CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages

Anas Fahad Khan[1], **Ana Salgado**[2], **Isuri Anuradha**[3], **Rute Costa**[2],
Chamila Liyange[4], **John P. McCrae**[5], **Atul Kr. Ojha**[5], **Priya Rani**[5],
Francesca Frontini[1]
[1]CNR-ILC, Italy, [2]CLUNL, NOVA University Lisbon, Portugal, [3]University of Wolverhampton, UK,
[4]University of Colombo, Sri Lanka, [5]University of Galway, Ireland
[1]{fahad.khan, francesca.frontini}@ilc.cnr.it, [2]{anasalgado, rute.costa}@fcsh.unl.pt,
[3]Isuri.Anuradha@wlv.ac.uk, [4]cml@ucsc.cmb.ac.lk,
[5]{atulkumar.ojha, priya.rani}@insight-centre.org, john.mccrae@universityofgalway.ie

Abstract

This paper introduces CHAMUÇA, a novel lexical resource designed to document the influence of the Portuguese language on various Asian languages, initially focusing on South Asian languages. Through the utilisation of linked open data and the OntoLex vocabulary, CHAMUÇA provides structured insights into the linguistic characteristics and cultural ramifications of Portuguese borrowings across multiple languages. The article outlines CHAMUÇA's potential contributions to the linguistic linked data community, emphasising its role in addressing the scarcity of resources for lesser-resourced languages and serving as a test case for organising etymological data in a queryable format. CHAMUÇA emerges as an initiative towards the comprehensive catalogisation and analysis of Portuguese borrowings, offering valuable insights into language contact dynamics, historical evolution, and cultural exchange in Asia, one that is based on linked data technology.

Keywords: portuguese, ontolex, language contact, lexicon

1. Introduction

In the current article, we introduce a novel lexical resource titled **Cultural Heritage and Multilingual Understanding through lexiCal Archives (CHAMUÇA)** that is currently under preparation. The intention behind the resource is to describe the impact that the Portuguese language has had on the lexicons of the languages of Asia, with an initial focus on those of South Asia. CHAMUÇA, when complete, will consist of lexicons of Portuguese borrowings in each of the target languages covered by the resource along with a Portuguese language lexicon containing detailed information on each single etymon mentioned in the other lexicons. CHAMUÇA will be published on both in TEI-XML and as linked open data; in the current submission, we will focus on the latter. As we detail below, CHAMUÇA is informed by a number of relevant lexical and scholarly sources including pre-existing dictionaries, research articles and monographs, however, it will be based directly on open-source lexical resources such as *Wiktionary* and *Wikidata*. In turn, it will be published with a Creative Commons Attribution licence. The intention is for CHAMUÇA to be an open-source lexical resource that will be expanded through crowdsourcing.

We begin this article by presenting the background to the project and motivating the need for such a resource in the first place. Then we will go into some more details on the planned resource itself, including the languages in which we will begin by covering and the kinds of information which we plan to include. We also highlight those aspects of CHAMUÇA which are potentially of most interest to the linguistic linked data community. In addition, an example is presented from the Portuguese and Hindi lexicon to illustrate the content of CHAMUÇA.

2. Historical and Linguistic Background

Portuguese has a lengthy history of influence in Asia, stemming from the presence of Portuguese traders and colonists on the continent, traceable back to the 15th century and figures such as Pêro da Covilhã and Vasco de Gama. It is arguable that, with the very obvious exception of English, no other modern European language has had as much impact as Portuguese on the lexicons of the languages of, at least, South Asia. This influence can often manifest itself culturally in interesting and perhaps unexpected ways. One such example is the lexical unit *balti* which refers to a variety of Punjabi cuisine which is popular in the United Kingdom[1].

This borrowing, which entered British English

[1] https://visitbirmingham.com/inspire-me/areas/balti

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

from Hindi/Urdu a few decades ago, ultimately derives from the Portuguese lexical unit *balde* ‘bucket’. A detailed history of language contact between Portuguese and the languages of Asia and the formation of Portuguese language creoles, as well as a survey of previous work in this area, can be found in Cardoso’s seminal article (Cardoso, 2016).

