DOI: https://doi.org/10.14232/wpcl.2022.7.4

Not sustainable but beautiful? – Some steps towards visual access to multidimensional data collections

Timm Lehmberg

1. Introduction

The following report provides insights in the data analysis and visualization practice of the long term language documentation project INEL.

The principles developed try to take into account the demands of data curation in this special type of projects that lie in the field of tension between sustainability and long term preservation on the one side and short term needs of visual access to the resources on the other side.

A central aspect is to rely on widely adopted and in some cases even proprietary tools and platforms that reduce the effort of interface building and generate maximum flexibility and scalability.

2. Initial Position and Requirements

The 18-year longterm project INEL (Grammar, Corpora, Language Technologie for Indigenous Northern Eurasian Languages) aims at the curation and analysis of language data coming from endangered languages/varieties of the Northern Eurasian Area[1].

Having started in 2016 the project generates deeply annotated digital language corpora and further resources which are made long term available both to the scientific and speaker communities as well as the interested public.

For this purpose, INEL is structured into language specific three-year sub-projects that, following predefined workflows, deal with the curation (in some cases new acquisition), time-aligned transcription, glossing and multi-layer annotation of language data in the respective language/variety. In this way up to the present-day corpora in Selkup, Dolgan, Kamas, and Evenki language have been finalized and published under open access conditions. (A more comprehensive description of the project aims can be found at Arkhipov and Däbritz 2018.)

However, beyond long term availability and sustainability the project considers its mission to provide visual interfaces that comply both to the demands of various target audiences and project research which of course puts high demands on its research data management.

While for example proven solutions to sustainability issues beyond other things result in the use of rather static storage types using established and in many cases generic data standards and formats (i.

[1] The project is funded by the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

Nicole Palliwoda (ed.): *Data Processing and Visualization in Variational Linguistics/Dialectology*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and methodology- Vol.7. 2022.

e. markup standards as defined by Text Encoding Initiative TEI[2]), the creation of visual interfaces[3] follows rather short term dynamics. This might be caused both by the rapid evolution in the area of visualization technology and also by the fact that approaches on data visualization and GUI creation often have to react to short term demands like scientific analyses or data search requirements. Furthermore, it is to mention that GUI creation requires a high level of personal effort which usually cannot be provided by time-limited and in most cases third-party funded research projects.

In order to reduce the effort of interface development and to stay flexible with respect to upcoming analyses and visualization-demands, a multi-level approach was chosen that on its ground level makes use of sustainable data structures and uses flexible tools for indexing and data analyses on top that allow for flexible and user friendly GUI creation.

In the following first the tools that are used for data processing and, based on this, the resources that are created and will be introduced. Based on these concrete examples for resource overarching visualization approaches and analysis will be presented.

3. Tools, platforms and data standards

EXMARaLDA

Depending on the nature of the primary data, a variety of tools like i. e. FLEx and ELAN are being used for the pre-processing of the language data (esp. transcription and glossing). However, finalization, analysis and publication is done exclusively with the help of EXMARaLDA4 (Schmidt and Wörner, 2014), a widely established framework of platform independent desktop applications and data formats to be used along with spoken language corpora.

The major reason for using EXMARaLDA is, besides its wide range of features and tools, the consequent use of XML-based and thus sustainable and interoperable data formats.

An important recent development that has to be mentioned in this context is the establishment of the e. ISO/TEI standard “Transcription of Spoken Language” on the base of the EXMARaLDA data formats.

Corpus Services

Based on the findings from longtime curation work at spoken language corpora the need for a modular and scalable tool collection arose, that would allow for an easification and, if possible, automation of recurring data checks and fixes. As a result Corpus Services, a collection of java based tools was developed initially at the Hamburg Centre for Language Corpora (HZSK) and in the following utilized and further developed in the projects CLARIAH-DE[5], CLARIN-D[6], INEL and QUEST[7]. It contains functionality used for data maintenance, curation, conversion, and visualization, primarily with a focus on data formats used in the environment of EXMARaLDA based spoken language corpora (see above).

[2] https://tei-c.org/

[3] In the following the acronym GUI for graphical user interfaces will be used for any kind of visual interface.

[4] https://exmaralda.org

[5] https://www.clariah.de/

[6] https://www.clarin-d.net/

[7] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html

In the INEL project with the help of corpus services nightly automated checks and fixes were implemented that produce extensive logging information on errors or weak points in the data and thus found a solid base for a high quality of the research data already in the process of creation.

