Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2901–2905
Marseille, 11–16 May 2020

Towards Flexible Cross-Resource Exploitation of Heterogenous Language Documentation Data

Daniel Jettka, Timm Lehmberg
Universität Hamburg, Institut für Finnougristik/Uralistik
Max-Brauer-Allee 60, D-22765 Hamburg
{daniel.jettka, timm.lehmberg}@uni-hamburg.de

Abstract

This paper reports on challenges and solution approaches in the development of methods for language resource overarching data analysis in the field of language documentation. It is based on the successful outcomes of the initial phase of an 18 year long-term project on lesser resourced and mostly endangered indigenous languages from the Northern Eurasian area, which included the finalization and publication of multiple language corpora and additional language resources. While aiming at comprehensive cross-resource data analysis, the project is simultaneously confronted with a dynamic and complex resource landscape, that especially results from a vast amount of multi-layered information stored in the form of analogue primary data in different widespread archives on the territory of the Russian Federation. The described methods aim at solving the tension between the needs for unification of heterogenous data sets and vocabularies on the one hand and maximum openness for the integration of future resources and the adaption of external information on the other hand.

Keywords: Digital infrastructure, Language Documentation, Linked Data, Knowledge Graphs

1. Introduction

Since its beginning in 2016 and over an expected runtime of 18 years the INEL project („Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages“)[1] aims at the discovery and documentation of language material for selected lesser documented and mostly endangered indigenous languages and varieties from the Northern Eurasian Area. In several subprojects, each focussing on one particular language, these resources are successively curated with digital methods, enriched with linguistic information, and are finally published in the form of linguistic corpora and other types of language resources. In association with this work comprehensíve linguistic description and analysis is executed over the entire runtime period and on all levels of language descriptions, also within satellite projects.

According to its extended areal, temporal, and also disciplinary focus the aims of INEL go beyond those of many other research projects in the field of language documentation. While the large-scale approach of the project allows for manifold cross-lingual, diachronic and interdisciplinary research on a very high level, the data situation that will be described in the following section puts new requirements to methods of data modelling and technology used within the project. Thereby, major challenges arise from the need for cross-resource data analysis and a rather complex resource landscape on the territory of the Russian Federation.

2. Language Resources and Digital Infrastructure

2.1 Complex Heterogeneous Information Sources

The amount of more or less structured collections of object and meta language data resulting from research on lesser documented languages all over the world is incalculable per se. In particular with regard to indigenous languages from the Northern Eurasian area a lot of research material has been created within the past two decades. It is held by different (both public and private) archives, mostly in Europe and on the territory of the Russian Federation. Like in many other cases, the most prominent type of primary data used for language description in the INEL project are audio recordings of spoken language, often stored in problematic quality on obsolete analogue media. However, there are numerous related collections of, for instance, manuscript data that originate from previous documentation and research projects. Due to the fact that the majority of indigenous languages were traditionally exclusively oral cultures, manuscripts typically do not play a primary role in the study of those languages. However, in particular handwritten notes that were created by researchers during fieldwork often are important information sources that, beyond other things, contain highly relevant information, ranging from object language data with attached translations and glossings, over lexical and grammatical descriptions to complex metadata, figural data and of course individual interpretation by the respective researcher.

As an example, Figure 1 shows excerpts of field notes from the archives of Angelina Ivanovna Kuzmina (volume 1, textbook 3, page 2), which are an integral part of the INEL Selkup Corpus (Brykina et al., 2018) and the *Heinrich Werner Archive* (Ket-Yugh materials volume 2, page 3a), containing figurative information on toponymy and semantics. Both archives are held by the Institute for Finno-Ugric/Uralic Studies (IFUU) at the Universität Hamburg.

Another exemplary information source that contains comprehensive and multi-layered information is the collection of toponyms of the *Dulson Archive at Tomsk State Pedagogical University*[2] (TSPU). It contains approximately 342,000 hand-written notes on toponyms used by the indigenous population of the Northern Eurasian region. The resource provides linguistic evidence of the presence, movement and contact of various language communities on the territory of the Russian Federation.

