A Linguistic Knowledge Graph of Word Borrowings from Portuguese

Anas Fahad Khan[1], Ana Salgado[2]
[1] CNR-ILC, Italy - fahad.khan@ilc.cnr.it
[2] NOVA CLUNL - Centro de Linguística da Universidade Nova de Lisboa, Portugal, Portugal - anacastrosalgado@gmail.com

ABSTRACT (ENGLISH)

We present ongoing work in the creation of a linguistic knowledge graph of word borrowings from Portuguese into South Asian languages. The first version of this resource, soon to be published, consists of lexicons of Hindi and Urdu words deriving from Portuguese. We start by describing our approach and the data collection process. Then we describe how different aspects of the data are represented as linked data.
Keywords: Linked Open Data; Semantic Web; Ontologies; RDF

ABSTRACT (ITALIANO)

Un Grafo di Conoscenza Linguistica sui Prestiti Lessicali dal Portoghese

Presentiamo un lavoro in corso per la creazione di un linguistic knowledge graph sulle influenze lessicali del portoghese nelle lingue dell'Asia meridionale, una risorsa che stiamo per pubblicare in una prima versione costituita da lessici di parole in hindi e urdu derivanti dal portoghese. Iniziamo descrivendo il nostro approccio e il processo di raccolta dei dati. Successivamente, spieghiamo come diversi aspetti dei dati vengono rappresentati come linked data.
Parole chiave: Linked Open Data; web semantico; ontologie; RDF

1. INTRODUCTION

Despite its being over a decade old, and despite being successfully used in important projects like Linking Latin (LiLa) (Passarotti et. al. 2020), linguistic linked open data (LLOD) is still a relatively underdeveloped and/or underutilised means of publishing linguistic data. This is unfortunate as LLOD is very well adapted for publishing numerous different kinds of linguistic dataset thanks to its graph-like structure, something which ultimately allows for the creation of extensive linguistic knowledge graphs[1] as in the case of the aforementioned LiLa project; it also has the advantage of supporting the enrichment of linguistic resources with non-linguistic resources via RDF's native linking mechanism. LLOD can, moreover, facilitate the development of language resources which better exploit LOD's technological (and other) strengths in making data more accessible/machine actionable. In this article, we highlight one kind of linguistic resource that seems particularly adapted to publication as an RDF linguistic knowledge graph: a language resource describing lexical borrowings from a single language into a number of other languages. Note that although we describe such resources as *linguistic* knowledge graphs, the kinds of information they contain are useful to researchers in a number of other disciplines too, including historians, and scholars in cultural studies. In this article we will concentrate on a particular example/use-case of a language resource of borrowings, one which the authors of the paper have been working on over the last few months, and the first version of which we have made available as RDF via Github,[2] CHAMUÇA. In particular, we will look at the different ways that RDF and the Semantic Web facilitate or afford the structured representation of various, salient aspects of our data, aspects which are typical of such data.

The rest of the paper is organised as follows. We begin by giving an overview of the project in Section 2. This is followed by a description of the data sources of our linguistic knowledge graph in Section 3. In Section 4, we describe our linguistic knowledge graph and focus on the affordances offered by the Semantic Web for representing the contents of our data.

[1] By linguistic knowledge graph we mean a knowledge graph whose principal scope is to describe linguistic phenomena.

[2] https://github.com/anasfkhan81/chamuca/tree/main/chamuca_lex_resource

1.1. TRACING THE IMPACT OF PORTUGUESE ON SOUTH ASIAN LANGUAGES: THE CASE OF HINDI AND URDU

The work described here began with a simple question: *how many words in the two South Asian languages Urdu/Hindi derive from Portuguese?* Although Portuguese does not seem, prima facie, to have had a profound impact on both languages -- certainly nowhere near as great as English has had -- there are a number of very popular words (e.g., the most common words in both languages for 'key', 'room' and 'church', all three of which are shared by both languages) which ultimately derive from the language, and which seem to evidence the great influence which Portuguese traders and colonists had had in Asia at the beginning of the Modern era. Indeed the more one looks into this question, the more it becomes clear that, leaving aside the obvious case of English, no other European language has had as great a direct impact on Urdu/Hindi as Portuguese -- something which seems to be common for many other Asian, and especially South Asian languages too (notable exceptions being South-east Asian languages such as Vietnamese which have received greater influence from French and Tagalog which was greatly influenced by Spanish). This led the authors to ask whether there were any modern electronic resources, and in particular any dictionaries, which could be consulted on this topic, and, additionally, whether such resources existed for other (South) Asian languages, in order to compare the levels of influence which Portuguese had had on nearby languages. The answer is that, although much of the relevant information *was* available distributed across different sites/resources/print volumes, in most cases it was **not** available as structured data, and was often not completely curated. There exists one well known scholarly work on this topic, 1913's *Influência do vocabulário português em línguas asiáticas* ('Influence of the Portuguese Vocabulary on Asian Languages'), a lexicon compiled by the early 20th century Goan scholar Sebastião Rodolfo Dalgado and which features Portuguese borrowings in 50 languages. However, this was over 110 years old by now and hadn't been updated.