Tsakorpus

The Tsakonian Corpus platform[8], is a corpus search platform that provides a web based search interface which not only allows intuitive multilayer search over annotations and glosses but also provides audio-aligned search results (if existing). The Tsacorpus backend is implemented with the help of *elasticsearch* and thus enables data querying that combines flexibility with respect to data models and schema with a high level of indexing and thus querying performance (see below). These major characteristics make Tsakorpus a powerful tool for the web-based corpus search in the project. An important step towards the seamless integration of INEL corpora into the Tsakorpos platform on has been done by Arkhangelskiy et al (2019) who define a workflow for the integration of audio-aligned transcripts formatted in the ISO/TEI standard for "Transcription of Spoken Language" mentioned above.

Elastic Stack

The Elastic Stack is a collection of open source software tools used for the analysis of textual data that is developed and distributed by the company elastic[9]. Its key component, the search engine elastic search (which is based on Apache Lucene[10]), is considered to be one of the most frequently used search engines in production environments. Further components used along with data analysis and visualization in the INEL project are *Logstash* (used for defining and executing ingest pipelines into elasticsearch) and *Kibana,* which provides comprehensive web based and interactive visualization and analysis capability that can easily be setup and scaled with respect to individual research issues. The most common fields of application in digital production environments for the elastic stack are customized search engines, analyses of large amounts of (often time based) data like contained log files and complex platform overarching data visualization. In light of the considerable effort of the development of graphic user interfaces (Lehmberg 2020) Kibana becomes a powerful tool for the out-of-the-box creation of intuitive and user centered interfaces.

4. Core and Accompanying Resources

The spoken language corpora described above form the core of the INEL resources. The use of the EXMARALDA System allows both for sustainable long term availability and the search capabilities provided by the EXMARaLDA System.

However, it is not surprising that in the framework of an 18 year long term initiative that focuses on data acquisition and curation at every point in time additional linguistic resources (like catalogue, lexical and geospatial data etc.) appear to be crucial for research.

Their necessary integration into the project resources or at least the correlation of the information they contain with the project data put high demands on data formats and workflows and, with respect to analysis and visualization issues, require a performant and flexible set of tools to be available.

[8] https://github.com/timarkh/tsakorpus

[9] https://https://www.elastic.co/

[10] https://lucene.apache.org/

In the following a brief overview on these accompanying resources followed by a selection of visualization approaches will be given.

Bibliographic and catalogue data

More than in many other areas of linguistic research, the documentation and analysis of minority languages not only relies on the analysis of the object language data itself but also on information gained from secondary sources of information. The variety of these resources is considered to range from unstructured (often analogue) language data collections to lexicon data, catalogue data and personal notes created by researchers in the past (first and foremost *manuscript fieldnotes*, cp. Sanjek 1990, Sanjek/Trattner 2016). Of course, an essential role is also played by references to secondary literature which, due to the specificity and low documentation of the languages in focus, can be hard to find in published form.

As a reaction to this in the INEL project comprehensive catalogue resources are being created that contain information on secondary resources. These linear and semistructured resources are both used for internal purposes and, in cases where their coverage, structuredness and well-formedness achieve a certain level, made available to the public.

Two prominent resources that have to be mentioned in this context are to mention in this context are the INEL research Bibliography[11] and the digital edition of the Kuzmina archive (Lehmberg 2020).

Geodata

Modeling, visualizing and correlating spatial data puts high demands on the data workflows to be chosen in minority language documentation and analysis. Central problems often arise from the vagueness and ambiguity of the data, variation in the naming of geographic entities and also the fact that spatial information may occur on all layers of a language resource.

As an example of the latter spatial information with different granularity may occur in metadata (both for exploration settings and speaker biography), in the object language itself and also in the form of external resources.

Considering this high level of variety a structured and standardized modeling of all types spatial seems to be problematic (and maybe even not desirable). At the same time the variety and richness of data contains a lot of explicit or implicit information which are relevant for sociolinguistics, typology and many other research areas. As an example, the use of certain toponymes for the description of geo entities may provide information on the presence of language communities in a particular area and time, migration processes, language contact and many more.

5. Visualization Issues

The creation of visual and in an ideal case interactive graphical user interfaces for querying and analysing language data usually requires a high amount of personal and technical effort. It becomes even more complex if the data to be analyzed is structured and formatted rather with a focus on sustainability and elaborated data structure than on efficient querying like in the case of several well established markup based data formats used along with linguistic data. As a common approach, various

[11] http://doi.org/10.25592/uhhfdm.731

steps of transformation and linearization are to be made in order to generate user-friendly output which complies with the respective research issues.