[1] https://inel.corpora.uni-hamburg.de/

[2] https://www.tspu.edu.ru/

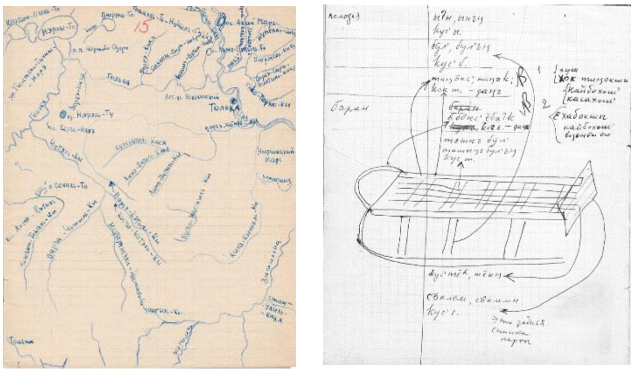


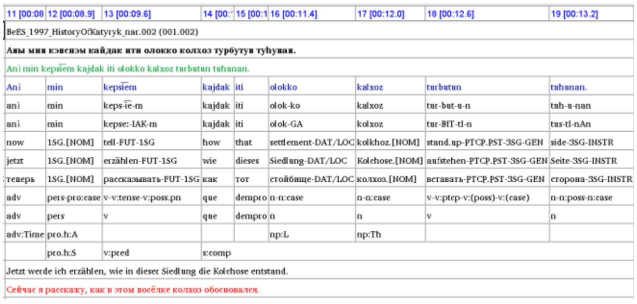
Figure 1: Manuscript fieldnotes from Kuzmina archive and Werner Archive

There is no doubt that these collections and the large number of further comparable resources are worth to be curated, digitized and made available with the help of well-established methods of manuscript research. However, in practice it seems to be common to only extract parts of information that is considered to be relevant and to leave the remaining information behind. (Exceptions may be some printed editions which are derivatives of fieldnotes, cp. Sanjek (1990); Sanjek, Tratner (2016)).

2.2 Exploitation of Language Resources

After the successful completion of the first three subprojects, corpora of the languages Selkup (Brykina et al., 2018), Dolgan (Däbritz et al., 2019) and Kamas (Gusev et al., 2018) were published under open access conditions at the CLARIN-D repository hosted by the Hamburg Centre of Language Corpora (HZSK)[3]. The corpora contain transcribed and richly annotated audio and partly also manuscript data. All annotation had to be carried out manually or in case of glossing semi-automatically on the base of lexical resources. This includes among other levels morphologic annotation and glossing, syntactic functions, code switching, borrowing etc. as well as translation layers into Russian, English and German (see also Arkhipov/Däbritz, 2018). The compilation of all three corpora followed identical workflows and processing pipelines that were based on preliminary work on the Nganasan Spoken Language Corpus (Wagner-Nagy et al. 2018) and refined in the INEL project.

Data models and transcription/annotation workflows used for the corpora are based on the EXMARaLDA-System



11 [00:08	12 [00:08.9]	13 [00:09.6]	14 [00:1	15 [00:1	16 [00:11.4]	17 [00:12.0]	18 [00:12.6]	19 [00:13.2]
BeES_1997_HistoryOfKatyryk_nar.002 (001.002)								
Аны мин кэпсиэм кайдак ити олокко колхоз турбутун туһунан.								
Ani min kepsīem kajdak iti olokko kalxoz turbutun tuhunan.								
Ani	min	kepsīem	kajdak	iti	olokko	kalxoz	turbutun	tuhunan.
ani	min	keps-īe-m	kajdak	iti	olok-ko	kalxoz	tur-but-u-n	tuh-u-nan
ani	min	kepse:-IAK-m	kajdak	iti	olok-GA	kolxoz	tur-BIT-tI-n	tus-tI-nAn
now	1SG.[NOM]	tell-FUT-1SG	how	that	settlement-DAT/LOC	kolkhoz.[NOM]	stand.up-PTCP.PST-3SG-GEN	side-3SG-INSTR
jetzt	1SG.[NOM]	erzählen-FUT-1SG	wie	dieses	Siedlung-DAT/LOC	Kolchose.[NOM]	aufstehen-PTCP.PST-3SG-GEN	Seite-3SG-INSTR
теперь	1SG.[NOM]	рассказывать-FUT-1SG	как	тот	стойбище-DAT/LOC	колхоз.[NOM]	вставать-PTCP.PST-3SG-GEN	сторона-3SG-INSTR
adv	pers-pro:case	v-v:tense-v:poss.pn	que	dempro	n-n:case	n-n:case	v-v:ptcp-v:(poss)-v:(case)	n-n:poss-n:case
adv	pers	v	que	dempro	n	n	v	n
adv:Time	pro.h:A				np:L	np:Th		
	pro.h:S	v:pred	s:comp					
Jetzt werde ich erzählen, wie in dieser Siedlung die Kolchose entstand.								
Сейчас я расскажу, как в этом посёлке колхоз обосновался.								

