Historiallisia ja oman aikamme suomalais-ugrilaisten vähemmistökielten korpuksia digitalisaation näkökulmasta / Historical and Modern Corpora on Finno-Ugric Minority Languages in the Context of Digitalizatio

(Erkkilä – Partanen – Virtanen)

Abstraktit

Timofey Arkhangelskiy (University of Hamburg), Nikolai Anisimov (Estonian Literary Museum) & Tatiana Anisimova (University of Tartu): ***Udmurt dialectal corpora: What, how, and for whom?***

Compiling sound-aligned dialectal corpora is labor-intensive, which is why there are few of them for the minority Uralic languages. Obviously, they are important for linguistic research, but they can be useful for other audiences as well. We will discuss our experience of a collaborative work between linguists and folklorists working at the intersection of folklore studies and cultural anthropology, focusing on Udmurt, a dialectally diverse language (Maksimov 2009). We will also consider Beserman. Specifically, we will discuss a Beserman multimedia corpus1 and a cross-dialectal Udmurt corpus. The latter requires further processing and will be merged with the existing (but tiny) sound-aligned Udmurt corpus (Arkhangelskiy & Georgieva 2018).

In the first part of our talk, we will focus on the technical side (‘how?’): transcribing in ELAN, rule-based annotation, handling of sensitive information, and publishing online in the Tsa-korpus platform.

In the second part, we will look closely at the content of the cross-dialectal corpus (‘what?’ and ‘for whom?’). It mostly consists of interviews related to the veneration of the dead and other phenomena, conducted by Nikolai Anisimov, who is a folklorist and a native speaker. These texts are of great value both linguistically and anthropologically. We will discuss how both linguists and folklorists benefit from such a collaboration and demonstrate specific tools we added to the Tsakorpus platform in order to improve the usability for folklorists and anthropologists (full-text view, support for topic tags).

Collaborative work of this kind results in a more accurate interpretation of folklore and mythological data, since a deep understanding of the cultural context, ritual practices, and the local tradition is complemented by linguistic analysis.

1 https://beserman.web-corpora.net/index_en.html

Finally, a corpus containing materials collected by folklorists or anthropologists will prob-ably be of interest to the language community—which is not always the case with corpora com-piled exclusively by linguists.

References

Arkhangelskiy, Timofey & Ekaterina Georgieva. 2018. Sound-aligned corpus of Udmurt dialectal texts. In *Proceedings of the Fourth International Workshop on Computatinal Linguistics of Uralic Languages*, 26–38. Helsinki, Finland: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0203.

ELAN (Version 7.0) [Computer software]. 2025. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Maksimov, Sergey. 2009. Kommentarij k kartam “Dialektnoe chlenenie udmurtskogo jazyka” i “Prinadlezhnost’ opornyx punktov k territorial’nym dialektam” [Comments for the maps “Dialectal subdivision of the Udmurt language” and “Distribution of pivot settlements by territo-rial dialects”]. In Rif Nasibullin, Sergey Maksimov, Vasiliy Semenov & Galina Otstavnova (eds.), *Dialektologicheskij atlas udmurtskogo jazyka: karty i kommentarii [Dialectological atlas of Ud-murt: Maps and comments], vol. 1, 42–48. Izhevsk: Reguljarnaja i xaoticheskaja dinamika.*

Timofey Arkhangelskiy (University of Hamburg) & Maria Brykina (University of Hamburg): ***Metadata handling in the INEL project***

The long-term project INEL aims at documentation of endangered or extinct languages, most of them Uralic. Due to the scarcity of data, the only viable data collection strategy is opportunistic: “include anything that is out there”. As a consequence, many of our nine corpora contain very heterogeneous data, making it difficult for linguists to correctly interpret the results of their queries.

Having extensive metadata and being able to use them in search and statistical queries is indispensable in our case. We will discuss the way metadata are structured and used in our corpora. First, we will outline our data model and format (CoMa XML) and present our web-based search platform, Tsakorpus. Then we will focus on four use cases from three corpora of Samoyedic lan-guages. In all of them, a simple corpus-wide search whould yield misleading results.

First, we will demonstrate on the examples of Kamas and Selkup that the proportion of Russian borrowings varies a lot between varieties within these languages.

Second, we’ll look at the anterior converb in Kamas. The corpus consists of two parts: early texts from full speakers and recordings of the last speakers (see Klumpp 2022). This converb only occurs in the first, smaller, part. It used to be frequent when the language was still spoken, but went extinct in the variety of the last speakers.

Third, there are seven words glossed as ‘river’ in Selkup. A metadata-based analysis re-veals that some of these terms are distributed across very different dialects (Glushkov et al. 2013).

