From language documentation data to LLOD: A case study in Turkic lemon dictionaries

Christian Chiarcos, Désirée Walther, and Maxim Ionov

Goethe-Universität, Applied Computational Linguistics,
Robert-Mayer-Straße 11-15, 60325 Frankfurt, Germany
{chiarcos,dwalther,ionov}@informatik.uni-frankfurt.de
http://acoli.informatik.uni-frankfurt.de

Abstract. In this paper, we describe the Lemon-OntoLex modeling of dictionaries created within language documentation efforts. We focus on exemplary resources for two less-resourced languages from the Turkic language family, Chalkan and Tuvan. Both datasets have been conveted into a Linked Data representation using the Lemon-OntoLex data model, with an extensible converter written in Python. We compare the conversion process for two both lexical resources, we analyze the difficulties we encountered during the conversion process and discuss the cases which caused the most common problems during the conversion. Furthermore, we evaluate the quality of converted dictionaries using specially designed SPARQL queries, and by manually checking random samples of the data. Finally, we describe the future application of this data within a lexicographic-comparative workbench, designed to facilitate language contact studies.

Keywords: Lemon-OntoLex model, Turtle, Turkic languages, SPARQL

1 Background and Motivation

Linguistic Linked Open Data (LLOD) has become widely popular in the language resource community in recent years, and with particular success in the area of machine-readable dictionaries, where the Lemon-OntoLex model now has become widely adopted. While Linked Open Data is also receiving substantial resonance in the areas of language documentation and typology for a considerable period now [6, 8, 13, 7], the application of Lemon-OntoLex has been rarely discussed in this context, so far. Here, we describe the application of Lemon-OntoLex to dictionary data from two exemplary low-resource languages from the Turkic language family.

The research described in this paper is conducted as part of the BMBF-funded Research Group "Linked Open Dictionaries (LiODi)" (2015-2020) at the Goethe-Universität Frankfurt, Germany, and our activities focus on uses of Linked Data to facilitate the integration data across different dictionaries, or between dictionaries and corpora. As a cooperation between researchers from

natural language processing and empirical linguistics, LiODi aims at developing methodologies and algorithmic solutions to facilitate research on comparative lexicography in the context of linguistic, cultural, sociological and historical studies. In particular, we develop a workbench that facilitates the cross-linguistic search of semantically and / or phonologically related words in various languages. LiODi is a joint effort of the Applied Computational Linguistics (ACoLi) lab at the Institute of Computer Science and the Institute of Empirical Linguistics at Goethe University Frankfurt, Germany, with a focus on Turkic languages (pilot phase, 2015-2016), resp. languages of the Caucasus (main phase, 2017-2020) and selected contact languages.

One main type of data in the project are dictionaries, and the conversion of an etymological dictionary of the Turkic languages has previously been described by [2]. Here, this approach is extended to another category of lexical data, dictionaries and word lists as created as part of language documentation efforts of two Turkic languages: Chalkan and Tuvan. By means of an implemented converter, XML-generated lexical language documentation data are converted into an RDF representation with a model based on the Lemon-OntoLex model. Subsequently, a check is made to determine the extent to which a generic converter for lexical resources is possible or how much effort is needed to expand the converter. Finally, evaluation steps are carried out using SPARQL queries on a SPARQL endpoint as well as a random check of the output data.

2 Data

The lexical resources described in the paper are a part of RELISH project[1] ([5]) and can be accessed with LEXUS tool[2] ([12]) which allows exporting in an XML format with LEXUS schema.

Chalkan language[3] is a variety of Northern Altai language, a Turkic language spoken in South Siberia ([11, p. 67]). A Nothern Altai have no fixed literary norms and many different dialects, mainly spoken in rural areas. It is heavily influenced by the Russian language. As a result, there are strong differences in the grammaticalization between the written and the spoken ([11, p. 74]).

Tuvan language also belongs to the Turkic language family. It is spoken in the Republic of Tyva, a part of Russian Federation, to the southwest of Tofalaria Siberia region. It is also spoken by some Mongol speakers ([4, p. 2]). Similar to other languages of the Altai region, Tuvan was strongly influenced by Russian languages. Also it had a strong Mongolian influence. Initially, Tuvan orthography was based on the Latin alphabet, but since 1943 it has only been written in Cyrillic ([3]).