For this reason, the INEL project follows the principle of a maximum adaptation of existing technology to be used for online-publication and visual (search-)interfaces. As an example, the INEL research Bibliography (see above) is stored and made long term available with the help of the BibTeX standard nut published for online-search as static HTML output with a Javascript based search interface, an output that can be created easily using the open-sourced reference manager tool *JabRef*[12].

As a further system that has been adapted in order to be used for out-of-the-box search, analysis and visualization the *Elastic Stack* (see section 2) was chosen. The following sections will provide a selection of insights into several application scenarios.

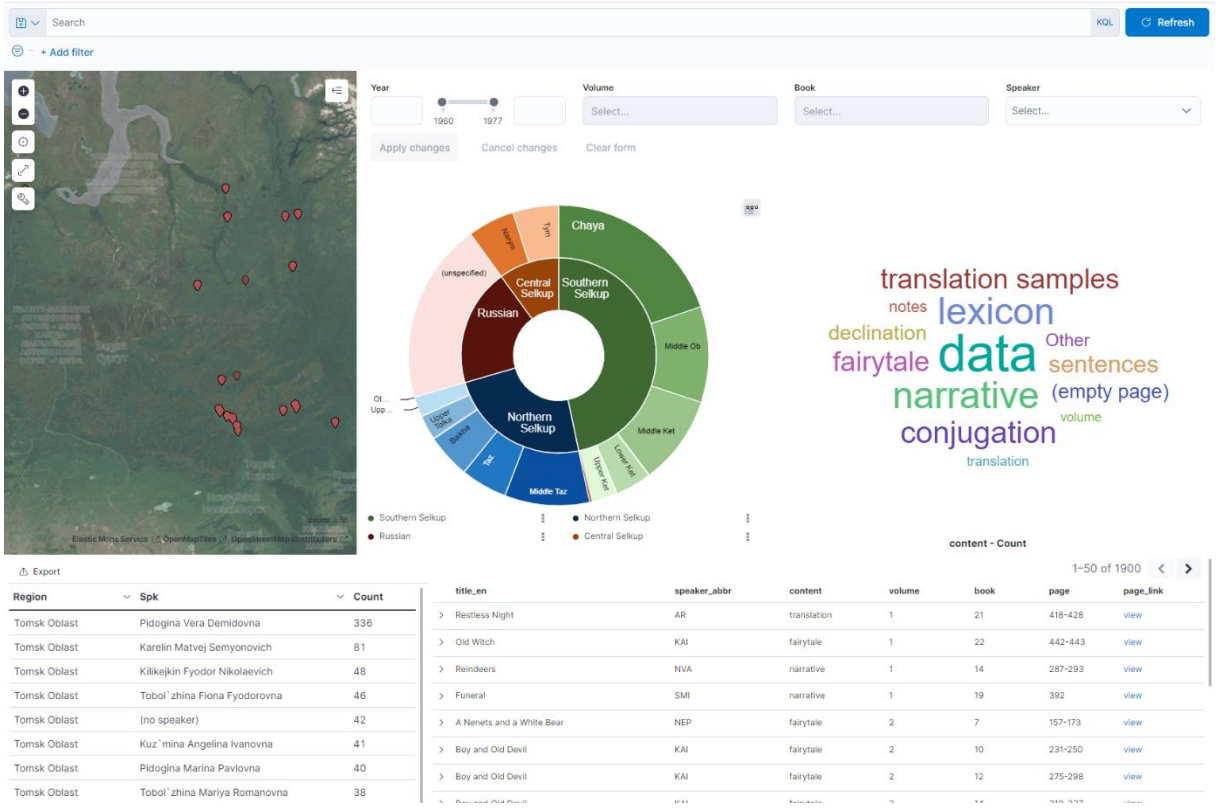
5 .1 Cross resource visualization and analysis

In the case of the abovementioned digital edition of the Kuzmina Archive that consists of digitized (scanned) manuscript pages and a well-structured and consistent tabular catalog (cp. Lehmberg 2020) for the first time in the INEL project a two-layered approach was chosen. To ensure long term availability the catalog is stored in TEI P5 compliant XML format along with scanned facsimiles in the research data repository of Universität Hamburg[13]. However, for analysis issues the catalog was ingested into the project's elastic cluster which then not only allows for effortless GUI creation with respect to the project's individual research demands but also straightforward correlation with other project resources (see Figure 1).