Figure 2: Audio-aligned Dolgan transcript and annotation/glossing-layers

(Schmidt and Wörner, 2014). This includes modeling of metadata for *subjects* (*"speakers"*, containing information on language background, socio-biographic and geographic information) and *sessions ("communications"*, containing also spatial data and references) modelled with the help of the *EXMARaLDA Corpus Manager* COMA (Wörner 2012).

To ease accessibility, all corpora are made available online and are searchable with the help of the *Tsakorpus* search platform via the project website. For this purpose, transcripts are converted to the ISO/TEI standard for *Transcription of Spoken Language*[4], indexed and stored in a document-based ElasticSearch index (Arkhangelskiy et al. 2019).

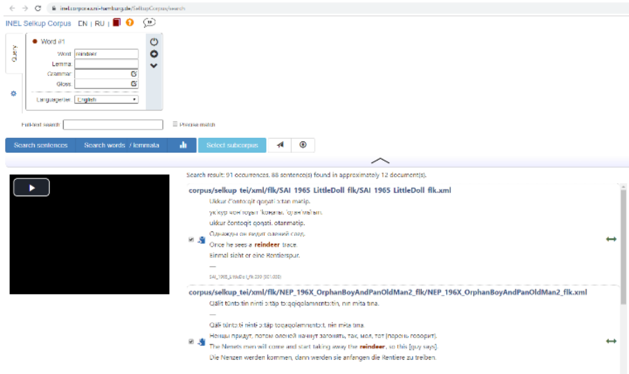


Figure 3: Online search for the INEL Selkup Corpus

In addition to the language corpora, catalogue resources were generated and made available online. They include among others a comprehensive indexed bibliography of relevant publications on the languages in the project focus that contains a constantly increasing number of (currently 2023) entries. Like all project resources it is available online, both for download in BibTeX format from the FDR@UHH[5] (Research Data Repository at Universität Hamburg) and in a full-text search on the INEL project website.

2.3 Implications

Due to the long-term and large-scale approach it can be expected that additional collections of possibly relevant primary data will appear over the entire project runtime. Thus, it is unforeseeable which data types they will contain and at what point in time they might become relevant for research conducted along with the project. The resources might contain highly specific information on different description layers that will not fit into the data models used by the project at the current point in time. Accordingly, in the initial project phase it turned out to be necessary to catalogue existing resources (in some cases just parts of them) to make them findable when they will become relevant.

Regarding the well-structured present resources compiled within the project the situation appears to be slightly different. Though the EXMARaLDA data model allows for correlated querying of object language, annotation and metadata, it is crucial not only to query project resources

3 https://corpora.uni-hamburg.de/

4 http://www.iso.org/iso/catalogue/detail.htm?csnumber=37338

5 https://www.fdm.uni-hamburg.de/en/

separately but also to enable the INEL infrastructure to support analyses on superordinate layers like i. e. correlating catalogue data with corpus metadata, object language data and/or its annotation layers. Thereby the applied method should also allow for referencing and integration of external and future data resources and/or their content.

To approach both problems in a sustainable way, the decisions in the area of data modeling that will have to be made will deal with the tension between unification of data sets and vocabularies on the one side and maximum openness to future resources and research queries on the other side. Therefore, as one solution it was decided to identify standardizable connecting information in the existing project resources and to make sure that connections to internal and external resources of any type can be represented and resolved in the future. However, it turned out to be necessary to refer to (or define) at least minimum standards, controlled vocabularies create unifiable (at least in parts) datasets/information models. The following section will describe the current state of referencing and correlation in the project.

3. Methods of Referencing and Correlation

As already indicated above, a comprehensible and straightforward approach is the identification of implicitly connective information which initially was not intended or at least not represented consistently as referential information. Entities that are considered to be predisposed could for instance be the following:

- **Individuals** such as *subjects* and *corpus editors* that are defined in the unified COMA metadata sets can be referenced to *authors, speakers, collectors* etc. in external resources.
- **Geo-Locations** that are also defined in the *communication/session* metadata entity as part of the COMA metadata can be referenced to any type of spatial data like included in bibliographies, toponym databases etc. by GPS-coordinates and established principles of geo-referencing.
- **Abstract data entities** like *communications/sessions*, may contain ID/IDREF information to any type of item
- **Genres** (in INEL: *conversation, folklore, song, narrative, translation and miscellaneous*) are defined as an approach to define a controlled vocabulary for the clustering of both internal and external resources.
- **Languages,** usually identified by ISO code 639-3[6]

The complexity of representing referential information naturally depends on the complexity of the underlying data model. When dealing with a single resource that consists of one or multiple uniform files, it is quite easy to define schemas (e. g. document grammars), that can be used for the (*intra-resource*) validation of consistent categories, controlled vocabularies, or *dataset-internal references* (ID/IDREF) between individual units inside the single dataset.