Finally, the choice between two apparently synonymous variants of a discourse particle translated as ‘now’ in Nganasan depends on speakers' individual preferences.

Extensive documentation of our data thus allows one to compensate for the lack of balance in the corpus and avoid making wrong conclusions.

References

INEL: The INEL corpora of indigenous Northern Eurasian languages. https://hdl.handle.net/11022/0000-0007-F45A-1.

Glushkov, Sergey, Aleksandra Baidak & Natalya Maksimova. 2013. Dialekty sel'kupskogo jazyka [Selkup dialects]. In N. Tuchkova, S. Glushkov, E. Kosheleva, A. Golovnyov, A.

Baidak & N. Maksimova (eds.), *Sel'kupy. Ocherki traditsionnoj kul'tury i sel'kupskogo jazyka* [Selkups: Surveys of traditional Selkup culture and language], 49–63. Tomsk: TSUP.

Klumpp, Gerson 2022. Kamas. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik (eds.), *The Oxford Guide to the Uralic Languages*. Oxford: Oxford University Press.

Riku Erkkilä (Helsingin yliopisto): *Suomalais-Ugrilaisen Seuran julkaisemat aineistot korpuksina*

Suomalais-Ugrilainen Seura (SUS) on julkaissut aineistoja lähes kaikista uralilaisista kielistä. Merkittävä osa näistä on kerätty SUS:n rahoittamilla kenttätyömatkoilla 1880-luvulta aina ensimmäiseen maailmansotaan asti (Salminen 2008: 26–36, 99–101), mutta myös myöhemmin kerättyjä aineistoja on julkaistu (esim. Munkácsi & Fuchs 1952; ks. myös Salminen 2008: 151–155).

Tässä esitelmässä pohdin sitä, voiko näitä aineistoja käyttää korpuksina. Korpuksella tarkoitan esitelmässä nimenomaan aineistokokoelmaa, jota käytetään modernissa korpuslingvistiikassa (Gilquin & Gries 2009: 6). Tällaiselle aineistolle on esitetty erilaisia kriteerejä, esimerkiksi kattavuus (representativeness; Biber 1993; Stefanowitsch 2020: 28–36; Ädel 2020: 4–6), eli se, että korpus sisältää mahdollisimman kattavasti erilaisia tekstejä, autenttisuus (Gilquin & Gries 2009: 6; Stefanowitsch 2020: 23–28), eli se, että aineiston kieli on tuotettu luonnollisessa tilanteessa, ja koko (Biber 1993: 247–248; Stefanowitsch 2020: 37–38; Ädel 2020: 8). Lisäksi saavutettavuus, esimerkiksi korpuksen digitaalisuus ja annotaatio (Gilquin & Gries 2009: 6)

Suurten tutkimusmatkojen aikana kerätty, samoin kuin ilmeisesti myöhempikin, on koottu alun perin Heikki Paasoselle esitettyjen kriteerien perusteella (Salminen 2008: 31–33) Nämä ovat eri mordva(laiskielte)n murteiden aineisto sanakirjaa varten, aineisto, josta käy ilmi kiel(t)en kieliopillinen rakenne sekä folkloretekstit (Korhonen 1986: 146). Kerätyt aineistot edustavat useimmista kielistä varhaisinta saatavilla olevaa aineistoa. Lisäksi informanteista on yleensä olemassa suhteellisen hyvät sosiolingvistiset tiedot, kuten ikä ja synnyinpaikka. Aineistojen koko vaihtelee suuresti. Esimerkiksi Paasosen Mordwinische Volksdichtung käsittää painettuna yhteensä yli 4000 sivua (luku sisältää tekstit ja niiden käännökset; Salo 2010: 78–79), kun taas esimerkiksi Kildinlappische Sprachproben (Itkonen & Lehtiranta 1985) on vain noin sata sivua käännökset mukaan luettuna. Ainakin osa teksteistä on kerätty autenttisissa tilanteissa (ks. Salo 2010: 67).

On ilmeistä, että SUS:n aineistot ovat varsin kattavia, mutta aineiston autenttisuus ja koko vaihtelevat. Aineistojen suurin heikkous korpusten kriteereihin nähden on, että suurin osa niistä ei ole digitaalisesti saatavilla, eikä niissä ole annotaatiota. Jos SUS:n aineistoja käytetään korpuksina, nämä rajoitteet on otettava huomioon.

Lähteet

Biber, Douglas 1993: *Representativeness in corpus design. – Literary and Linguistic Computing* 8 (4) s. 243–257.