In the current work, we focus on borrowings into pre-existing Asian languages resulting, directly or indirectly, from this historical contact rather than on Portuguese creoles. These borrowings range from a handful of lexical units in languages such as Tibetan to languages with hundreds of Portuguese borrowings. It is interesting to note that although Hindi and Urdu, two of the most widely spoken languages in South Asia, only feature a few dozen borrowings from Portuguese (and these are generally shared by both languages), a good number of these are common everyday words: e.g., those for key (*chabi*), room (*kamra*), and even the word for English (*ingrez*). Other languages, such as Sinhala and Malayalam exhibit a much more substantial Portuguese lexical influence, reflecting a greater level of contact with Portuguese traders and colonists. Cardosos’s article (Cardoso, 2016), and indeed research in this area in general, is heavily in debt to the work of the turn of the century scholar Sebastião Rodolfo Dalgado, and in particular his lexicon of Portuguese borrowings in Asian languages (Dalgado, 1913), a work which has had a significant influence on CHAMUÇA.

3. CHAMUÇA as Lexical Resource

3.1. The Why and How

Many interesting questions arise from the borrowings discussed in the previous section, considering various linguistic, historical, and cultural factors. While it is true that some of the information that could be used to respond to such questions is currently only available in print (non-digitized) resources or behind paywalls, a lot of it is currently available online and, in many cases, under an open license via sites as Wiktionary and Wikipedia[2]. In this latter case, however, the information can either be incomplete, or unavailable in a structured form that can be easily queried using formal languages such as SPARQL. This is where CHAMUÇA enters the scene. The idea is precisely to create a structured lexical resource of Portuguese borrowings into Asian languages: one that is initially bootstrapped using open publicly available sources. In particular, we will make use of Wiktionary, and its RDF version DBnary (Sérasset, 2015), for basic linguistic and grammatical information. This will be augmented by further relevant lexical information, such as corpus frequency data for the borrowed words using contemporary corpora for the languages in question, example sentences, more detailed domain label information, and alternative etymologies. It is important to emphasise that the authors of this submission – who are also the core contributors to this work – include not only speakers of the languages covered by the first version of CHAMUÇA but linguists and lexicographers who have worked with the languages in question as experts (including Portuguese) and will be able to curate the information that is included in CHAMUÇA, thereby adding scholarly value to the resource.

We have initiated our work on CHAMUÇA by focusing on the South Asian languages Urdu/Hindi, Sinhala, Tamil, Gujarati and Bengali. The plan is to open CHAMUÇA up to crowd-sourcing (initially via Github) to allow the addition of more words, more lexical information and more languages (again, this information will be checked and curated by the experts working on CHAMUÇA). The plan would be eventually to create an updated version of Dalgado’s lexicon of Portuguese borrowings in Asian languages. One could ask whether such a resource is really necessary in the age of LLMs. However, after having carried out several experiments with ChatGPT, we found that it was very often unreliable with the kind of lexical information we were interested in; in short, then, the answer is yes.

From a high-level, architectural, perspective, CHAMUÇA is a *lexical resource*, where we understand this term as it is defined in the 2008 version of the Lexical Markup Framework standard (Francopoulo, 2013), that is, as a container for one or more lexicons. In our case, each separate lexicon belongs to a different language and consists of lexical units borrowed from Portuguese, or at least units which can plausibly be said to have been borrowed from Portuguese (since some words have conflicting etymologies)[3]. We decided to publish our resource in linked data because aside from the more general benefits of publishing data in a structured format and using a recognised standard[4], the graph-based RDF model seems to be ideal for a resource structured in the way that CHAMUÇA is –

[2] see for instance `https://en.wikipedia.org/wiki/List_of_Sinhala_words_of_Portuguese_origin`

[3] That is, aside from the obvious case of the CHAMUÇA lexicon for Portuguese which contains lexical information on the Portuguese etymons which are featured in the other CHAMUÇA lexicons.

[4] Benefits which we would also have from publishing the resource solely in TEI-XML, a format which humanists and especially lexicographers tend to be more comfortable with, or at least less suspicious of.