We decided to fill the existing gap by building our own language resource, **Cultural Heritage and Multilingual Understanding through lexiCal Archives (CHAMUÇA)**, an RDF linguistic knowledge graph consisting of lexicons of Portuguese borrowings in various different Asian languages (beginning with the languages of South Asia). CHAMUÇA will be published (in different versions) with an open license, and will be based, as far as possible, on already existing resources. As of the time of writing, we have compiled the Hindi and Urdu lexicons and an indexical Portuguese lexicon, these are available as RDF graphs on Github[3] and we are planning to publish them as linked data on our triple store, as one of the use-cases of the Italian national project, H2IOSC,[4] shortly. We chose these two languages because of the linguistic competences of the authors as well as their popularity as languages on the subcontinent, and the interesting status of the two in relation to one another (many linguists regard them as two different registers of the same language). However, we are in discussion with experts of languages such as Bengali, and Sinhala for the next updated version of our graph.

2. DATA SOURCES

We derived an initial list of over a hundred entries for Hindi and Urdu on the basis both of Dalgado's monumental study as well as from Wiktionary and Wikipedia; these entries were cross-checked and enriched with respect to a number of several different lexical sources (including those made available by the Digital Dictionaries of South Asia[5] platform). Each entry in our lexicons includes basic lexical information such as *lemma*, *gender, part of speech*, and other morphological variants of the lemma (nouns in Urdu/Hindi have 6 morphological variants, Portuguese has 2; however, in the rare cases of verbs/adjectives we have come across we haven't listed all their forms); there is also link from each entry to its supposed Portuguese etymon (as well as information on other potential etymologies, taken from different sources). Note that Dalgado uses his own Roman alphabet transliterations for all of the languages featured in his lexicon. In many cases, it was easy to find Hindi/Urdu lemmas in their Devanagari/Nastaliq (respectively) forms; in a handful of cases it was much less straightforward and

[3] https://github.com/anasfkhan81/chamuca

[4] https://www.h2iosc.cnr.it/

[5] https://dsal.uchicago.edu/dictionaries/

required consulting and comparing a number of lexicographic sources. One of the authors of the current work is chief editor for the dictionaries of the Academia das Ciências de Lisboa and was able to contribute with resources on Portuguese etymons. The result of this data collection phase was a fairly comprehensive list of around 100 Hindi lemmas, circa 80 Urdu lemmas and around 50 Portuguese lemmas with relevant lexical/linguistic information.

3. BUILDING A LINGUISTIC KNOWLEDGE GRAPH OF PORTUGUESE BORROWINGS

Our idea, to reiterate, was to bring together and curate already existing information, enrich it (especially with reference to more recent scholarship) and make it available as linked data, namely, as an RDF linguistic knowledge graph; and in terms of the latter, the more (suitable/relevant) links there are between the different nodes in our data, the better. In what follows, we look at how we converted our original data (stored as TSV files) into linked data and how we tried to take advantage of the different expressive and technological possibilities offered by the Semantic Web. The modelling of basic linguistic information via the OntoLex vocabulary[6] together with the lexinfo ontology[7] (the latter acting as a Data Category Registry for the former) was fairly straightforward.[8] As an example in Figure 1 we present part of an entry from the CHAMUÇA Hindi lexicon, namely the entry for अलमारी [almari] meaning 'cupboard', a word ultimately derived from the Portuguese *armario*, with some of this basic information shown.

```
:अलमारी_entry a ontolex:LexicalEntry;
      lexinfo:etymologicalRoot pt_lex:armário ;
      lexinfo:gender lexinfo:feminine ;
      lexinfo:partOfSpeech lexinfo:commonNoun ;
      ontolex:canonicalForm :अलमारी_lemma ;
      ontolex:lexicalForm :अलमारियो_vp_form_अलमारी,
         :अलमारियों_op_form_अलमारी,
         :अलमारी_dp_form_अलमारी,
         :अलमारी_os_form_अलमारी,
         :अलमारी_vs_form_अलमारी ;
```