[1] `https://tla.mpi.nl/relish/`.
[2] `https://tla.mpi.nl/tools/tla-tools/older-tools/lexus/`.
[3] Sometimes the language name is spelled *Chelkan*.

3 Modeling

The modeling is based on the Lemon-OntoLex model. Additionally, the vocabularies SKOS ([10]), LexInfo ([9]) and RDFS are used. Furthermore, a new namespace is created, which contains lexical entries, forms, senses and concepts.

According to the Lemon-OntoLex model, *Form*, *Lexical Sense*, *Lexical Entry*, *Lexical Concept* and *Concept Set* are represented as classes. The lexical form contains the written and the phonetic representation. The lexical sense encompasses only the range of use in a note of the SKOS vocabulary, if this information is present. The decision to outsource the semantic meaning into a concept was also made to enable a semantic search based on a lexical set of concepts. Thus, words with the same semantic meaning refer to a common concept. This concept refers not only to the relevant lexical entries, but also to the corresponding lexical senses. The core of the Lemon-OntoLex model was extended by homonym relations between lexical entries. The decision to use the Lexinfo vocabulary to represent homonym relations instead of other possible ways (e.g. using the Lexico-Semantic Relations of the vartrans module) was dictated by the fact that we were already using the LexInfo vocabulary to represent the parts of speech. Our goal was to use as few vocabularies as possible.

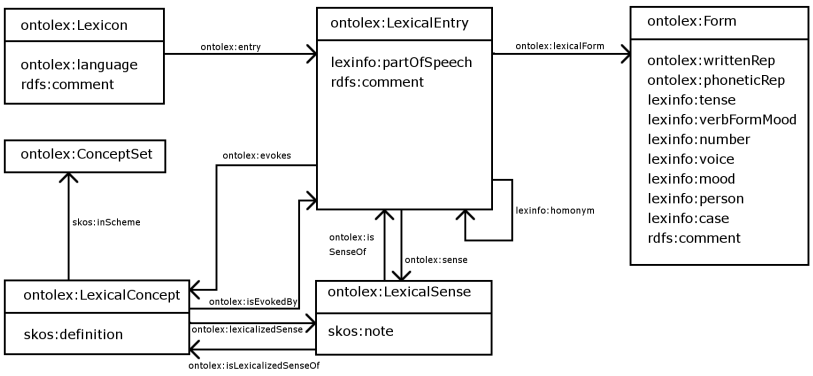


Fig. 1. Common data model of Tuvan and Chalkan based on the Lemon-OntoLex model.

4 Implementation

The main goal of the converter is the implementation of the modeling decisions of the Chalkan lexical resource as well as their reuse for the implementation of the

modeling of the Tuvan data. The extension to Tuvan seemed sophisticated with regard to the differentiated features and different element hierarchies between the two lexical resources, which are in XML format.

In the lexical resource Chalkan a division of lexical entries is necessary, since several parts of speech were originally assigned to a lexical entry. A challenge is, above all, the assignment of abbreviated parts of speech to the correct forms in the LexInfo vocabulary. We used a dictionary for those abbreviations in order to be able to reuse them for other lexical resources. Compared to the lexical resource Tuvan, however, it always has one lexical form.

Based on the described experiences a generic converter can be only implemented without the consideration of specific exceptions and leading to a loss of information.

However, abbreviations and lexicon-specific exceptions are implemented for this reason.

Due to the differentiation of lexical entries, the creation of several forms due to the separation of lexeme chains and the formation of different lexical senses, concepts and homonym relations for a single entry, this converter is extremely multifaceted precisely because of the differentiated features of both lexical resources. Due to this reason, it is impossible to apply the converter to the new lexicons right away, but the modular architecture makes it easy to adapt it to the specifics of a new resource.

Above all, the modularization of individual steps of the modelling provides a good overview, is easily expandable and existing modules can be reused. The converter automatically recognizes the lexical resource and encapsulation of the name spaces required for the output file provides an acceptable overview.

5 Evaluation

5.1 Important Facts of the output files

The Chalkan dataset was converted to 3 165 lexical entries. Newly created lexical entries were linked using homonym relations. This resulted in 208 new entries, which means that the total number of lexical entries in the resulting dataset is 3 373. A former lexical entry has been divided between two and four times, and thus has a maximum of three homonym relations for four divisions.

