Solberg, Per Erik ; Sonnenhauser, Barbara ; Sourov, Shafi ; Sprugnoli, Rachele ; Stamou, Vivian ; Steingrímsson, Steinþór ; Stella, Antonio ; Stephen, Abishek ; Straka, Milan ; Strickland, Emmett ; Strnadová, Jana ; Suhr, Alane ; Sulestio, Yogi Lesmana ; Sulubacak, Umut ; Suzuki, Shingo ; Swanson, Daniel ; Szántó, Zsolt ; Taguchi, Chihiro ; Taji, Dima ; Tamburini, Fabio ; Tan, Mary Ann C. ; Tanaka, Takaaki ; Tanaya, Dipta ; Tavoni, Mirko ; Tella, Samson ; Tellier, Isabelle ; Testori, Marinella ; Thomas, Guillaume ; Tonelli, Sara ; Torga, Liisi ; Toska, Marsida ; Trosterud, Trond ; Trukhina, Anna ; Tsarfaty, Reut ; Türk, Utku ; Tyers, Francis ; Þórðarson, Sveinbjörn ; Þorsteinsson, Vilhjálmur ; Uematsu, Sumire ; Untilov, Roman ; Urešová, Zdeňka ; Uria, Larraitz ; Uszkoreit, Hans ; Utka, Andrius ; Vagnoni, Elena ; Vajjala, Sowmya ; Vak, Socrates ; van der Goot, Rob ; Vanhove, Martine ; van Niekerk, Daniel ; van Noord, Gertjan ; Varga, Viktor ; Vedenina, Uliana ; Venturi, Giulia ; Villemonte de la Clergerie, Eric ; Vincze, Veronika ; Vlasova, Natalia ; Wakasa, Aya ; Wallenberg, Joel C. ; Wallin, Lars ; Walsh, Abigail ; Washington, Jonathan North ; Wendt, Maximilan ; Widmer, Paul ; Wigderson, Shira ; Wijono, Sri Hartati ; Wille, Vanessa Berwanger ; Williams, Seyi ; Wirén, Mats ; Wittern, Christian ; Woldemariam, Tsegay ; Wong, Tak-sum ; Wróblewska, Alina ; Wu, Qishen ; Yako, Mary ; Yamashita, Kayo ; Yamazaki, Naoki ; Yan, Chunxiao ; Yasuoka, Koichi ; Yavrumyan, Marat M. ; Yenice, Arife Betül ; Yıldız, Olcay Taner ; Yu, Zhuoran ; Yuliawati, Arlisa ; Žabokrtský, Zdeněk ; Zahra, Shorouq ; Zeldes, Amir ; Zhou, He ; Zhu, Hanzhi ; Zhu, Yilun ; Zhuravleva, Anna ; Ziane, Rayan, 2023-11, Universal Dependencies 2.13, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5287.