[12] https://www.jabref.org/

[13] https://www.fdr.uni-hamburg.de/

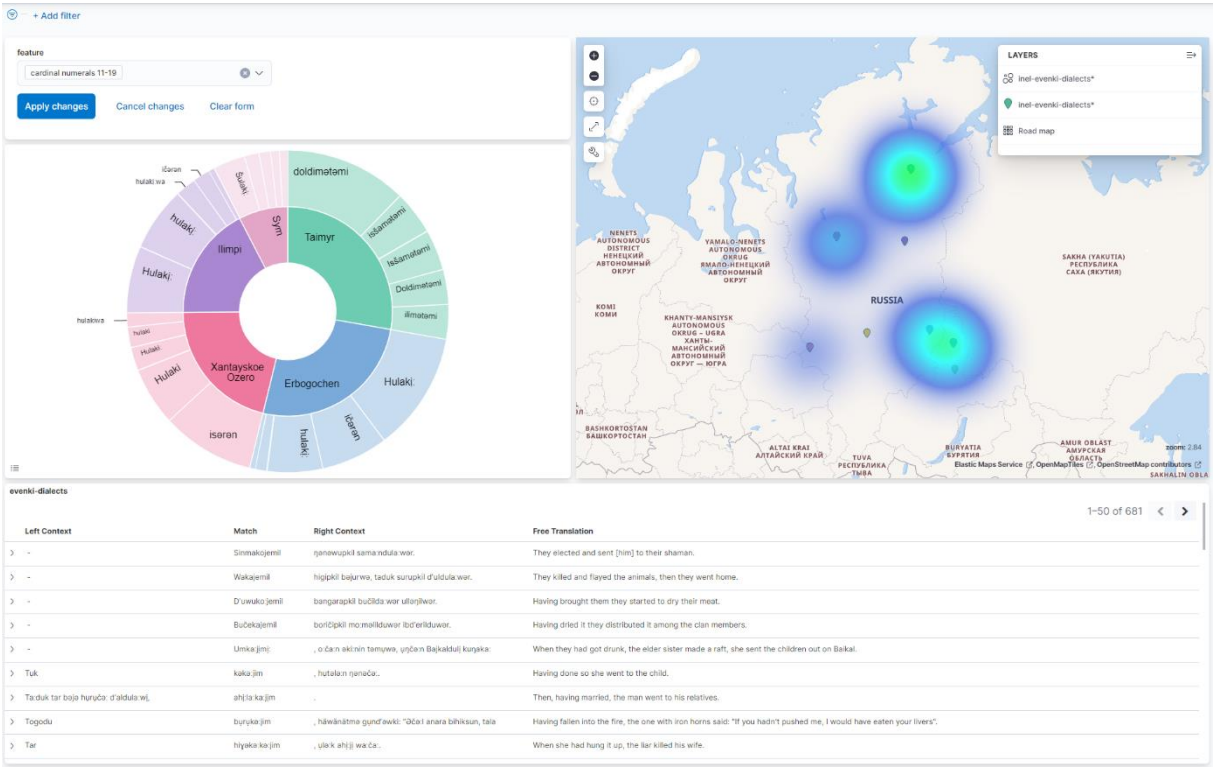
Figure 1 *Dashboard for straightforward access to the catalogized manuscript data of the 'Kuzmina Archive'*



After successful implementation further approaches on ingesting linear structured data, primarily derived from the INEL corpora, were taken. An essential role in this context was played by the tool EXAKT (EXMARaDLA Analyse und Konkordanzprogramm) a standalone tool which provides multilayer search using regular expressions over corpora stored in the EXMARaLDA format. As one central output format it provides linear structured concordances, possibly containing an arbitrary number of additional columns which for instance may contain annotation or metadata values associated with the respective match. These concordances that usually carry information corresponding to specific research issues found a perfect base to be ingested into a data analysis framework like the above mentioned elastic stack. In doing this, they can be correlated with other data to get new insights into the corpora with a minimum of conversion effort.

While earlier approaches of resource overarching analysis in the INEL project aimed at the definition of "linking" data (cp Jettjka/Lehmberg 2020) this method turned out to be much more flexible and applicable because it allows it to react to visualization demands already in the framework of the process of corpus creation. The following figure shows an exemplary visualization generated on the base of the INEL Evenki Corpus.