The number of triples that was modeled for Chalkan is 68 286, providing a wide range for search queries. Above all, instead of abbreviations, the respective part of speech was assigned to the LexInfo vocabulary ([9]).

Tuvan data has some peculiarities within its 7 482 lexical entries. Several written representations contained chains of different lexemes separated by commas after the initial conversion. This lead to their assignment to the same entry. The maximum number of a lexeme chain is eleven. The number of new lexical entries resulting from a lexical entry by division is two, which have a homonym relation with each other. A total of 17 additional lexical entries were generated during the modeling of Tuvan data, which means that the total number of lexical

entries amounts to 7 499. The data is represented by 136 580 triples, almost all modeling decisions for Chalkan could be applied to Tuvan. Lexeme chains form the exception among others.

5.2 Conversion evaluation

Because of the fact that two dictionaries are available as a result, the queries do not have to be restricted to a lexicon, but can specifically address both lexicons.

After looking into the lexical resources, the question arises whether there are different entries that have the same definition and have not yet been merged into one lexical concept. Therefore, the resource was stored in two separate entries. The example "hunting", which is frequent in languages due to the cultural reasons, is investigated. If the SPARQL query shown in Fig. 2 provides results, the concepts concerned can be merged and the lexical entries refer to the same lexical concept.

```
1   PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo# >
2   PREFIX skos: <http://www.w3.org/2004/02/skos/core# >
3   PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex# >
4
5   SELECT ?entry ?concept ?definition
6   WHERE {
7         ?concept skos:definition "hunting"@en ;
8            skos:definition ?definition .
9         ?concept ontolex:isEvokedBy ?entry .
10  }
```

Fig. 2. Checks whether "hunting" is associated with multiple concepts as well as entries.

entry	concept	definition
ontology:tuvan...4cb8_0	ontology:concept_tuvan...c94cca	"hunting"@en
ontology:tuvan...2a00_0	ontology:concept_tuvan...452a10	"hunting"@en

Fig. 3. Output of query in Fig. 2.

The result of the query (Figure 3) shows that two lexical entries with two different lexical concepts for the definition "hunting" exist in the Tuvan dictionary.

In addition, it was examined whether a word in a language can have several meanings, which was the case in Chalkan.

For the definition of "father", it was examined whether the spelling is the same or similar in both dictionaries. In fact, there were even two matches, reflecting the similarity of both languages.

Additionally, sample-based check could be carried out because the IDs of the lexical resource were transferred to the output file. The original identifiers from the lexical resource were used as part of the URI of each lexical entry, form and sense. Using these IDs, each form, sense, and lexical entry could be examined by random sampling and the correctness could be determined.

6 Application

The primary goal of the LiODi project is to develop a workbench and associated methodologies to facilitate language contact studies in Eurasia and the Caucasus area. The **Comparative-Lexicographical Workbench** (Fig. 4) provides novel search functionalities extending the functionality of existing platforms, form-based search and gloss-(meaning-)based search, currently applied to the Turkic language family and its contact languages.

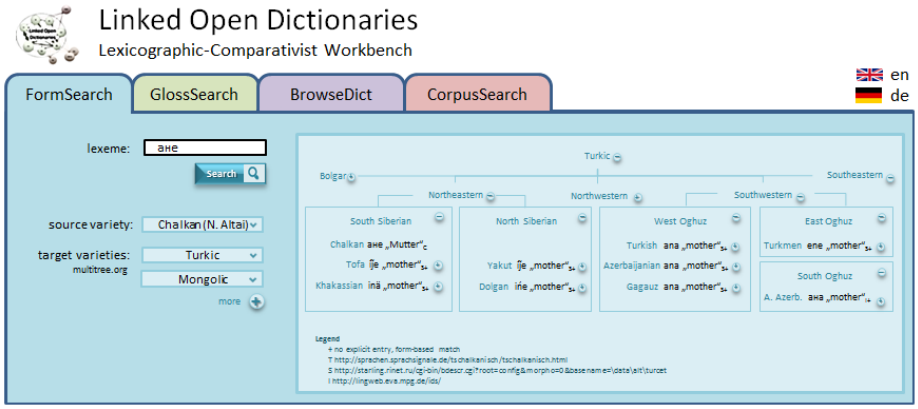


Fig. 4. Design study: Form-based search in the Comparative-Lexicographical Workbench

– *gloss-(meaning-)based search*
  Dictionary lemmas are complemented with a gloss paraphrasing their meaning. Linked Data allows transitive search over sequences of bilingual dictionaries (e.g., Chalkan-Russian-English).
– *form-based search*
  Given a lexeme in a particular language, say, Chalkan, and a set of related

languages, say, the Turkic languages in general, the system retrieves phonologically similar lexemes for the respective target languages.