Taking into account, however, that some categories or vocabularies should be used consistently across different kinds of resources, and moreover, units from different datasets should be made referable by one another, the necessary manifestation of dataset-overarching, or *dataset-external references* adds a considerable amount of complexity.[7] In this case, certain mechanisms have to be applied to either make sure that the information on controlled vocabularies, IDs, etc is shared and kept synchronous between several different schemas, or selective (*inter-resource*) validation has to be applied to individual (referential) parts of the resources.

The technical realization and utilization of interoperable resources is independent from whether the reference data (category lists, vocabularies, norm data, IDs, etc) originates from external sources (e.g. VIAF, GND, DBPedia, Wikidata) or from internal definitions. Of course the use of external standard identification schemes is desirable because it allows for the establishment of connections to further related information (as Linked Open Data), but the necessary identification schemes for this cannot be applied to all relevant data in the project, e.g. because unpublished texts do not have ISBNs, or only specific researchers can be found in authority data. Where possible standard identification should be used, but to make project-wide referencing feasible, internal identification schemes have to be applied as well.

4. Exploitation of Linked Resources

The previous sections have indicated that the INEL infrastructure is designed to principally cover any dataset that can be of interest for the documentation of the languages in focus. In order to make resource-overarching sense of the data, however, particular units in the datasets are controlled by identification and referencing methods and so a set of language resources is created which are interoperable on different levels. Following this approach, a successively growing knowledge graph is built, from which data can be extracted on the basis of the shared information, which in turn can provide different analytic views on the entirety of the INEL data.

To demonstrate the potential of this approach, a sample graph has been extracted that shows the references between parts of the INEL bibliography (see Figure 4), the research data catalogue (see Figure 5) and corpus metadata (see Figure 6). It presumes the unique identification of bibliographic items that are referenced from descriptions of research data (in the catalogue) and corpus metadata.

[6] https://iso639-3.sil.org/code_tables/639/data

[7] The necessary use of different software for the data entry, enrichment, and curation of different language resources adds even more to the complexity.

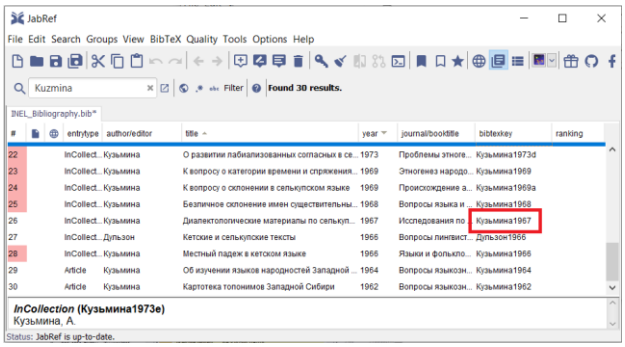


Figure 4: Definition of publications IDs in bibliography

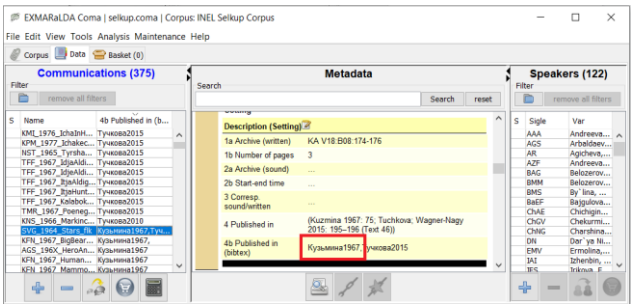


Figure 5: References to publication IDs in corpus metadata

Figure 6: References to publication IDs in the research data catalogue

The resulting derived resource on the one hand shows which data from the resource catalogue and which sessions from the corpora have been published in certain bibliographic items. On the other hand, it demonstrates that some have actually been published in different bibliographic items. To make the knowledge graph visually and structurally accessible, it has been instantiated in a graph database, which can be browsed and queried via an online user interface[8] (see Figure 7).

