Lemon-aid: using Lemon to aid quantitative historical linguistic analysis

Steven Moran
University of Zurich
University of Marburg
steven.moran@uzh.ch

Martin Brümmer
University of Leipzig
AKSW
bruemmer@informatik.uni-leipzig.de

Abstract

In this short paper, we describe how we converted dictionary and wordlist data made available by the QuantHistLing project into the Lexicon Model for Ontologies. By doing so, we leverage Linked Data to combine disparate lexical resources – more than fifty lexicons and dictionaries – by converting the lexical data into an RDF model that is specified by Lemon. The resulting new Linked Data resource, what we call the QHL dataset, provides researchers with a *translation graph*, which allows users to query across the underlying lexicons and dictionaries to extract semantically-aligned wordlists.

1 Introduction

There is an increasing amount of research that applies quantitative approaches to historical-comparative linguistic processes, including diverse areas such as: statistical tests for genealogical relatedness (Kessler, 2001), methods for phylogenetic reconstruction (Holman et al., 2011; Bouckaert et al., 2012), phonetic alignment algorithms (Kondrak, 2000; Prokić et al., 2009), and automatic detection of cognates (Turchin et al., 2010; Steiner et al., 2011), borrowings (Nelson-Sathi et al., 2011), and proto-forms (Bouchard-Côté et al., 2013). However, before any of these steps within the pipeline of computational historical linguistics can be undertaken, lexical data from secondary resources such as dictionaries and wordlists, or from tertiary resources like online lexical databases, must be collected, digitized and collated. The promise of the automatization of time-consuming tasks, such as lexical comparison, phonetic alignments and similarity judgements, is providing a resurgence of historical-comparative analysis, the goal of which is to identify the genealogical relatedness of languages and ultimately inform the prehistory of native peoples and their migrations. By linking data on these low-resource languages to the Linguistic Linked Open Data cloud (LLOD), and thus to the Linked Open Data cloud (LOD), we are also following in the practice and vision of the Semantic Web – open data sharing.

In the following sections we describe the QHL project's lexicon and wordlist format and how we converted the data into our ontological model specified in Lemon (McCrae et al., 2010; McCrae et al., 2011). The resulting resource allows users to query across what are originally disparate paper lexicons and dictionaries to extract semantically-aligned wordlists for historical-comparative analysis. We provide some examples in SPARQL.

2 Data

2.1 Source

The Quantitative Historical Linguistics (QuantHistLing) research unit aims to uncover and clarify phylogenetic relationships between native South American languages using quantitative methods.[1] There are two main objectives of the project: digitalization of lexical resources on South American languages and the development of computer-assisted methods and algorithms to quantitatively analyze the digitized data. The project aims to digitize around 500 works, most of which are currently only available in print and many of which are the only resources available for the languages that they describe. The list of the languages, language families and the data that has so far been digitized is available online.[2]

The QuantHistLing project has a simple data output format that contains metadata (prefixed with "@") and tab-delimited lexical out-

[1] http://quanthistling.info/

[2] http://quanthistling.info/index.php?id=resources

put. An example is given in Figure 1. The first row following the metadata contains the data header with the fields: QLCID, HEAD, HEAD_DOCULECT, TRANSLATION, TRANSLATION_DOCULECT, which correspond respectively to the internal QLC unique identifier, the headword in the dictionary, the *doculect* of the headword (or in other words the language in which this particular document describes), the translation for the given headword, and the doculect that the translation is given in. For each resource a data dump with the same format is provided by the project.

2.2 Conversion

We convert the QLC data into Linked Data that conforms to the Lemon model with a simple Python script. Lemon is an ontological model for modeling lexicons and machine-readable dictionaries for linking to the Semantic Web and the Linked Data cloud.[3] It is based on the Lexical Markup Framework (LMF) (Francopoulo et al., 2006) and uses the idea of data categories (Romary, 2010), like ISOCat (Kemps-Snijders et al., 2008), which include uniquely identified concepts that are useful for computational tasks (McCrae et al., 2011).

The benefits of modeling lexical data in Lemon are multi-fold. Internal to the Lemon mission are the benefits from overcoming the challenges that the model was designed to meet:[4]

- RDF-native form to enable leverage of existing Semantic Web technologies (SPARQL, OWL, RIF etc.).
- Linguistically sound structure based on LMF to enable conversion to existing offline formats.
- Separation of the lexicon and ontology layers, to ensure compatability with existing OWL models.
- Linking to data categories, in order to allow for arbitrarily complex linguistic description. In particular the LexInfo vocabulary is aligned to Lemon and ISOcat.
- A small model using the principle of least power - the less expressive the language, the more reusable the data.

