Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), pages 11–14
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon

Francisco Mondaca[1], **Felix Rau**[2]
[1]Cologne Center for eHumanities - University of Cologne
[2]Data Center for the Humanities - University of Cologne
Albertus-Magnus-Platz, 50923 Cologne, Germany
{f.mondaca, f.rau}@uni-koeln.de

Abstract

The Cologne Digital Sanskrit Dictionaries (CDSD) is a large collection of complex digitized Sanskrit dictionaries, consisting of over thirty-five works, and is the most prominent collection of Sanskrit dictionaries worldwide. In this paper we evaluate two methods for transforming the CDSD into Ontolex-Lemon based on a modelling exercise. The first method that we evaluate consists of applying RDFa to the existent TEI-P5 files. The second method consists of transforming the TEI-encoded dictionaries into new files containing RDF triples modelled in OntoLex-Lemon. As a result of the modelling exercise we choose the second method: to transform TEI-encoded lexical data into Ontolex-Lemon by creating new files containing exclusively RDF triples.

Keywords: tei, ontolex-lemon, lexicog, rdfa, sanskrit

1. Sanskrit Lexicography

Sanskrit (ISO 639-3 san) is a classical language from South Asia. It is the liturgical language of Hinduism and some branches of Buddhism and was the literary and scientific language of South Asia well into modern times. As a consequence, Sanskrit has a 4000 year long history.

Sanskrit belongs to the Indo-Aryan branch of the Indo-European language family and it is the only attested form of Old Indo-Aryan. Based on internal diachronic developments, it is conventionally divided into Vedic Sanskrit (early Old Indo-Aryan, 2000 BCE–600 BCE) and Classical Sanskrit (later Old Indo-Aryan) after the Vedic period (Masica, 1991).

As the oldest attested form of Indo-Aryan, Sanskrit constitutes one of the oldest attested Indo-European laguages and is central to our understanding of this language family. From the 19th century onward, philological research produced a vast array of Sanskrit text editions, grammatical descriptions, and dictionaries. In particular, the Große Petersburger Wörterbuch (Böhtlingk and Roth, 1855) and Monier-Williams' Sanskrit-English Dictionary, (Monier-Williams, 1899) are among the most important bilingual lexicographical works of the 19th century, if not in general.

The corpus of scientific Sanskrit dictionaries consist of over thirty-five works and includes mono- and bilingual general dictionaries as well as more specialised thematic works. Some of these work, such as Grassmann's dictionary (Grassmann, 1873), are specific to one text. Others are covering only one specific variety of Sanskrit. For example, the Buddhist Hybrid Sanskrit Dictionary (Edgerton, 1953) covers the distinct variety of Sanskrit used in some early schools of Buddhism (Burrow, 2001). Other dictionaries aiming at covering the whole lexical range of 4000 years of language history, resulting in lexicographical challenges modern Sanskrit lexicography still has to address (Lugli, 2018).

The entries in the more extensive dictionaries – in particular in the Große Petersburger Wörterbuch (Böhtlingk and Roth, 1855) and Monier-Williams' Sanskrit-English Dictionary, (Monier-Williams, 1899) – are highly structured and complex. These lexicographical microstructures pose a challenge to all attempts of developing general schemes and vocabularies for entry structures and are good test cases for whether a model can cover complex bilingual, multi-writing system entries.

2. About the Cologne Digital Sanskrit Dictionaries (CDSD)

2.1. Overview

The Cologne Digital Sanskrit Dictionaries[1] is the most prominent collection of digitized Sanskrit dictionaries available on the Internet. This project was initiated in 1994 when XML did not exist yet and Sanskrit had no proper Unicode support. Sanskrit is traditionally written in a variety of local scripts, but is now generally printed in Devanagari, a the North Indian script that is most prominently used to write Hindi and Nepali. While support for Devanagari was already included in Unicode 1.0.0 in 1993, full coverage of the characters needed to encode Sanskrit texts and lexicographic resources was only achieved in 2009 when the Vedic Extensions were added with Version 5.2. to the Unicode standard. As a consequence, an ASCII-based encoding scheme was developed in 1994 specifically for this collection, in order to encode strings in Devanagari, or to encode its Roman script transliteration (later standardized as ISO-15919). In 2003, when XML and Unicode were already part of the technologies available for serializing language resources, the CDSD offered different web applications for accessing its dictionaries. The CDSD collection consists of more than thirty-five Sanskrit dictionaries, mostly bilingual dictionaries covering different modern European languages. The CDSD web portal offers different web applications to access each dictionary. Also from the CDSD web portal each dictionary can be downloaded. The

[1] `https://www.sanskrit-lexicon.uni-koeln.de`

dictionaries can be accessed on GitHub[2] in their source format. Their XML-encoded versions can also be accessed via web APIs[3] provided by the API framework Kosh [4].