Figure 2 *Dashboard visualizing the distribution of dialect features of Evenki*



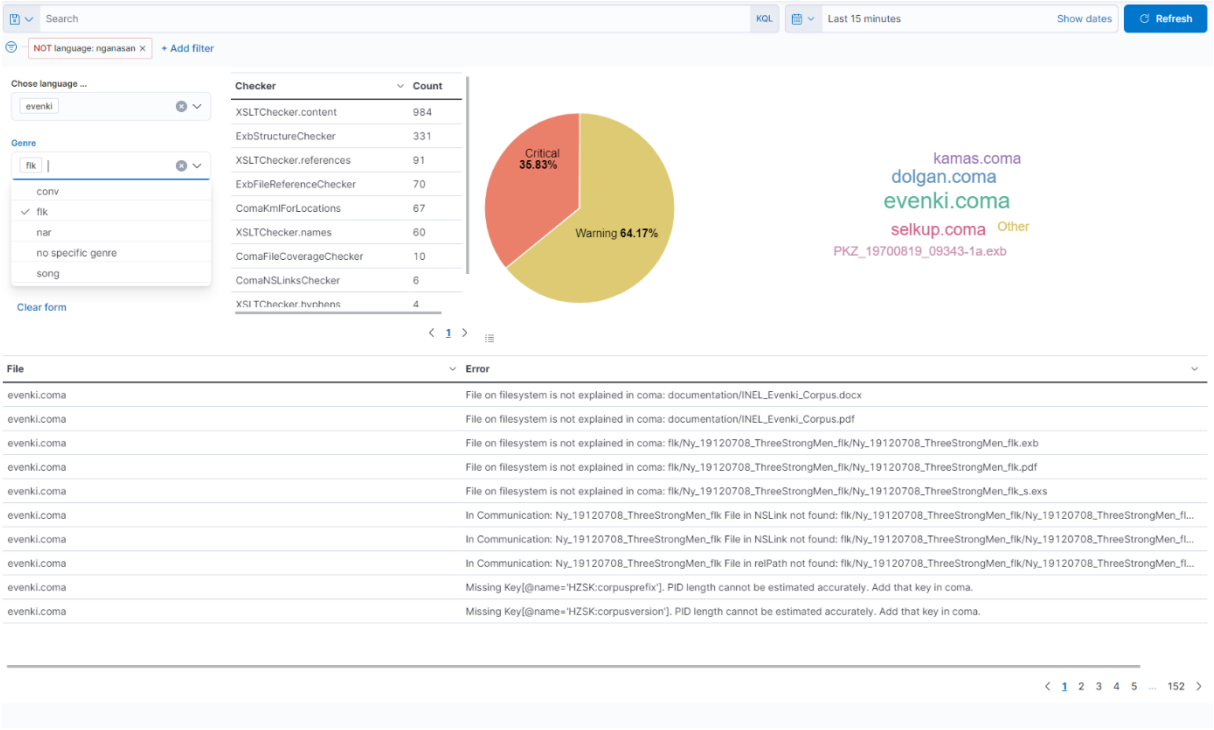
5.2 Visualization for automatic Quality Control and Consistency Checking

As mentioned above, to allow for already using the corpora and resources within the process of their creation, detailed workflows for all necessary steps (including, primary data curation, annotation and glossing, for a detailed description cp. Arkhipov/Daebritz 2018) where defined[14].

Furthermore, deriving the principles of *continuous integration* from software development projects (like described by Ferger/Hedeland 2020), daily consistency checks and (in many cases automatic) fixes are applied to the data. A central function in this process is taken over by the abovementioned software framework *Corpus Service* (see section 2). The output of these daily checks is stored in the form of comprehensive error lists containing information on inconsistencies in metadata, data structure, noncompliance of project specific conventions and further categories. In an initial step a hard coded html output and in addition XML based error lists that easily can be imported into the EXMARaLDA Partitureditor for error correction was chosen as output format. Facing the need for filtering, searching and a more intuitive access to the log files that results from the immense amount (in some rare cases more than 1000) of entries and also their redundancy, it was decided to index and visualize them with the help of the elastic stack. As a result, not only corpus overarching error search and filtering capability was added, also error prone files are highlighted automatically which led to an immense improvement of the user experience (see Figure 3).

[14] Workflowmanagement and -monitoring are proceeded with the help of an Issue tracing System

Figure 3 *Graphical Overview, results of INEL consistency checks*



5.3 Visualization for manual consistency checking

Both core and accompanying project resources contain information that is connected to entries in other project resources or appear in different resources on different layers. As an obvious example, named entities that carry geographic information occur in corpus metadata (for speakers and sessions), catalog resources and also in speaker utterances themselves. In order to be used for georeferencing and geovisualization and further resource overarching analyses they need to be harmonized with respect to writing (i. e. cyrillic writing and its latin transliteration) and the geo coordinates they refer to.