We chose to model lexicons in Lemon instead of the Graph Annotation Format (GrAF) (Ide and Suderman, 2007) and the Lexicon Interchange FormaT (LIFT)[5] because of Lemon's tight integration with Semantic Web technologies, which allows us to add lexical data to the Linked Open Data cloud (LOD) and the Linguistic Linked Open Data cloud (LLOD). From the perspective of linguistics researchers, mapping dictionary and wordlists data to the LLOD has many advantages:

- Data that is linked is available on the Web in a standard format and accessible via the (L)LOD.
- Data are queryable through a SPARQL endpoint.
- The use of an ontology and Linked Data addresses the problem of merging disparate dictionary entries using senses and meaning mappings, including leveraging other sources such as Wordnet and domain-specific ontologies.

2.3 Ontology

Figure 2 illustrates our model implementation of the Lemon model with the QHL data.[6] Subjects, predicates and objects are clearly labeled. Currently the dataset contains 3,828,420 triples and we have made links to Lexvo,[7] a pivot for linguistic resources in the LLOD, via ISO 639-3 language name identifiers (de Melo, Submittied). There are currently 216 language links to Lexvo and thus numerous entries to other language resources.

3 Application

A major goal in historical-comparative linguistics is the identification of cognates, i.e. sets of words in genealogically related languages that have been derived from a common word or root (e.g. English 'is', German 'ist', Latin 'est', from Indo-European 'esti'). Modeling dictionaries and lexicons in a pivot ontology using overlaps in translations is

[3] The Lemon developers are also active in the W3C Ontology-Lexica Community Group, whose goal is to "develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies". See: http://www.w3.org/community/ontolex/.

[4] http://lemon-model.net/

[5] https://code.google.com/p/lift-standard/

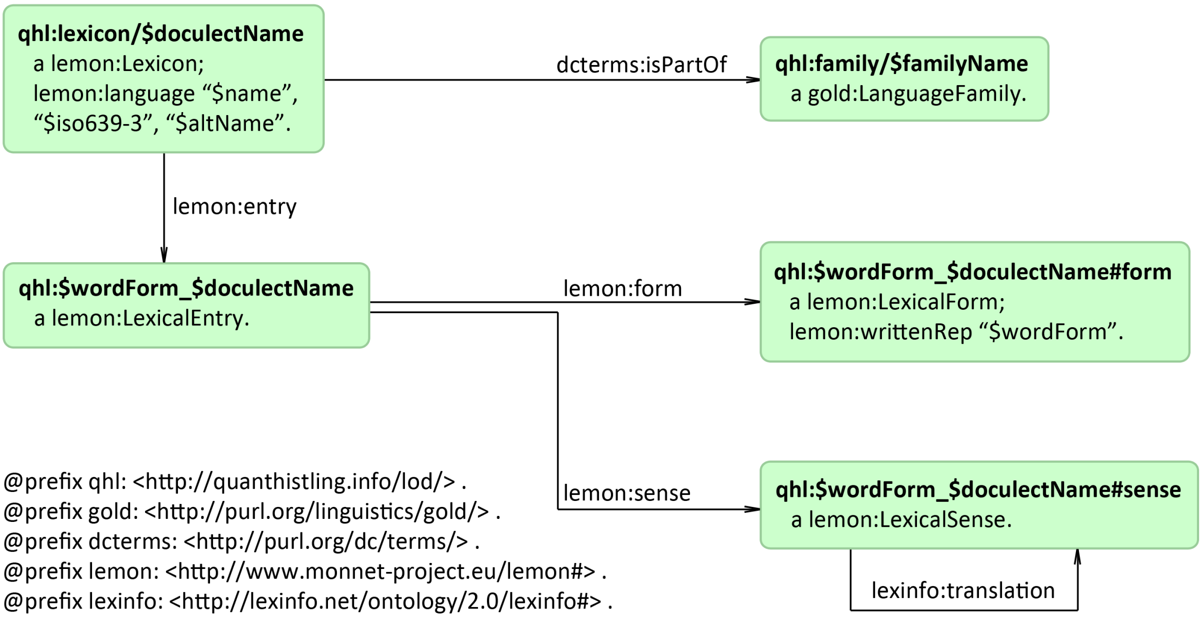
[6] Our version of the Linked Data is available here: http://linked-data.org/datasets/qhl.ttl.zip.