Figure 1. Part of the RDF encoding for the entry अलमारी

Here, one can see reference to the gender, and part of speech of the word, a link to the resource representing the lemma and the word's other morphological variants together with a link to the sense of the word; this is the minimum standard information for all the words in our Portuguese, Hindi and Urdu lexicons. There is also a link to the proposed Portuguese etymon which belongs to the Portuguese lexicon *pt_lex* and a link to the corresponding Urdu entry via the RDFS *seeAlso* property.

Our data focuses on describing a network of relationships between words of different languages; the graph-based nature of RDF makes it a natural fit for this (relationships between words viewed as nodes). The LiLa project, cited above, was an important step in revealing the potential of linked data for creating rich linguistic knowledge graphs (LKG) on the basis of a specific conceptual architecture. In the case of LiLa this is based on the existence of a hub component of the LKG, *the lemma bank* (Mambrini and Passarotti, 2023). In our case we have a slightly different architecture. Our entire dataset is a Lexical Resource,[9] that is a container for a number of lexicons, each one of which contains borrowed words from

[6] OntoLex or OntoLex-Lemon is a popular RDF based lexicon for encoding lexical resources, https://www.w3.org/2016/05/ontolex/

[7] https://lexinfo.net/

[8] However, as we begin to look at other languages, especially non-Indo European ones, it may prove somewhat more challenging, since OntoLex has up till now been used to model a fairly limited selection of languages.

[9] We take the definition of Lexical Resource used in for instance LMF (Francopoulo, 2013).

Portuguese, and an indexical Portuguese lexicon which acts as a hub. We would like to propose this as a sort of basic pattern[10] for similar work in lexicons of borrowings or in language resources for contact linguistics. The use of these and similar patterns would offer an extra level of interoperability in addition to the use of the RDF data model and shared vocabularies and ontologies.

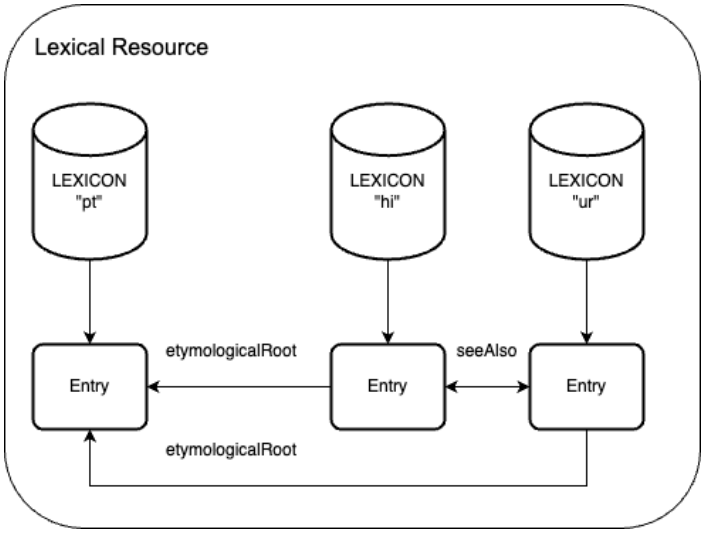

Figure 2. The overall architecture of our resource (we plan to expand this to add more languages)

We decided to focus on exploiting the semantic affordances made available by the Semantic Web to ensure that our LKG would be easier to navigate, and to enable it to answer a wider variety of questions. For instance, we use the popular Semantic Web knowledge organisation vocabulary SKOS[11] to encode a list of basic domains as concepts, and to annotate different entries in our CHAMUÇA lexicons for the domains with which they are associated (using the lexinfo *domain* property). We ended up with 24 domains (this number will undoubtedly expand as we include other languages): *administration, animal husbandry, architecture, botany, calendar, cleaning and hygiene, clothing, culinary, furniture, healthcare, household, materials, metrology, military, music, nationalities, nautics, occupations, proper names, religion, sensory perception, sewing, trade, transportation*. For instance, in our Hindi/Urdu lexicons we have several examples of entries which are related to the Christian religion, these include **गिरजा/**گِرْجا ('church'), **पादरी**پادری ('priest') and **क्रूस**/ ('cross'); each of these entries is associated with the domain *religion* (the latter represented as a SKOS concept). We followed the best practices suggested by (Khan et al, 2023) for adding domain labels to linked data lexicons here.

