Towards an Old Permic Universal Dependencies Treebank

Niko Partanen[1*], Jack Rueter[1], Rogier Blokland[2]
1 University of Helsinki, Finland
2 Uppsala University, Sweden

*Corresponding author: Niko Partanen niko.partanen@helsinki.fi

Abstract
Old Permic, also known as Old Komi, is an extinct variety of Komi that was spoken in the late Middle Ages in the lower Vychegda river basin in Northeastern European Russia, in an area that currently is not Komi-speaking. This language variety is attested in fragmentary records from the 14th to 17th century written both in the Old Permic alphabet and in Cyrillic. These records are of significant importance for research on the history of the Komi language. Here we introduce our attempt towards a new Universal Dependencies treebank that will eventually contain the existing corpus of Old Permic in a structured and CoNLL-U annotated format. This will be the first time this material is being made openly available in digital format, and our contribution describes the current state of the art and remaining challenges.

keywords
Uralic; Permic languages; Old Permic; Old Komi; CoNLL-U; Universal Dependencies

INTRODUCTION

Komi is a pluricentric Uralic language spoken in North-eastern European Russia. There are two main varieties with separate written standards, Permian Komi and Zyrian Komi[1], and a severely endangered less-documented variety called Yazva Komi. The largest variety, Zyrian Komi, has approximately 180,000 speakers. This paper focuses on the Zyrian Komi variety, which has existing written records from the 14th century onwards. The variety used in these early written records is usually known as Old Permic, and it consists of about 830 words (about 240 in the Old Permic script and nearly 600 in Cyrillic) in running text of religious nature; they are all translations of Church Slavonic texts[2] (Baker 1983; Baker 1985: 24-29; Rédei 1993), and primarily found written in codices and on old icons[3]. We would prefer the term 'Old Komi'[4], as this variety can clearly be considered a Zyrian Komi dialect, although one that differs to some degree from contemporary dialects (Lytkin 1952: 119-121; Ljašev 1980). Needless to say, in order to maintain continuity with earlier research, we will here follow the earlier convention. The region where Old Permic is attested is no longer Komi-speaking, although, historically, it was known to have been so.

[1] The terms 'Komi-Permyak' and 'Komi-Zyrian' are also used; speakers of either variant will, however, usually just say they speak 'Komi'. Hence, our terminology provides a clearer indication of the pluricentricity of the language.
[2] Lytkin (1889: 30–37) shows for a number of Old Permic texts from which Church Slavonic texts they were translated.
[3] In the 2000s, a few publications have appeared using the Old Komi script, e.g. [illegible], Antologiâ finno-ugorskoj poèzii Terent′ev, Semën (2011a); Terent′ev has also (re)published Russian translations of Omar Khayyam's poetry using his own adaptation of the Old Komi script (Terent′ev 2011b) .
[4] For a differing view, see Lytvynenko & Griščenko 2022: 146.

Besides Old Hungarian (first texts from the 12th century) and individual Finnic records (first texts from the 13th century), the texts in Old Permic are one of the oldest written sources we have of any Uralic language. They are obviously of utmost importance for historical research on the Permic languages, and they are in many ways a unique early linguistic record within this language family. There are texts in Old Permic written both in the Old Permic script (14th-16th century)[5] and in Cyrillic script (16th-17th century), and they contain phonological features that are not present in contemporary dialects, as well as archaic lexicon.[6]

Recent years have seen the emergence of many treebanks of historical language varieties. These include, for example, the Old East Slavic Birchbark treebank[7] and the Old Turkic treebank (Derin & Harada 2021). The Old Permic treebank will fit very well in this context, increasing the diversity of written records produced in Northern Eurasia in the Middle Ages.

Figure 1 below shows a 14th century Troitsa icon, on which the text in Old Permic script can be found. Other texts in Old Permic script have not, to our knowledge, been digitized in such high resolution, and the only images available are hand-drawn reproductions. There are no existing original texts in the Old Permic script. As we currently do not have high quality scans or images of all individual texts, work towards a more comprehensive Old Permic corpus is still at an incomplete state. We are presently working with current reproductions with the goal of assigning each pixel region that corresponds to a character with the correct Unicode symbol, whereas Old Permic is one of the many Komi alphabets supported (cf. Rueter & Ponomareva 2019). This would make it clear which characters can be read and how much interpretation is needed, the level of which could also be annotated separately per character. The condition of individual words and characters necessarily leaves some space for guesswork and varying interpretations.