2.2. Searching for Sustainability and Interoperability

During the LAZARUS project (2013-2015)[5] in order to provide a sustainable and interoperable format for the CDSD collection, a common TEI-P5[6] schema[7] was developed. Three dictionaries were transformed into TEI-P5: the two most complex dictionaries both from a content and a layout perspective (Monier-Williams, 1899; Böhtlingk and Roth, 1855) and one English-Sanskrit dictionary (Apte, 1884). During the VedaWeb project (2017-2020)[8], four dictionaries of the CDSD collection (Apte, 1890; Edgerton, 1953; Grassmann, 1873; Macdonell and Keith, 1912) have been transformed into TEI-P5 employing the schema developed during the LAZARUS project. VedaWeb offers a digital edition of the Rigveda, the most ancient Indo-Aryan text. VedaWeb is an API-driven project. On the one hand, the project offers its textual data through a REST API [9]. On the other hand, the project offers its lexical resources via REST and GraphQL APIs[10]. One of the main features of VedaWeb is that each token of the Rigveda points to an entry in Grassmann's dictionary (Grassmann, 1873). This Sanskrit-German dictionary has been specially compiled for the Rigveda with the goal of defining every token present on it. The VedaWeb app calls the Grassmann GraphQL API and displays its respective information.

3. Transforming TEI-encoded dictionaries into Ontolex-Lemon

3.1. Where to start?

Taking into account the experiences gained during the projects LAZARUS and VedaWeb, we decided to begin the transformation of the available TEI-P5 dictionaries into OntoLex-Lemon (McCrae et al., 2017) with the most complex Sanskrit-English dictionary: Monier-Williams (Monier-Williams, 1899). This Sanskrit-English dictionary is considered to be the most detailed Sanskrit dictionary compiled in the English language. It is also a constant reference for Sanskrit scholars. For these reasons it was chosen to be the basis for creating a TEI-schema that would be applied to other dictionaries of the collection. And it will be the basis of this new transformation scenario.

3.2. Existing transformation methods

There are two main approaches for transforming TEI-encoded data into a Ontolex-Lemon compliant version. The first approach consists in extracting the lexical data contained in the TEI file and create a new file with this data modelled in Ontolex-Lemon. This method has been applied previously to the *Dictionnaire étymologique de l'ancien français* (Tittel and Chiarcos, 2018). Tittel and Chiarcos employ XSLT Stylesheets for the transformation. The same technology is applied at the tei2ontolex GitHub repository[11] developed by the European Lexicographic Infrastructure Project (ELEXIS) (Declerck et al., 2019), where researchers John P. McCrae and Laurent Romary, experts in Ontolex-Lemon and TEI respectively, have worked together.

To create new files modelled in Ontolex-Lemon would be less verbose, because it would leave the TEI file with its tags and attributes as it was originally encoded. But this method would also duplicate the amount of files containing lexical data to be curated. It would also require to synchronize both the TEI and Ontolex-Lemon serialized versions.

A second approach consists in employing RDFa within the source TEI file (Chiarcos and Ionov, 2019), i.e. modelling TEI and Ontolex-Lemon within the same file. This method would simplify at first sight the task of encoding the existent lexical data with Ontolex-Lemon. However, an issue to consider when applying RDFa within TEI files is that while this method is W3C-compliant it is not TEI-endorsed (Chiarcos and Ionov, 2019).

Another issue that arises when modelling digitized dictionaries with Ontolex-Lemon derives from the constraint that it allows a single part-of-speech (POS) per entry. In this regard TEI-P5 and TEI-Lex0 [12] are flexible because they allow multiple POS per lexical entry. Consequently, the structure of a printed dictionary is respected in TEI. On the contrary, in OntoLex-Lemon the lexicographic structure must be split when an entry contains more than one POS. This is a negative aspect when modelling and encoding digitized dictionaries because it makes a transformation scenario more complex and verbose than necessary.