The same applies for the naming of individuals and bibliographic references. In cases where the necessary harmonization of these entries cannot be done by automatic consistency checks and fixes, indexing and creating aggregation based visualization allows for straightforward manual consistency checking by project members.

6. Conclusion and outlook

As shown in this report it seems to be a suitable approach to meet the visualization needs and requirements in scenarios like given in the INEL project by complying to established and widely adopted tools and standards on the one hand and making use of existing (often proprietary) out-of-the-box solutions for data analysis and visualization on the other hand.

In doing so it becomes possible to build the bridge between workflows that take into account the sustainability and long term availability of the data (which in many cases does not include user friendly visual access) and the everyday visualization demands resulting from for example from consistency checking and analysis. Further steps to be performed in the INEL project will be the correlation of indexed Tsakorpus search data (already stored in the form of elasticsearch indexes) and further (i. e. lexical) data.

However, though the principles and practices described here have been proven successful in short term visualization they have a number of significant disadvantages with respect to long-term

availability. By their very nature, visual interfaces created with the help of proprietary software frameworks like the elastic stack require continuous maintenance and sometimes also financial effort. This becomes even more crucial in cases where visualizations being referenced in publications (ideally using persistent identifiers or at least permalinks) have to be kept long term available.

References

Arkhangelskiy, Timofey, Ferger, Anne and Hedeland, Hanna: Uralic multimedia corpora 2019: ISO/TEI corpus data in the project INEL: *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, 115–124. https://aclanthology.org/W19-0310.pdf.

Arkhipov, Alexander and Däbritz, Chris Lasse 2018: Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology* (3): 9–18. https://doi.org/10.23951/2307-6119-2018-3-9-18.

Ferger, Anne, Hedeland, Hanna, Jettka, Daniel, and Pirinen, Tommi 2020: *Corpus Services* (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.4725655.

Hedeland, Hanna and Ferger, Anne 2019: Towards Continuous Quality Control for Spoken Language Corpora. In: *Proceedings of the 14th International Digital Curation Conference* (IDCC19), https://doi.org/10.2218/ijdc.v15i1.601.

Hedeland, Hanna and Ferger, Anne 2020: Towards Continuous Quality Control for Spoken Language Corpora. *International Journal for Digital Curation*, 15 (1). https://doi.org/10.2218/ijdc.v15i1.601.

ISO/TC 37/SC 4. 2016: *Language resource management* – Transcription of spoken language. Standard ISO 2462:2016, International Organization for Standardization, Geneva, CH. http://www.iso.org/iso/cataloguedetail.htm?csnumber = 37338.

Jettka, Daniel and Lehmberg, Timm 2020: Towards Flexible Cross-Resource Exploitation of Heterogeneous Language Documentation Data. *Proceedings of the 12th Language Resources and Evaluation Conference*, 2901–2905.

Lehmberg, Timm 2020: Digitale Edition des Kuzmina Archivs. *Finnisch-Ugrische Mitteilungen* 44: 121–130.

Sanjek, Rogier (ed.) 1990: *Fieldnotes*. The Makings of Anthropology. Ithaca, London: Cornell University Press.

Sanjek, Rogier and Tratner, Susan (eds.) 2016: *Fieldnotes*. The Makings of Anthropology in the digital world. Philadelphia: University of Pennsylvania Press.

Schmidt, Thomas and Wörner, Kai 2014: EXMARaLDA. In Durand, Jacques, Gut, Ulrike, and Kristoffersen, Gjert (eds.): *Handbook on Corpus Phonology*. Oxford: Oxford University Press, 402–419.

Szeverényi, Sándor and Wagner-Nagy, Beáta 2002: The History of Samoyed Toponymic Research *Onomastica Uralica* 2: 253–259.

Wagner-Nagy, Beáta, Szeverényi, Sándor, and Gusev, Valentin 2018: User's Guide to Nganasan Spoken Language Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and methodology* Volume 1. https://ojs.bibl.u-szeged.hu/index.php/wpcl/issue/view/810

Wagner-Nagy, Beáta and Arkhipov, Alexandre 2019: *INEL Bibliographie* (Version 1.0) [Data set]. http://doi.org/10.25592/uhhfdm.731.