We aim to collect all available Old Permic materials, both those in Old Permic script and in Cyrillic, and publish them as an annotated Universal Dependencies treebank. This corpus is described in the current study. For the wider context, however, we also aim to contextualize the background of Old Permic in slightly more detail in this paper so that the material can be maximally useful and well understood.

Figure 1. The 14th century Troitsa icon with Komi text (Стефан Пермский, Public domain, via Wikimedia Commons)

The texts are all of a religious nature, and have existing parallels among the Russian recensions of Old Church Slavonic religious material. This is also important for the interpretation of the texts. The texts themselves are not always easy to read, and many characters are close to one

[5] It was also occasionally used by Russian scribes in the 16th century to write Russian for cryptographic purposes (Lytkin 1952: 75-87; Baker 1983: 87-88).
[6] For a brief English-language overview of Old Permic see Baker 1985: 24-29.
[7] https://github.com/UniversalDependencies/UD_Old_East_Slavic-Birchbark

another and show clear variation between different spellings of the individual words, but comparison with other variants is usually helpful for arriving at a specific reading.[8]

We believe that corpora of historical languages are also an important data source for digital humanities research. In recent years the availability of such materials has increased significantly, and they have been used in a wide range of disciplines. They are also good candidates for different experiments in complex collaborative research undertakings, as the texts are no longer under copyright, and have also often already been studied by several generations of scholars. However, how to represent these materials in a digital format in a way that is accurate with regard to the original sources, whilst showing respect to the work done before, is not a question that we at present necessarily know how to answer.

I WRITING SYSTEM

The Old Permic script is read from left to right, and it consists of 38 characters. Five of the letters are combining. The texts, as is also common for Russian writing from the period, are all in scriptio continua, i.e. word or sentence boundaries are not marked; only some wider spaces occur occasionally to denote a longer pause. The script was used between the 14th and 16th centuries, and its creation is usually attributed to Stephen of Perm (c. 1340–1396). There are two texts in Old Permic that can be directly dated; one to 1486 (cf. Grinščenko & Ponarjadov 2021: 12) and one to 1510 (cf. Lytkin 1952: 32). The exact time period when the script was used still remains a topic of further investigation, but can with certainty be said to have been from the end of the 14th century till the middle of the 16th century. This dating is supported by the linguistic features of Old Permic: as the divergence from modern Komi dialects is relatively minor, the script cannot be much older.

The relationship of Old Permic and Komi is also a facet for further discussion of which language code should be used to represent Old Permic. In our current preliminary version the code ‘kv’ is used, referring to Komi in general, whereas the other treebanks use the language codes ‘kpv’ and ‘koi’, referring to Zyrian and Permian Komi, respectively. As the relationship to Zyrian Komi seems still dialectal, the code ‘kpv’ could be the most correct, but the question must be asked whether there could be a danger of confusion in such distinct materials if they are not distinguished in any manner at the language code level from other Komi materials. There is no language code for Old Permic, but nevertheless, before the official UD release this problem must be solved in some manner.

The Old Permic Unicode block was proposed to be included in Unicode in 2011, which was accepted in 2014, and currently there are several fonts that support these characters. An Android mobile keyboard has also been published[9]. The script is not actively used in modern communication, although small clusters of enthusiasts do exist at least in Russia and Finland.

The script has been compared to Cyrillic and Greek characters (Penttilä 1924; Stipa 1960), as well as, less convincingly, to e.g. the Old Turkic script (Ponarjadov 1996) and the Phoenician alphabet (Turkin 1996). The possible influence of Komi traditional written signs, known as pas ‘(property) sign’, has also been extensively discussed (cf. Korolev & Savel’eva 1996), but to the authors’ knowledge, as of yet, no thorough comparative study has been conducted of how

[8] Penttilä (1924: 37) points out that copies of the same text often differ significantly in terms of graphemic legibility.
[9] https://play.google.com/store/apps/details?id=com.majbyr.keyboard

the different scripts and signs exactly relate to each other at the character level, and which influence is visible where.