This issue is addressed by the Ontolex-Lemon Lexicography Module (lexicog) [13]. Lexicog does respect the structure of a digitized dictionary. However, it achieves this basically from a layout perspective. It does create a parallel structure linked to Ontolex-Lemon core's module where the lexicographic information, e.g. POS, is encoded. Complexity and verbosity are thus not reduced.

The terms complexity and verbosity are here employed from the perspective of a human that reads a file and seeks to elucidate the model behind it. While XML is the most verbose of all RDF serializations (Cimiano et al., 2020), the same applies to Turtle or other RDF serializations when modelling digitized dictionaries in Ontolex-Lemon. In our opinion the complexity and verbosity that emerges when modelling digitized dictionaries in Ontolex-Lemon has its

[2] https://github.com/sanskrit-lexicon/csl-orig/tree/master/v02

[3] https://cceh.github.io/kosh/docs/implementations/cdsd.html

[4] https://kosh.uni-koeln.de

[5] https://cceh.uni-koeln.de/lazarus

[6] https://tei-c.org/guidelines/p5

[7] https://github.com/cceh/c-salt_dicts_schema

[8] https://vedaweb.uni-koeln.de

[9] https://vedaweb.uni-koeln.de/rigveda/swagger-ui.html

[10] https://cceh.github.io/c-salt_sanskrit_data

[11] https://github.com/elexis-eu/tei2ontolex

[12] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[13] https://www.w3.org/2019/09/lexicog

origin in establishing that an entry can have only a single part-of-speech.

3.3. Modelling Ontolex-Lemon with RDFa

As seen in the previous section the two presented methods for converting TEI-encoded data into Ontolex-Lemon consist in creating new files containing exclusively RDF triples or in applying RDFa within the same TEI file.

Figure 2 shows an entry in Monier Williams that has been partially modelled in RDFa using Ontolex-Lemon. Figure 1 shows the entry as it has been modelled in TEI-P5. The first problem that arises in figure 2 is that morphosyntactic information encoded as part-of-speech (POS) in Ontolex-Lemon can not be related to an `ontolex:LexicalSense`. This must be related to an `ontolex:LexicalEntry`. A possible solution for applying RDFa would be to modify the existent XML-TEI structure, i.e. create new XML nodes.

Another issue when applying RDFa and Ontolex-Lemon relates to choose an external ontology or vocabulary for encoding POS. When encoding lexical data with Ontolex-Lemon, usually the lexinfo[14] ontology is employed. In figure 2, the POS 'mfn.' means 'masculine, feminine and neutral'. There is no such category in `lexinfo:partOfSpeech` for this POS. A possible solution would be to create a vocabulary containing the POS to be found in all the Sanskrit dictionaries of the collection and later map the values of this vocabulary to existing values of `lexinfo:partOfSpeech`.

4. Conclusion

At first sight RDFa seemed to tackle our requirements better than creating new files containing RDF triples when transforming TEI source files into Ontolex-Lemon. But a brief modelling exercise showed that employing RDFa required adding new elements into the TEI source files. These structural modifications to the TEI files would unnecessarily complicate the maintenance of these files over time. Therefore, the method to follow will be to create completely new files modelled in Ontolex-Lemon. To this end, we will follow the experiences made during the transformation of the *Dictionnaire étymologique de l'ancien français* (Tittel and Chiarcos, 2018) into Ontolex-Lemon, as well as the current development of the tei2ontolex[15] repository.

5. Bibliographical References

Apte, V. S. (1884). *The Student's English-Sanskrit Dictionary*. Arya Bhushana, Poona.

Apte, V. S. (1890). *The Practical Sanskrit-English Dictionary*. Shiralkar, Poona.

Burrow, T. (2001). *The Sanskrit Language*. Motilal Banarsidass Publ.

Böhtlingk, O. and Roth, R. (1855). *Sanskrit Wörterbuch. Herausgegeben von der kaiserlichen Akademie der Wissenschaften, bearbeitet von Otto Böhtlingk und Rudolph Roth*. Eggers, St-Petersburg.