Finally, we have enriched our KG with etymological information. Quite a few of the words in our Hindi/Urdu lexicon have alternative etymologies (i.e., without Portuguese etymons), and in some cases the Portuguese origin of the word is quite doubtful. Etymologies are far from simple to represent in RDF, however, since each etymology is effectively an individual word history and there is no agreed upon best practice for how to represent even very simple narratives as RDF graphs. There have been proposals for encoding this in a 'deeper' way, in the sense of attempting to explicitly represent as much of the relevant information contained in an original textual version of the etymology as possible as structured data in RDF; we decided, however, that we would give the etymology as a string in a standard form. These etymologies, each of which is associated with an individual entry via the lexinfo:etymology property, are derived from a number of sources, including Dalgado and (McGregor, 1993), provenance information which we mark explicitly in our string etymologies. For instance, in the case of the Urdu word مرمر [marmar] ('marble') we have two contrasting etymologies, one from Portuguese that was proposed by Dalgado and another from Persian, suggested both by McGregor and Wiktionary; these alternative etymologies are represented in the string etymology divided by the ';'. We use the same standard format for representing all of our etymologies as strings, with each etymology divided from its predecessor by the

[10] In this we are inspired by Ontology Design Patterns (Gangemi & Presutti, 2009), but we would like to apply this concept to LLOD.

[11] https://www.w3.org/2004/02/skos/

semi-colon and the source given in abbreviated form between brackets, i.e., "etymology 1 (source: X1) ; etymology 2 (source: X2, X2')....; etymology k (source: XK,...,XK')".

```
:مرمر_entry a ontolex:LexicalEntry;
    lexinfo:etymologicalRoot pt_lex:mármore ;
    lexinfo:etymology
           "mármore (Source: Dalgado); Persian
      مرمر  (Source: Wiktionary, McGregor)" .
```

Figure 3. The etymological part of a the entry for مرمر

4. CONCLUSIONS

In this article we have looked at some of the features of the linguistic Knowledge Graph we have been creating. As we have mentioned above, the Urdu/Hindi lexicons are almost complete and can be consulted on Github.[12] This first version is shortly due to be made available via a public SPARQL endpoint as one of the use cases of the H2IOSC project, and deposited in a CLARIN repository. Looking ahead, we are working on a next version of CHAMUÇA that will incorporate borrowings from Portuguese into other South Asian languages, including Bengali and Sinhala. We are actively collaborating with linguistic experts to ensure the accuracy and relevance of the newly integrated data. In addition, we aim to refine our methods for representing etymological uncertainty and to improve the alignment of CHAMUÇA with other linguistic knowledge graphs, thereby enhancing its interoperability within the broader LOD ecosystem.

ACKNOWLEDGEMENTS

This work is partly supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 "Education and Research" Component 2 "From research to business" Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures" Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe" - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

REFERENCES

Dalgado, S. R. (1913). Influência do vocabulário português em línguas asiáticas. Coimbra.

Francopoulo, G. (Ed.). (2013). LMF: Lexical markup framework. Wiley.

Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In S. Staab & R. Studer (Eds.), Handbook on ontologies (pp. 221–243). Springer.

Khan, F., Salgado, A., Costa, R., Ramos, M., Carvalho, S., Silva, R., & Almeida, B. (2023). Encoding domain labels in RDF: Guidelines [GitHub repository]. GitHub. https://github.com/anasfkhan81/EncodingDomainLabelsRDF/blob/main/Guidelines.md

Mambrini, F., & Passarotti, M. C. (2023). The LiLa Lemma Bank: A knowledge base of Latin canonical forms. Journal of Open Humanities Data, 9, 28. https://doi.org/10.5334/johd.145

McGregor, R. S. (1993). The Oxford Hindi-English dictionary. Oxford University Press.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through lemmas: The lexical collection of the LiLa knowledge base of linguistic resources for Latin. Studi e Saggi Linguistici, 58(1). https://doi.org/10.4454/ssl.v58i1.277

[12] https://github.com/anasfkhan81/chamuca/tree/main/chamuca_lex_resource