The Old Permic script was used to write the first texts in Old Komi, but later, after knowledge of the script declined, Old Komi texts were transliterated into Cyrillic; there are two extant texts in Cyrillic totalling almost 600 words, and there is therefore much more material in Old Komi in Cyrillic than in the Old Permic script. The texts in Cyrillic will also be included as they clearly represent the same language form.[10] Inclusion of texts in the Old Permic corpus will essentially be done based on language, not the script, although the Old Permic script is very distinctive for this language variety. Other Komi treebanks all use contemporary orthographies, with UD_Komi_Zyrian-IKDP also containing dialectal materials written in Cyrillic transcription that still closely follows the current orthography.

To our knowledge, no one has yet published all the Old Permic texts in the original script using the currently available Unicode encoding. Our aim is to complete this in a reasonable time frame. At the same time, due to the difficulties in reading and interpreting the characters, it is clear this will not be the final word on the matter. In the current version, complete renditions of all the texts in the Old Permic script have not been included as our work is still in progress. Be as it may, the currently completed individual portions give a good illustration of what can be done, and they also provide a location for further discussion of the chosen conventions.

II CORPUS

Texts are included in Old Permic script or Cyrillic alphabet, depending on the text; for texts in Old Permic also Lytkin's 1952 Cyrillic transliteration will be included. All texts found before the 1950s have been exhaustively described in Lytkin 1952. Since then, there have been two major discoveries: a text with 18 words published in Sidorov in 1962, and a number of texts with a total of 24 words published in Grinščenko & Ponarjadov in 2021.

The word form corresponds to the string that is present in the original sources. The word tokenization follows Lytkin's conventions. We agree that it could also be possible to represent each text without word boundaries, following the scriptio continua of the original texts, but as we operate with a treebank structure, divisions into words and sentences are essential, and in any case we are working with an interpretation of the text as we are publishing it in a highly complex annotated structure. The original line boundaries are marked as ‹/› in the word form and lemma columns, and in the MISC-column the feature value is represented by ‹NewLine=Yes›. This follows the conventions of other treebanks of historical texts, e.g., as in Old East Slavic Birchbark.

The goal of these conventions is that the corpus user should be able to retrieve a digital representation of the entire original text from the material in a condition where they can read in the original sources, and represented in the characters that can be considered as correct or plausible interpretations. Combined with information about missing or destroyed characters, this also gives a good starting point for situations where the text becomes more legible. Lately, there have been great advances in image restoration using, e.g., image processing algorithms (Knox et al. 2008) or x-ray imaging and AI (Parsons et al. 2023), so one wonders if parts of the Old Permic texts which are illegible now could eventually be retrieved with modern technology. Even in parts where the text on icons has been damaged, it is worth pondering if the pigments

[10] The cryptographic texts in Russian in the Old Permic script are obviously not included.

used have left some detectable traces in the wood layer beneath. In a wider context we believe our annotation format serves as a good example in republishing electronically annotated versions of ancient texts without loss of information. The problem remains that we have but few texts in high resolution (the abovementioned Troitsa [Figure 1] and the texts in Grinščenko & Ponarjadov 2021), and no easy way to present or obtain high resolution images of the other ones.

In lemmatization, the forms are given in the contemporary Zyrian Komi orthography. When the corresponding word does not exist in the standard language, but does occur in dialects, the form follows the correspondences deductible as presented in the recent Zyrian Komi dialect dictionary (Beznosikova et al. 2012). The dialect underlying the Old Permic corpus seems to be close to the contemporary Udora and Lower Vychegda dialects to which it has been geographically closest as well. As the current Komi dialect treebanks also contain lemmas in the standard Zyrian Komi (Partanen et al. 2018), this choice seems to be well founded for our purposes here.

Part-of-speech tagging follows the tagset in Universal Dependencies project, as used in the Permian Komi and Zyrian Komi treebanks (Partanen et al. 2018), included in the UD release 2.5. (Zeman et al. 2019). In addition, the morphological analysis follows the UD conventions in the Komi and other Uralic language treebanks. The recent efforts to systematize the annotation conventions within Uralic languages in general (Partanen & Rueter 2019), and in more detailed issues such as numerals (Rueter et al. 2021), will be contemplated as thoroughly as possible. Old Permic itself does not seem to contain morphosyntactic features entirely absent from modern Komi dialects, that is to say, our current level of annotation at least has not encountered any. One morphological difference with regard to current Komi treebanks would be the seeming lack of certain tense oppositions that would correspond to the contemporary present and future tenses. The same pattern found in Old Permic is also present in the modern Zyrian Komi dialect of Udora (Partanen & Kellner 2021: 178). The current annotation conventions, however, are Tense=Pres and Tense=Fut, although this may not be the optimal solution.