[7] http://www.lexvo.org/

Figure 1: QLC data format

```
@date: 2012-11-23
@url: http://www.quanthistling.info/data/source/aguiar1994/dictionary-329-369.html
@source_title: Analise descritiva e teorica do Katukino-Pano
@source_author: de Aguiar, Maria Sueli
@source_year: 1994
@doculect: Katukina, n/a, Katukina, Panoan
@doculect: Portugues, por, Portugues, Panoan
QLCID HEAD HEAD_DOCULECT TRANSLATION TRANSLATION_DOCULECT
aguiar1994/329/1 ai Katukina presente Portugues
aguiar1994/329/2 aima Katukina solteiro Portugues
aguiar1994/329/3 ain Katukina esposa Portugues
aguiar1994/329/4 ainnan Katukina cipo para cesta Portugues
aguiar1994/329/5 ainnan Katukina casado Portugues
aguiar1994/329/6 aka Katukina soco Portugues
aguiar1994/329/7 akaai Katukina tomar Portugues
```

Figure 2: Implementation of QHL data in Lemon



qhl:lexicon/$doculectName
a lemon:Lexicon;
lemon:language “$name”,
“$iso639-3”, “$altName”.
dcterms:isPartOf
qhl:family/$familyName
a gold:LanguageFamily.
lemon:entry
qhl:$wordForm_$doculectName
a lemon:LexicalEntry.
lemon:form
qhl:$wordForm_$doculectName#form
a lemon:LexicalForm;
lemon:writtenRep “$wordForm”.
lemon:sense
qhl:$wordForm_$doculectName#sense
a lemon:LexicalSense.
lexinfo:translation
@prefix qhl: <http://quanthistling.info/lod/> .
@prefix gold: <http://purl.org/linguistics/gold/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#> .

one way to merge several resources into one RDF graph for querying and extracting semantically-aligned wordlists, which can then be used as input into computational historical linguistics tools such as LingPy (List and Moran, 2013).[8]

As a first step, we have converted the QHL data into Linked Data and it is available online through a SPARQL endpoint.[9] Querying the combined dictionaries and lexicons is straightforward, as shown in example 1, which returns us all triples.

(1)
```
select * where
{GRAPH
<http://quanthistling.info/lod/>
{?s ?p ?o}
}
```

Next we limit the query in example 2 to the set of languages in our translation graph that contain written forms for the lexical sense "casa". The query returns pairs of words, but one can programmatically expand it by using the *wordForm2* and inserting it in the filter clause.

(2)
```
PREFIX lemon:
<http://www.monnet-project.eu/lemon#>
PREFIX lexinfo:
<http://lexinfo.net/ontology/2.0/
lexinfo#>
select ?wordForm1 ?language1
?wordForm2 ?language2 where
{GRAPH
<http://quanthistling.info/lod/>
{
?word1 a lemon:LexicalForm;
lemon:writtenRep ?wordForm1.
?entry1 lemon:form ?word1;
lemon:sense ?sense1.
?language1 lemon:entry ?entry1.
?sense1 lexinfo:translation ?sense2.
?word2 a lemon:LexicalForm;
lemon:writtenRep ?wordForm2.
?entry2 lemon:form ?word2;
lemon:sense ?sense2.
?language2 lemon:entry ?entry2.
FILTER(str(?wordForm1)="casa")
}
}
```

Regarding our use of *sense*, the Lemon documentation states: "The sense object represents a mapping between a lexical entry and an ontology entity." The "ontology entity" that the Lemon authors use as an example is a link to the corresponding DBpedia or Wiktionary entry, where a description of the meaning can be found. While the principle is sound, this information is not contained in our data. Hence that is why there is no more information in our #sense resources. If a reference to an ontology entry is to be added later, it can be easily done so by adding it as a property of the #sense resource (for example as owl:sameAs, dcterms:references, etc.). However, if we have only strings in languages that are very rare, how are we to add an ontology entry? For most of the entries, there will be no corresponding entry. In fact, suppose we find the translation of an entry in a poorly documented language into a richer-resourced language (e.g. Katukina to Portuguese), we would not know if the Portuguese sense is a proper description of the sense of the work in Katukina. Moreover, the links would be sparse and some, if not many, would be wrong due to missing information. Therefore, our modelling follows the Lemon cookbook (examples 29, page 18) for good reason: the translation of a word is neither a translation of its wordform or representation nor is it a translation of its lexical entry. It is thus linguistically sound to say the "sense" of a word like "casa" is translated into another language, but its word form or entry is not.