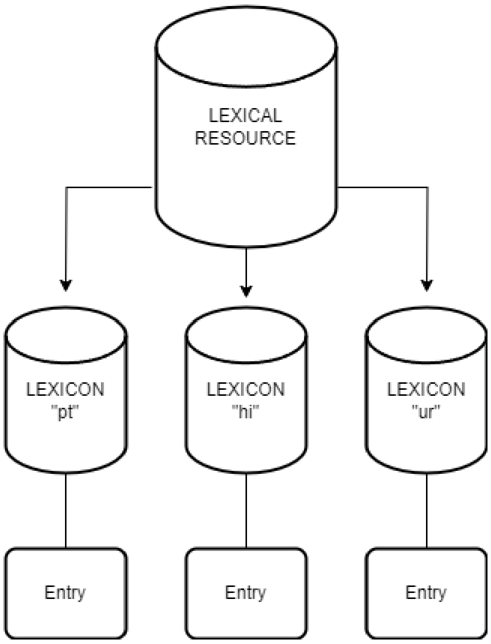

Figure 1: CHAMUÇA as a lexical resource

this becomes especially clear when one considers the power of the graph traversal-based SPARQL query language. In addition, RDF makes it easy to link to other kinds of resources (we intend to link CHAMUÇA to other, non-linguistic linked data resources including historical/geographical ones). In modelling CHAMUÇA as linked data, and more generally, as a structured dataset in the first place, we began by thinking about the kinds of questions (competency questions) that a user might ask of such a resource, e.g., those relating to which domains the borrowed words tend to belong to in a given language and how this changes across languages (and what this can tell us about the particular historical conditions of cultural contact for that given language) or those relating to the extent to which phonological, grammatical and semantic features are preserved (such as gender) or altered in different languages. This determined both what we intended to include as well as how it would be structured. At the time of writing, we have generated the first version of our Portuguese, Hindi and Urdu lexicons in RDF. Before we describe the dataset itself, we note the following points of interest for the linguistic linked data community:

- CHAMUÇA will cover (non-European and in some cases non-European and non-Indo-European) languages that currently don't have many resources dedicated to them in the LLOD cloud (as well as being lesser-resourced languages more generally). Building OntoLex lexicons for these languages will help us, among other things, in understanding the extent to which different kinds of linguistic phenomena associated with these languages can be described by this model.
- CHAMUÇA is a kind of specialised lexical resource (a lexical resource consisting of lexical borrowings from a single language that has a strong cultural and historical interest) that so far has not been represented in the LLOD cloud, and which hasn't yet been covered in any existing OntoLex reports or sets of guidelines and best practices.
- CHAMUÇA will serve as a test case for the structuring of etymological information in a way that can be easily queryable.
- CHAMUÇA will allow us to further develop previous work on domain labelling[5] carried out by some of the authors of the current submission as part of a Short Term Scientific Mission for the Nexus Linguarum COST action[6] – since we plan to add domain labels explicitly to our data, informed by the approach set out in (Salgado, 2022).

In particular, we intend to contribute to current efforts in the BPMLOD W3C group[7] on the creation of guidelines and best practices for LLD for tasks related to each single point listed above (Khan et al., 2022). In particular, we intend to create a series of metadata patterns for specifying the relationship of single resources with others both within the resource (in our case a single lexical resource and component lexicons) and external resources from which a given LLD lexical resource has been derived.

3.2. Generating a First Version of Chamuça

As a first experiment, we converted our initial dataset, composed of lexicons for three languages, Portuguese, Hindi and Urdu into linked data using the OntoLex vocabulary; for now the information in these lexicons derives principally from Wiktionary, although as mentioned above we plan to augment this with additional information in future. Our data was originally stored as a TSV file which was used to generate the RDF sources (and which will be used to generate the TEI-XML too) via a Python script[8]. The result is a first

[5] https://github.com/anasfkhan81/EncodingDomainLabelsRDF/blob/main/Guidelines.md

[6] https://nexuslinguarum.eu/

[7] https://www.w3.org/community/bpmlod/

[8] We intend to make the RDF files available by the time of the workshop, for various logistical reasons we weren't able to make them available by the time of submission.

version of Chamuca-RDF which consists of four separate files `chamuca_lexical_resource`, `chamuca_pt_lexicon`, a lexicon of Portuguese etymons, `chamuca_hi_lexicon`, a lexicon of Portuguese borrowings into Hindi, and `chamuca_ur_lexicon`, a lexicon of Portuguese borrowings into Urdu. As mentioned above `chamuca_lexical_resource` is a container for the three OntoLex lexicons, and will contain lexicons for other languages when they are ready. Since there is no specific class for lexical resources in OntoLex we have made `chamuca_lexical_resource` a subclass of `DCAT:dataset` from the Data Category Vocabulary[9]. We link `chamuca_lexical_resource` to its component lexicons using the Dublin Core `hasPart`.

```
:chamuca_lexical_resource a dcat:dataset ;
    dct:hasPart
        chamuca_hi_lex:,
        chamuca_ur_lex: ;
        chamuca_pt_lex: ;
    dct:language
        ”hi”, ”pt”, ”ur” ;
    dct:license
<https://creativecommons.org/licenses/by/4.0/>;
    dct:title
        ”chamuça”@eng .
```