One difference between the Old Permic treebank and the other Zyrian Komi treebanks is that the morphological analysis produced by an FST is not included in the MISC column. The main reason is that there is no analyser of this type for Old Permic, and as the number of texts is very small and do not represent a currently spoken dialect, it does not seem necessary to integrate it either in the GiellaLT (Pirinen et al. 2023; Moshagen et al. 2023) infrastructure that is mainly targeted toward language maintenance efforts. Similarly, tools that have been very important in treebank creation for the Uralic languages, such as UralicNLP (Hämäläinen 2019), cannot be used due to these constraints.

The annotation of dependencies is currently in progress, and we have not yet encountered any annotation problems unknown to other existing treebanks of Komi. The texts are, in the end, relatively straightforward Komi, although it must be emphasised that if some words are to be read differently, the result would oftentimes see a new syntactic structure.

As sentence boundaries are not indicated in the Old Permic texts, they must be deduced from the context. This allows multiple interpretations, which in turn results in different tree structures in the annotations; we have mainly followed the interpretation of Lytkin 1952, here. The treebank is characterized by a fairly large number of lists and repetitions, which seems to follow from the religious nature of the material. There are often annotations with the APPOS relation.

In some instances, the word order is not natural for Komi, and appears to follow the original Church Slavonic source. The possessive constructions annotated with NMOD regularly have a word order opposite to the one expected based on other Komi treebanks.

III CONCLUSION

We hope this work leads the way to further research on Old Permic. There is ample scope for future critical editions and new analyses of these texts, especially as our knowledge of Komi dialects and other Permic varieties increases. The language resource presented in this paper concentrates primarily on making these materials available in a new digital form, with basic linguistic analysis that is complete, though not exhaustive.

The complete treebank will be made available in the Universal Dependencies release (v2.15), which is scheduled for November 15, 2024. The development version can be accessed in the Universal Dependencies project's GitHub repository.[11]

Further work that could complement the current landscape of Komi treebanks would include different early Komi texts spanning newer periods of the written Komi tradition (from the 18th century onwards). Having more openly available materials on Komi dialects would also be important. There is a discontinuity between the Old Permic script and later orthographies used for Komi, but through the contemporary dialectal variation these registers are still interwoven, despite interruptions in the written record.

References

Baker, R. W. Slavonic influence upon the language of the Old Permian texts. *Finnisch-Ugrische Forschungen* 45: 82-106, 1983.

Baker, R. W. *The development of the Komi case system*. Suomalais-Ugrilaisen Seuran toimituksia 189. Helsinki: Suomalais-Ugrilainen Seura. 1985.

Baraksanov, G. G. *Важ коми гижöдъяс*. – Syktyvkar. 1992.

Beznosikova, L. M., Zaboeva, N. K., Ajbabina, E. A. & Kosnyreva, R. I. *Коми сёрнисикас кывчукöр I-II. Словарь диалектов коми языка I-II*. Сыктывкар: ООО Издательство «Кола». 2012.

Derin, M. O., & Harada, T. Universal Dependencies for Old Turkish. In Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021) (pp. 129-141), 2021.

Grinščenko, A. I. & Ponarjadov, V. V. Новые находки памятников древнепермского языка и письма. *Урало-алтайские исследования* 4(43) (pp. 7-34), 2021.

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

Knox, K. & Easton, R. & Christens-Barry, W. Image restoration of damaged or erased manuscripts. *16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, August 25–29, 2008.

[11] https://github.com/UniversalDependencies/UD_Komi-OldPermic/tree/dev

Korolev, K.S. & Savel'eva, Ė.A. К проблеме происхождения коми письменности. Бараксанов, Г.Г., Тираспольский, Г. И. & Напалков, А.Д. (отв. ред.). *Стефан Пермский и современность.* Сыктывкар: Российская академия наук. Уральское отделение. Коми научный центр. (pp. 24–29), 1996.

Ljašev, V.A. *Диалектное членение древнекоми языка*. Серия препринтов. "Научные доклады". Выпуск 60. Сыктывкар: Академия наук СССР, Коми филиал, 1980.