Chiarcos, C. and Ionov, M. (2019). Linking the TEI. Approaches, Limitations, Use Cases. Digital Humanities 2019, July.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham, Switzerland.

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K., and Wissik, T. (2019). ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data.

Edgerton, F. (1953). *Buddhist Hybrid Sanskrit Grammar and Dictionary*. Yale Univ. Press, New Haven.

Grassmann, H. G. (1873). *Worterbuch zum Rig-veda*. O. Harrassowitz, Wiesbaden.

Lugli, L. (2018). Drifting in Timeless Polysemy: Problems of Chronology in Sanskrit Lexicography. *Dictionaries: Journal of the Dictionary Society of North America*, 39(1):105–129, August.

Macdonell, A. A. and Keith, A. B. (1912). *Vedic Index of Names and Subjects*. J. Murray, London.

Masica, C. (1991). *The Indo-Aryan Languages*. Cambridge University Press, Cambridge.

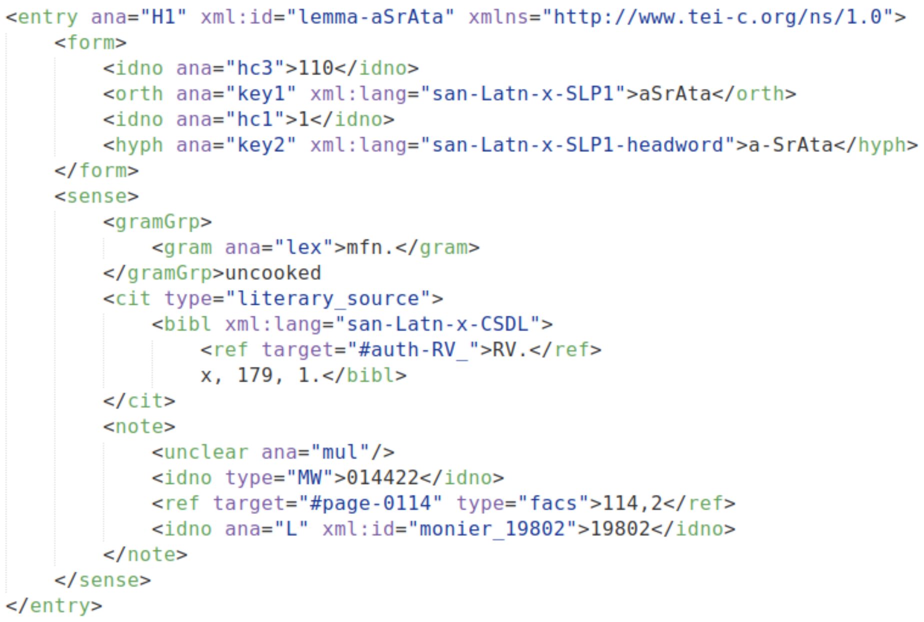
McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*.

Monier-Williams, M. (1899). *A Sanskrit-English dictionary*. The Clarendon Press, Oxford.

Tittel, S. and Chiarcos, C. (2018). Linked open data for the historical lexicography of old french. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