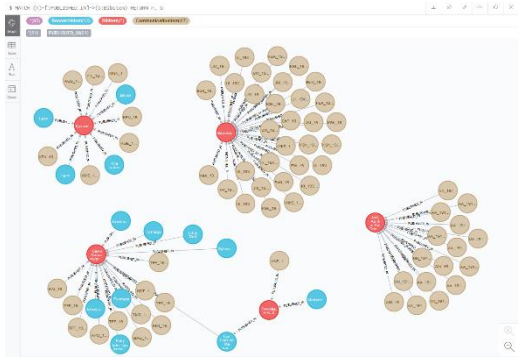


Figure 7: Extracted Knowledge Graph

[8] Neo4j browser, see https://neo4j.com/

The sample above is only one of many imaginable ways to represent the extracted information, and it could be derived in any data format and instantiated in any suitable framework for further visualization and analysis.
Furthermore, the example only superficially indicates the potential of the approach which becomes more apparent when taking into account that a lot of further referential information can be used for the exploitation across resources (e.g. languages, subjects, locations, etc.) as it allows the inclusion and/or concentration on several other categories that establish relations between entities in the INEL data.

5. Conclusion

Due to the fact that cross-resource data analyses require the existence of curated and well-structured language resources whose labour- and time-intensive creation was the focus of the initial project phase of the INEL project, the approaches described here are still in an early state. A crucial result, however, is the requirement to identify data structures and entities, that have the potential both of internal and external cross-resource referentiality, as early as possible to grant for maximum openness to future content and data analysis.
Especially in connection with long-term approaches that aim at language documentation this insight should have strong impact on the conception of data structures, controlled vocabularies and standardization in general.
At the same time the definition of referential information plays a crucial role in the area of resource-internal and resource-external validation. For instance, in the INEL project automated mechanisms for the synchronization of geolocations as defined in a KML-based resource and the location metadata in different resources were already implemented successfully, and additional validation mechanisms aiming at other information types will follow.

6. Acknowledgements

This publication has been produced in the context of the joint research funding of 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.

7. Bibliographical References

Arkhangelskiy, T., Ferger, A. and Hedeland, H. (2019). Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In: Proceedings of 'The fifth International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2019)', January 7—January 8, 2019 Tartu, Estonia

Arkhipov A. and Däbritz C. L. (2018). Hamburg corpora for indigenous Northern Eurasian languages. In: Tomsk Journal of Linguistics and Anthropology, (3), pp 9–18. https://doi.org/10.23951/2307-6119-2018-3-9-18

ISO/TC 37/SC 4. 2016 Language resource management – Transcription of spoken language. Standard ISO

2462:2016, International Organization for Standardization, Geneva, CH.

Sanjek, R., editor, (1990). Fieldnotes. The Makings of Anthropology. Ithaca, London: Cornell University Press.

Sanjek, R. Tratner, and Susan W., editors, (2016). Fieldnotes. The Makings of Anthropology in the digital world. Philadelphia: University of Pennsylvania Press.

Wagner-Nagy, B. and Szeverényi S. and Gusev, V. (2018). User's Guide to Nganasan Spoken Language Corpus. In: Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 1.

Wörner K. (2012). Finding the balance between strict defaults and total openness: Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager. In: Multilingual Corpora and Multilingual Corpus Analysis, Vol. 14, pp. 383–400. John Benjamins.

8. Language Resource References

Brykina, M., Orlova, S. and Wagner-Nagy, B. (2018). INEL Selkup Corpus. Version 0.1. Publication date 2018-12-31. Archived in Hamburger Zentrum für Sprachkorpora. http://hdl.handle.net/11022/0000-0007-CAE5-3. In: Wagner-Nagy, B., Arkhipov, A., Ferger, A., Jettka, D. and Lehmberg, T., editors, (2018). *The INEL corpora of indigenous Northern Eurasian languages*.

Däbritz, C. L., Kudryakova, N. and Stapert E. (2019). INEL Dolgan Corpus. Version 1.0. Publication date 2019-08-31. http://hdl.handle.net/11022/0000-0007-CAE7-1. Archived in Hamburger Zentrum für Sprachkorpora. In: Wagner-Nagy, B., Arkhipov, A., Ferger, A., Jettka, D. and Lehmberg, T., editors, (2018). *The INEL corpora of indigenous Northern Eurasian languages*.

Gusev, V., Klooster, T. (2018). INEL Kamas Corpus. Version 0.1. Publication date 2018-12-31. http://hdl.handle.net/11022/0000-0007-CAE6-2. Archived in Hamburger Zentrum für Sprachkorpora In: Wagner-Nagy, B., Arkhipov, A., Ferger, A., Jettka, D. and Lehmberg, T., editors, (2018). *The INEL corpora of indigenous Northern Eurasian languages*.