Gilquin, Gaëtanelle – Gries, Stefan Th. 2009: Corpora and experimental methods. A state-of-the-art review. – *Corpus Linguistics and Linguistic Theory* 5 (1) s. 1–26. https://doi.org/10.1515/CLLT.2009.001.

Itkonen, Erkki – Lehtiranta, Juhani (toim.) 1985: *Kildinlappische Sprachproben*. Suomalais-Ugrilaisen Seuran Toimituksia 191. Helsinki: Suomalais-Ugrilainen Seura.

Korhonen, Mikko 1986: *Finno-Ugrian language studies in Finland 1828–1918*. The history of learning and science in Finland 1828–1918 11. Helsinki: Societas Scientiarum Fennica.

Munkácsi, Bernhard – Fuchs, D. R. (toim.) 1952: *Volksbräuche und Volksdichtung der Wotjaken*. Suomalais-Ugrilaisen Seuran Toimituksia 102. Helsinki: Suomalais-Ugrilainen Seura.

Salminen, Timo 2008: *Aatteen tiede. Suomalais-Ugrilainen Seura 1883–2008*. Suomalaisen Kirjallisuuden Seuran Toimituksia 1172. Helsinki: Suomalaisen Kirjallisuuden Seura.

Salo, merja 2010: Heikki Paasonen – Mittavan aineiston kerääjä ja keräyttäjä sekä jälkipolvien työllistäjä. – Paula Kokkonen & Anna Kurvinen (toim.), *Kenttäretkistä tutkimustiedoksi* s. 57–106. Uralica Helsingiensia 4. Helsinki: Suomalais-Ugrilainen Seura.

Stefanowitsch, Anatol 2020: *Corpus linguistics. A guide to the methodology*. Textbooks in Language Sciences 7. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822.

Ädel, Annelie 2020: Corpus compilation. – Magali Paquot & Stefan Th. Gries (toim.), *A practical handbook of corpus linguistics* s. 3–24. Cham: Springer. https://doi.org/10.1007/978-3-030-46216-1_1.

Olesya Khanina (University of Helsinki) & Andrey Shluinsky (Humbold Universität zu Berlin): ***Tracing a long path to the Enets corpus: between fieldwork and legacy data***

We will provide an overview of how the glossed Enets corpus (Shluinsky et al. 2024) was created. The corpus is a digital dataset, available online. Arkhipov & Shluinsky (2025) is a user's guide to the corpus, which documents its structure and scope, as well as gives credit to all people who contributed to its creation for over more than a half of a century: speakers, archive owners, linguists, and technical assistants. However, it does not cover many other practical aspects that have been raised by the workshop organisers as worthy of a scholarly discussion: e.g. digitization as a process, obtaining financial support, time scales, challenges of locating legacy materials scattered over thousand of kilometers, negotiation with archive owners, ethical and copyright-related issues, etc. A separate topic is multi-faceted consolidation and integration of different elements of the legacy data and our own extensive modern recordings, which led to a single corpus with unified conventions.

We will present all stages of the workflow and all participants that contributed to the corpus, and mention separately some challenges we faced and solutions we found. We will also uncover connections and continuity between seemingly distinct types of data: unnamed audio tapes in one archive could feature a story represented simultaneously in writing in other archives. We would digitize the audio and time-align it with the transcription, an amalgam of the previous transcriptions and our own editing based on the actual sound. This way, the more

we worked on the corpus, the less clear-cut categorizations would remain, precluding unambigous attribution of any given record of a corpus to a single source.

The figure summarizes some aspects of work on the corpus, as it developped with time (exemplars of Enets speech are in bold, software used for their editing is in italics).

References

Arkhipov, Alexandre & Shluinsky, Andrey. 2025. *User's Guide to INEL Enets Corpus.* (Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 11). Szeged: University of Szeged.

Shluinsky, Andrey; Khanina, Olesya; Wagner-Nagy, Beáta. 2025. *INEL Enets Corpus*. Version 1.1. Publication date 2025-12-31. https://hdl.handle.net/11022/0000-0008-005C-1. Archived at Universität Hamburg. In: The INEL corpora of indigenous Northern Eurasian languages. https://hdl.handle.net/11022/0000-0007-F45A-1

Shluinsky, Andrey & Wagner-Nagy, Beáta. 2024. Enets, In: Behnke, A., & Wagner-Nagy, B. (Eds.). 2024. *Clause Linkage in the Languages of the Ob-Yenisei Area: asyndethic constructions*. Leiden, The Netherlands: Brill.