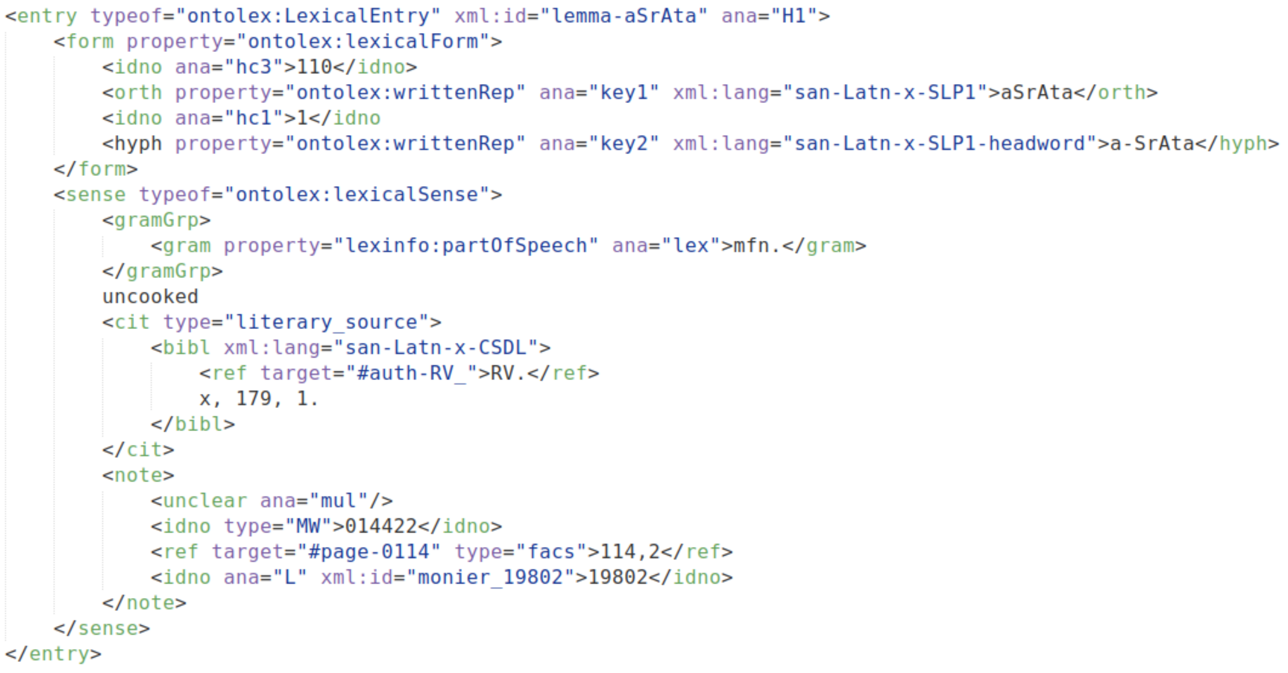
[14] `https://lexinfo.net`

[15] `https://github.com/elexis-eu/tei2ontolex`



```
<entry ana="H1" xml:id="lemma-aSrAta" xmlns="http://www.tei-c.org/ns/1.0">
    <form>
        <idno ana="hc3">110</idno>
        <orth ana="key1" xml:lang="san-Latn-x-SLP1">aSrAta</orth>
        <idno ana="hc1">1</idno>
        <hyph ana="key2" xml:lang="san-Latn-x-SLP1-headword">a-SrAta</hyph>
    </form>
    <sense>
        <gramGrp>
            <gram ana="lex">mfn.</gram>
        </gramGrp>uncooked
        <cit type="literary_source">
            <bibl xml:lang="san-Latn-x-CSDL">
                <ref target="#auth-RV_">RV.</ref>
                x, 179, 1.</bibl>
        </cit>
        <note>
            <unclear ana="mul"/>
            <idno type="MW">014422</idno>
            <ref target="#page-0114" type="facs">114,2</ref>
            <idno ana="L" xml:id="monier_19802">19802</idno>
        </note>
    </sense>
</entry>
```

Figure 1: Entry ’aSrata’ in Monier Williams modelled in TEI-P5



```
<entry typeof="ontolex:LexicalEntry" xml:id="lemma-aSrAta" ana="H1">
    <form property="ontolex:lexicalForm">
        <idno ana="hc3">110</idno>
        <orth property="ontolex:writtenRep" ana="key1" xml:lang="san-Latn-x-SLP1">aSrAta</orth>
        <idno ana="hc1">1</idno
        <hyph property="ontolex:writtenRep" ana="key2" xml:lang="san-Latn-x-SLP1-headword">a-SrAta</hyph>
    </form>
    <sense typeof="ontolex:lexicalSense">
        <gramGrp>
            <gram property="lexinfo:partOfSpeech" ana="lex">mfn.</gram>
        </gramGrp>
        uncooked
        <cit type="literary_source">
            <bibl xml:lang="san-Latn-x-CSDL">
                <ref target="#auth-RV_">RV.</ref>
                x, 179, 1.
            </bibl>
        </cit>
        <note>
            <unclear ana="mul"/>
            <idno type="MW">014422</idno>
            <ref target="#page-0114" type="facs">114,2</ref>
            <idno ana="L" xml:id="monier_19802">19802</idno>
        </note>
    </sense>
</entry>
```

Figure 2: Entry ’aSrata’ in Monier Williams modelled in TEI-P5 and partially with errors in Ontolex-Lemon with RDFa