Lytkin, G.S. *Зырянский край при епископах пермских и зырянский язык*. Санктпетербург: Типография императорской академии наук, 1889.

Lytkin, V.I. *Древнепермский язык. Чтение текстов, грамматика, словарь*. Москва: Издательство Академии наук СССР, 1952.

Lytvynenko, V. V. & Griščenko, A.I. Unlocking two marginalia in Old Permic script in a fifteenth-century Slavonic manuscript (Russian State Library, Volok. 437) with Athanasius' *Orations Against the Arians*. *Byzantinoslavica, Revue internationale des études byzantines 80/1-2: 146–162, 2022.*

Nørstebø Moshagen, S., Pirinen, F., Antonsen, L., Gaup, B., Mikkelsen, I. L. S., Trosterud, T., Wiechetek, L. & Hiovain-Asikainen, K. The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP. In Hurskainen, A., Koskenniemi, K., Pirinen, F. A., Ranta, A., Listenmaa, I., Axelson, E., Hardwick, S., Lindén, K., Moshagen, S. N. and others, *Rule-Based Language Technology*, NEALT Monograph Series, 2, 2023.

Parsons, S., Parker, C. S., Chapman, C., Hayashida, M., & Seales, W. B. EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT. *arXiv preprint arXiv:2304.02084*, 2023.

Partanen, N., Blokland, R., Lim, K., Poibeau, T., & Rießler, M. The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium* (pp. 126-132), 2018.

Partanen, N., & Kellner, A. On the interplay between tense marking, aspect and temporal continuity in Udora Komi. *Finnisch-Ugrische Forschungen*, 2021.

Partanen, N., & Rueter, J. Survey of Uralic Universal Dependencies development. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)* (pp. 78–86), 2019.

Penttilä, A. Muutamia paleografisia huomioita Pyhän Tapanin syrjäniläisestä kirjaimistosta. *Turun historiallinen arkisto* 1 (pp. 32–45), 1924.

Pirinen, F. A., Moshagen, S. N., & Hiovain-Asikainen, K. GiellaLT---a stable infrastructure for Nordic minority languages and beyond. In *The 24rd Nordic Conference on Computational Linguistics*, 2023.

Ponarjadov, V.V. О возможной связи стефановской письменности с древнетюркской руникой. Савельева, Е.А. (отв.ред.). *Христианизация коми края и ее роль в развитии*

государственной культуры. Том II. Филология. Этнология. Сыктывкар: Коми научный центр УрО Российской академии наук. (pp. 238-241), 1996.

Rédei, K. *Óegyházi szláv szemantikai és szintaktikai hatás az ózürjén nyelvben*. Akadémiai székfoglaló 1992. január 6. Budapest: Akadémiai Kiadó, 1993.

Rueter, J., Partanen, N., & Pirinen, T. A. Numerals and what counts. In *Fifth Workshop on Universal Dependencies. The Association for Computational Linguistics*, 2021.

Rueter, J., & Ponomareva, L. Komi Latin Letters, Degrees of UNICODE Facilitation. In *Proceedings of the Language Technologies for All (LT4All)* (Sivut 251–254). European Languages Resources Association (ELRA), 2019.

Sidorov, A. S. Новые памятники древнекоми письменности (С комментариями, подстрочными примечаниями и заключением В. И. Лыткина). *Вопросы финно-угорского языкознания. Вып. 1*. Moskva. (pp. 178–211), 1962.

Stipa, G. Der Ursprung der permischen Schrift. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 110 (Neue Folge 35) (pp. 342–364), 1960.

Terent′ev, Semën (compl). [illegible] / *Antologiâ finno-ugorskoj poèzii*. – Syktyvkar, 2011a.

Terent′ev, Semën. [illegible]. Omar Khajam. *Rubaj*. Syktyvkar, 2011b.

Turkin, A. Некоторые размышления о происхождении древнепермской письменности. Савельева, Е.А. (отв.ред.). *Христианизация коми края и ее роль в развитии государственной культуры. Том II. Филология. Этнология*. Сыктывкар: Коми научный центр УрО Российской академии наук. (pp. 278–281), 1996.

Zeman, D.; et al. Universal Dependencies 2.5, 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/11234/1-3105, 2019.