Ovsjannikova, Maria & Khanina, Olesya. 2018. What happens to a language when it is not spoken any more (evidence from Forest Enets non-finite forms) [Что происходит с языком, когда на нем перестают говорить? (данные нефинитных форм лесного диалекта энецкого языка)], In Semenova, Ks. P. (Ed.). *Small languages in big linguistics [Малые языки в большой лингвистике]*. Conference proceedings. Moscow: Buki Vedi, 151-158.

Elena Lazarenko (Universität Hamburg), Aleksandr Riaposov (Universität Hamburg): ***The INEL corpora of indigenous Northern Eurasian languages: current technical challenges and solutions***

In this talk, we present INEL, a long-term project, the aim of which is to develop and publish deeply annotated digital multimodal corpora of Indigenous Northern Eurasian Languages. So far, it has resulted in 9 published corpora, 6 of which represent Uralic languages.

The process of creating a modern corpus is long and arduous. Within the framework of the project, we developed an efficient general workflow. It starts with multifarious source data (printed materials, manuscripts, fieldwork notes, audio recordings) and ends with a thoroughly annotated corpus that has an online search functionality. We are going to provide an illustrative example of the main workflow stages, discussing our solutions and remaining challenges:

1. Digitization. Transkribus has proven to be a useful tool for OCR (Optical Character Recognition) and HTR (Handwritten Text Recognition) tasks. However, some printed volumes that use Finno-Ugric Transcription remain a tough nut to crack even for state-of-the-art OCR models. An extreme example of that is Juraksamojedische Volksdichtung (Lehtisalo 1947), in which over 100 unique characters and diacritics were used to render Tundra and Forest Nenets. As a result of our digitization efforts, we published a Lehtisalo-2.0 Transkribus model that was extensively trained using the book.

2. Continuous curation of digital corpora. SIL FieldWorks Language Explorer (FLEx) and the EXMARaLDA package (Schmidt & Wörner 2014) are pieces of software that are integral to our workflow during the annotation stage. Both use different flavors of XML to store the data. Since a lot of the work is done manually, often by several project members simultaneously, it has become imperative for us to ensure that the data are in order before they are ready for publication. To that end, we implemented an infrastructure that uses Git and Corpus Services (Riaposov & Lazarenko 2024) for daily maintenance of corpus files. A battery of scripts performs fixes essential to the data consistency, highlights the issues that need to be addressed by the linguists, and keeps the data synchronized across the team.

3. Usability of the corpora. Every published INEL corpus is available online via the Tsakorpus platform (https://tsakorpus.readthedocs.io), which provides a convenient search interface. The

platform itself regularly gets new features. For instance, a recently published corpus-based dictionary of Forest Enets (Shluinsky 2025) was digitized and thus made interoperable with the INEL Enets corpus; dictionary entries contain links to corpus search results featuring the headword, and vice versa.

References

Lehtisalo, T. (1947). *Juraksamojedische Volksdichtung*. Helsinki: Suomalais-Ugrilainen Seura.

Schmidt, T & Wörner, K. (2014). *EXMARaLDA.* In: Handbook on Corpus Phonology, pp. 402-419. Oxford University Press.

Shluinsky, Andrey. (2025). *Forest Enets Dictionary* (Studia uralo-altaica 58). Szeged: University of Szeged. DOI: 10.14232/sua.2025.58.

Riaposov, Aleksandr & Lazarenko, Elena. (2024). Corpus Services: A Framework to Curate XML Corpus Data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 4030–4035, Torino, Italia. ELRA and ICCL.*

The INEL corpora of indigenous Northern Eurasian languages. https://hdl.handle.net/11022/0000-0007-F45A-1.

Niko Partanen (University of Helsinki), Jack Rueter (University of Helsinki) & Kristiina Ojala (University of Helsinki): ***From text collections to modern language corpora: Technical considerations and lessons learned***

In recent years, the Finno-Ugrian Society has been working on the digitization of published text collections. The work has primarily taken place within a pilot project funded by the Finnish Association for Scholarly Publishing, which concluded in 2025. Within the project, several text collections in Komi, Erzya, Moksha, and Livonian were processed and brought into the infrastructure at the Language Bank of Finland. Digitization work is done primarily to make the materials more accessible and useful. Thereby, we also need to evaluate how this infrastructure serves the needs of contemporary researchers. Questions gain new complexity when the needs of additional annotations and reuse in general are considered.

We describe the methods and workflow used to carry out the digitization project. Reflecting on the methods used, and even though the results are good, the undertaking was more challenging than anticipated, and some choices may need to be revisited later. Livonian materials have already been briefly discussed by Partanen et al. (2025), but our scope is broader, and we focus on specific questions within the larger infrastructure. These are the character choices in representing the Finno-Ugric transcription in Unicode, and issues related to normalization and restructuring of the materials when they are brought into the Language Bank of Finland's infrastructure.