Building on the former query, one can also add a node, as illustrated in example 3:[10]

(3)
```
PREFIX lemon:
<http://www.monnet-project.eu/lemon#>
PREFIX lexinfo:
<http://lexinfo.net/ontology/2.0/
lexinfo#>
select ?wordForm1 ?language1
?wordForm2 ?language2 ?wordForm3
?language3
WHERE
{GRAPH
<http://quanthistling.info/lod/>
{?word1 a lemon:LexicalForm;
lemon:writtenRep ?wordForm1.
?entry1 lemon:form ?word1;
lemon:sense ?sense1.
?language1 lemon:entry ?entry1.
?sense1 lexinfo:translation ?sense2.
?word2 a lemon:LexicalForm;
lemon:writtenRep ?wordForm2.
?entry2 lemon:form ?word2;
lemon:sense ?sense2.
?language2 lemon:entry ?entry2.
?sense2 lexinfo:translation ?sense3.
?word3 a lemon:LexicalForm;
lemon:writtenRep ?wordForm3.
?entry3 lemon:form ?word3;
lemon:sense ?sense3.
?language3 lemon:entry ?entry3.
FILTER (str(?wordForm1)="casa")
}
}
```

Of course this query can be easily extended to in-

[8] http://lingpy.org/

[9] http://qhanthistling.info/lod/

[10] Note that the filter in this query is computationally expensive and at the moment certain queries may time out as we try and increase server capacity.

corporate entire wordlists, such as the Swadesh list (Swadesh, 1952) or Leipzig-Jakarta list (Tadmor et al., 2010).

Again we emphasize that the combination of disparate data from many dictionaries and lexicons is a first step in a computational historical linguistics pipeline: the results are given in the source documents' orthographic representations and therefore they must be normalized into an interlingual pivot, such as the International Phonetic Alphabet, if phonetic or phonemic analysis is to be applied to the data. This would be the next step before producing phonetic alignments and cognate judgements based on metrics and algorithms for calculating lexical similarity.

4 Conclusion

From data being digitized and extracted from print resources, we are creating machine-readable lexicons that are both interoperable with each other (we link semantic senses using the Lemon ontology model) and with other linguistics sources (we use standard language code URIs used by other Linked Data resources in the LLOD).

Future work may proceed in a number of directions, such as:

- building algorithms that identify semantically similar translation-pairs from terse translations, e.g. identify that doculect translations like "coarsely grind", "grind up, crush well", "grind lightly (chili pepper, millet for a quick snack)", "grind lightly (groundnuts) with stones" for different languages can be mapped to a simpler form such as "to crush/grind" for initial comparative analysis

- using NLP Interchange Format (Hellmann et al., 2012) to keep track of where information in the dictionaries comes from – or in other words, use NIF combined with Lemon to annotate the QHL data sources for provenance

- linking to other resources that contain other linguistic and non-linguistic information (e.g. typological data and geographic variables that provide useful information for determining the genealogical and geographical relatedness of languages)

References

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *PNAS*, 110(11):4224–4229.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, Aug.

Gerard de Melo. Submittied. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web Journal*.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria, et al. 2006. Lexical markup framework (lmf). In *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Sebastian Hellmann, Jens Lehmann, and Sören Auer. 2012. Linked-data aware uri schemes for referencing text fragments. In *Knowledge Engineering and Knowledge Management*, pages 175–184. Springer.

Eric. W. Holman, Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dimitry Egorov. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Nancy Ide and Keith Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8. Association for Computational Linguistics.

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. Isocat: Corralling data categories in the wild. In *LREC*.

Brett Kessler. 2001. *The significance of word lists*. CSLI Publications, Stanford.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 288–295, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johann-Mattis List and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the Association for Computational Linguistics*.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez Pérez, Jorge Gracia, Laura

Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2010. The lemon cookbook. Technical report, CITEC, Universität Bielefeld.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer.

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B*, 278(1713):1794–1803.

Jelena Prokić, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 18–25, Stroudsburg, PA. Association for Computational Linguistics.

Laurent Romary. 2010. Standardization of the formal representation of lexical information for nlp. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*.

Lydia Steiner, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Uri Tadmor, Martin Haspelmath, and Bradley Taylor. 2010. Borrowability and the notion of basic vocabulary. *Diachronica*, 27(2):226–246.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, 3:117–126.