Both search functionalities aim to detect candidate cognates. The data provided by Starling represents a gold standard, but can also be directly integrated into the search process:

In Fig. 5, we query for Chalkan *ana* and possible cognates from Turkic (as an inherited word) or Mongolic (as a possible source of loan words). The results are organized according to the taxonomic status of the varieties in `www.multitree.org`. They include a gloss from a Chalkan dictionary (marked by subscript C), but in addition provide form-based matches (subscript +) from the Starling dictionaries (S), e.g., with Turkish *ana* and its etymologically corresponding forms, etc.

A prototype of this workbench is available[4]. Albeit still limited in coverage and functionality, it illustrates a core strength of the Lemon-Ontolex-based approach: Given a number of bilingual dictionaries, we can use identical SPARQL fragments to retrieve word lists over which then a transitive closure can be calculated.[5] This is illustrated in Fig. 5 with a screenshot of the workbench prototype for the Chalkan word *küski* and the corresponding SPARQL query,[6] with corresponding properties in different Lemon dictionaries being highlighted.

The development of Lemon towards a community standard is still in progress, even though the publication of the W3C community report in May 2016[7] set a stable milestone. While many early adopters of Lemon still use older specifications or resource-specific extensions (e.g., DBnary), it is expected that these will eventually converge towards the current specification. For example, DBnary still uses the older Lemon-Monnet model at the time of writing, but is currently in migration to Lemon-OntoLex (Gilles Serraset via the OntoLex mailing list, 2017-03-10). In the course of this process, we will eventually be able to apply the *same SPARQL property path* to internal and external resources to retrieve bilingual word pairs that may then be the basis for online transitive search across dictionaries.

Even more so, Lemon (and the SPARQL concept of federation with the keyword `SERVICE`) allows us to evoke remote data sources directly. However, querying a local graph instead is normally a more scalable solution.

7 Summary and Outlook

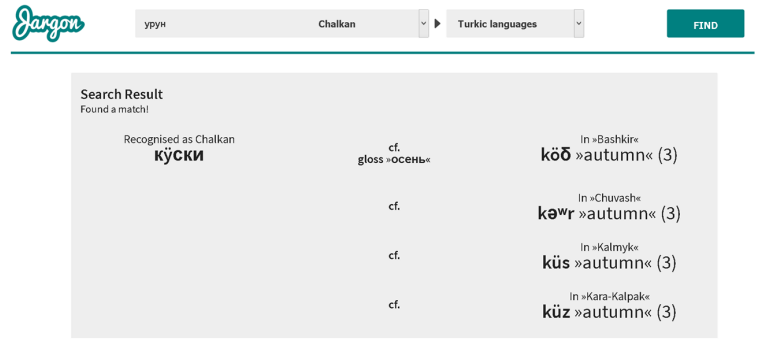
We described the application of Lemon-OntoLex to two dictionaries compiled in the context of language documentation efforts, an area where the appli-

[4] http://dbserver.acoli.cs.uni-frankfurt.de:5000/

[5] At the moment, this is done on the fly, with greater amounts of data in the system, optimizations will become necessary.

[6] Note that this query has been slightly simplified with respect to prefix declarations, matching typed and untyped strings, and different Lemon namespaces.

[7] `https://www.w3.org/2016/05/ontolex/`.




```
SELECT DISTINCT ?ru ?en_gloss ?trk
WHERE
{
 GRAPH liodi:chalkan {
   ?ru ^skos:definition/^lemon:sense/lemon:canonicalForm/lemon:writtenRep "кӱски"@atv.
 }
 GRAPH dbnary:rus {
   ?t dbnary:isTranslationOf/lemon:canonicalForm/lemon:writtenRep ?ru.
   ?t dbnary:targetLanguage lexvo:eng.
   ?t dbnary:writtenForm ?en_gloss.
}
GRAPH liodi:starling {
   ?en ^skos:definition/(^lemon:reference)?/^lemon:sense/lemon:canonicalForm/lemon:writtenRep ?trk.
 }
}
```