The text collections often represent spoken language and connect to many spoken language corpora. Many of them have been created in recent years, or they are in the late stages of preparation (Jouste et al. 2022). In some situations, there are also existing recordings of the text collections, which are available in the collections of some other organizations. The next logical and generally beneficial step would be to bring these together. This will bring the work further from digitization towards the creation of entirely new corpora, datasets, and collections, which in turn offer new potential uses. These include, e.g., speech recognition tools and augmented dictionaries and analysers for morphological annotation, on the one hand, but also the contemplation of further questions, on the other, for example: How do we navigate this landscape, making sure that everyone's work is recognized and scarce resources are not used in duplicate efforts?

Lähteet

Jouste, M., Mettovaara, J., Morottaja, P., & Partanen, N. (2022). Archive infrastructure and spoken language corpora for Saami languages in Finland. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference* (DHNB 2022). RWTH Aachen University.

Partanen, N., Rueter, J., & Ernštreits, V. (2025). Digitization Work at the Finno-Ugrian Society: Livonian Case Study. In *Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages* (pp. 112-122).

Susanna Virtanen (Tartu Ülikool & Helsingin yliopisto): *Kannisto, Munkácsi ja muut – mansin kielen tutkijan digitaaliset aineistot*

Mansin kielen tutkimuksen perustan ovat luoneet Artturi Kannisto ja Bernát Munkacsi, jotka keräsivät pitkillä kenttätyömatkoillaan (Munkácsi 1888–1889; Kannisto 1901–1905) monipuoliset kansanrunousaineistot (Kannisto 1951, 1955, 1956, 1958, 1959, 1963; Munkácsi 1892a, 1892b, 1893, 1896). Kokoelmina julkaistut aineistot edustavat laajaa murrevarianttien kirjoa. Lisäksi mainittujen tutkijoiden kenttätyöaineistoista koottiin paksut sanakirjat. Vielä 2000-luvun alussa nämä kokoelmat olivat tutkijoiden käytössä vain painetussa muodossa, kunnes niitä alettiin pikkuhiljaa digitoida.

Esitelmässäni käsittelen sekä mainittujen kokoelmien eri digitoituja versioita ja niiden käytettävyyttä että uudempia, nykykieltä edustavia media-aineistoja. Esitykseni perustuu hyvin omakohtaiseen kokemukseen lähes 30 vuoden ajalta. Esittelen mm. Münchenin Ludwig Maximilian -yliopiston ylläpitämän obinugrilaisten kielten tietokannan (OUDB), Suomalais-Ugrilaisen Seuran digitoimia aineistoja, Szegedin yliopistossa FinUgRevita-hankkeessa

tuotetun verkkosanakirjan sekä prof. Ulla-Maija Kulosen (nyk. Forsberg) Akatemia-hankkeen (2003–2004) piirissä toimitetut digitoidut aineistot. Keskityn aineistojen toimivuuteen sekä käytettävyyden ja saavutettavuuden haasteisiin. Pyrin esitelmälläni vastaamaan kysymykseen, mitä digitaalisia aineistoja mansin tutkijalla on tällä hetkellä käytössään, mitä lähiaikoina on odotettavissa, ja mitä voisimme vielä tarvita.

Lähteet:

Kannisto, Artturi 1951: *Wogulische Volksdichtung I. Texte Mütischen Inhalts.* Toim. Matti Liimola. SUST 101. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi 1955: *Wogulische Volksdichtung II. Kriegs- und Heldensage.* Toim. Matti Liimola. SUST 109. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi 1956: *Wogulische Volksdichtung III. Märchen*. Toim. Matti Liimola. SUST 111. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi 1958: *Wogulische Volksdichtung IV. Bärenlieder.* Toim. Matti Liimola SUST 114. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi 1959: *Wogulische Volksdichtung V. Aufführungen beim Bärenfest*. Toim. Matti Liimola. SUST 116. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi 1963: *Wogulische Volksdichtung VI. Schicksalslieder.* Toim. Matti Liimola. SUST 134. Helsinki: Suomalais-Ugrilainen Seura.

Munkácsi, Bernát 1892a: *Vogul népköltési gyűjtemény I*. Budapest: Reguly Társaság.

Munkácsi, Bernát 1892b: *Vogul népköltési gyűjtemény II*. Budapest: Reguly Társaság.

Munkácsi, Bernát 1893: *Vogul népköltési gyűjtemény III*. Budapest: Reguly Társaság.

Munkácsi, Bernát 1896: *Vogul népköltési gyűjtemény IV*. Budapest: Reguly Társaság.