Fig. 5. Transitive query for Chalkan *küski* via Russian-English DBnary to the Starling Turkic etymological dictionary, (a) Workbench visualization, (b) SPARQL query (slightly simplified)

cation of Lemon-OntoLex has rarely been discussed before. We focus on exemplary resources for two less-resourced languages from the Turkic language family, Chalkan and Tuvan. Both datasets have been converted into a Linked Data representation using the Lemon-OntoLex data model, with an extensible converter written in Python. Finally, their application within a comparative-lexicographical workbench has been described, where Linked Data permits to formulate transitive queries over dictionaries from entire language families.

A proof-of-principle implementation of this workbench is currently available (`http://dbserver.acoli.cs.uni-frankfurt.de:5000/search/?query=%D0%B0%D1%80%D1%8B&originLang=&targetLang=trk`). Using the Chalkan-Russian data described in this paper, the Russian-English DBnary[8] and the English-Turkic etymological dictionary described in [1], it performs a transitive search for cognate candidates across two pivot languages (Russian and English): Lemon-based transitive sense links yield semantically corresponding forms, and the result set is ordered according to phonological (graphological) similarity with the requested word per sense. Top-level matches are thus most likely cognate candidates. For a limited number of small dictionaries with few thousand words as those created as part of language documentation efforts described here, our vanilla system is actually able to perform an effective online search without any further optimization. With more data being produces as part of the project, scalability issues are a likely area of future studies.

With the publication of this paper, the converter and the Chalkan data will be published under open licenses. For the Tuvan data, we are still waiting for legal clearance, but the original XML that serves as a basis for conversion, is, however, publicly available from The Language Archive[9]. With the converter provided, its Linked Data edition can be locally recreated.

Acknowledgments

The research described in this paper was conducted in the project 'Linked Open Dictionaries (LiODi, 2015-2020)', funded by the German Ministry for Education and Research (BMBF) as an Early Career Research Group on eHumanities.

[8] `http://kaiko.getalp.org/sparql`.
[9] `https://tla.mpi.nl/relish/`.

Bibliography

[1] Frank Abromeit and Christian Fäth. Linking the tower of babel: Modelling a massive set of etymological dictionaries as rdf. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 11, 2016.

[2] Frank Abromeit, Christian Chiarcos, Christian Fäth, and Maxim Ionov. Linking the Tower of Babel: Modelling a massive set of etymological dictionaries as RDF. In John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, and Sebastian Hellmann, editors, *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, pages 11–19, Portoroz, Slovenia, 2016.

[3] Simon Ager. Tuvan, 2016.

[4] G.D.S. Anderson. *Auxiliary Verb Constructions in Altai-Sayan Turkic*. Turcologica Series. Harrassowitz, 2004. ISBN 9783447046367.

[5] Helen Aristar-Dry, Sebastian Drude, Menzo Windhouwer, Jost Gippert, and Irina Nevskaya. „rendering endangered lexicons interoperable through standards harmonization": the relish project. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 766–770. European Language Resources Association (ELRA), 2012.

[6] Scott Farrar and William D Lewis. The gold community of practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, 41(1):45–60, 2007.

[7] Robert Forkel. The Cross-Linguistic Linked Data project. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 60–66, Reykjavik, Iceland, May 2014.

[8] William D Lewis and Fei Xia. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319, 2010.

[9] John McCrae, Philipp Cimiano, and Paul Buitelaar. Lexinfo ontology 2.0, 2010.

[10] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system namespace document - html variant, 2009.

[11] Irina Nevskaya. *Locational and directional relations and tense and aspect marking in Chalkan, a South Siberian Turkic language*, pages 67–75. Studies in Lanuage Companion Series. John Benjamins Publishing Company, 2014. ISBN 9789027269362.

[12] Jacquelijn Ringersma and Marc Kemps-Snijders. Creating multimedia dictionaries of endangered languages using lexus. In *Interspeech 2007: 8th Annual conference on the International Speech Communication Association*, pages 65–68. ISCA-Int. Speech Communication Assoc, 2007.

[13] Andrea C Schalley. Tyto – A collaborative research tool for linked linguistic data. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann,

editors, *Linked Data in Linguistics*, pages 139–149. Springer, Heidelberg, 2012.