In order to show the relationships between separate lexicons and the kinds of information which this first iteration of the language resource contains, we look at a single entry in Portuguese and its corresponding entry in the Hindi lexicon. The entry for *câmara* meaning ’chamber’ (at least in its primary sense `câmara_sense_1`) in `chamuca_pt_lex` is as follows:

```
:câmara_entry a ontolex:LexicalEntry,
        ontolex:Word ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:partOfSpeech
        lexinfo:commonNoun ;
    ontolex:canonicalForm :câmara_lemma ;
    ontolex:lexicalForm :câmara_plural ;
    ontolex:sense :câmara_sense_1,
        :câmara_sense_2,
        :câmara_sense_3,
        :câmara_sense_4,
        :câmara_sense_5,
        :câmara_sense_6,
        :câmara_sense_7 .
```

The entry for कमरा (*kamra*) ’room’ the Hindi word corresponding to *câmara* is as follows:

```
:कमरा_entry a ontolex:LexicalEntry,
    ontolex:Word ;
    lexinfo:etymologicalRoot
        chamuca_pt_lex:câmara ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:partOfSpeech
        lexinfo:commonNoun ;
    rdfs:seeAlso
        chamuca_ur_lexicon:kamra ;
    ontolex:canonicalForm
        :कमरा_lemma ;
    ontolex:lexicalForm
        :कमरे_dp_form_कमरा,
        :कमरे_os_form_कमरा,
        :कमरे_vs_form_कमरा ;
        :कमरो_vp_form_कमरा,
        :कमरों_op_form_कमरा ;
    ontolex:sense
        :कमरा_sense .
```

From the preceding, one can see that the word switched its grammatical gender in entering Hindi, this is not unusual since the ’-a’ ending in Hindi and Urdu is usually associated with masculine nouns (with the opposite being true in Portuguese). Our immediate plans are to add a fuller etymology for each Portuguese etymon, as well as having an example sentence for each word in the target languages along with corpus frequency and attestation data, using the Frequency Attestation and Corpus module of Ontolex, currently under development.

4. Future Work and Conclusion

In this article, we have introduced our ongoing development of CHAMUÇA, a novel lexical resource documenting the Portuguese influence on various Asian languages, with an initial focus on South Asian languages. By leveraging linked data principles and the OntoLex vocabulary, we have structured CHAMUÇA to facilitate accessibility, interoperability, and queryability. Through our efforts, we have transformed initial datasets into Chamuça-RDF, comprising lexicons for Portuguese, Hindi, and Urdu. This structured representation will potentially enable us to explore relationships between lexicons and delve into borrowed word domains across languages. Moving forward, CHAMUÇA holds the promise of being a valuable resource for linguistic research, historical inquiry, and cultural understanding. Ultimately, CHAMUÇA is intended to stand as a testament to the collaborative efforts of linguists, lexicographers, and language enthusiasts in preserving and exploring the rich tapestry of linguistic interactions between Portuguese and Asian languages.

[9] https://www.w3.org/TR/vocab-dcat-3/

5. Acknowledgements

The research of Ana Salgado and Rute Costa is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020; 10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020; 10.54499/UIDP/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

John P. McCrae, Atul Kr. Ojha and Priya Rani would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics.

References

Hugo C Cardoso. 2016. O português em contacto na ásia e no pacífico. *Manual de linguística portuguesa*, pages 68–97.

Philipp Cimiano, John P McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Community report. *Final community group report, 10 may 2016, W3C*.

Sebastião Rodolfo Dalgado. 1913. *Influência do vocabulário português em línguas asiáticas:(abrangendo cêrca de cinquenta idiomas)*. Impr. da Universidade.

Gil Francopoulo. 2013. *LMF lexical markup framework*. John Wiley & Sons.

Fahad Khan, Christian Chiarcos, Thierry Declerck, Maria Pia Di Buono, Milan Dojchinovski, Jorge Gracia, Giedre Valunaite Oleskeviciene, and Daniela Gifu. 2022. A survey of guidelines and best practices for the generation, interlinking, publication, and validation of linguistic linked data. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 69–77, Marseille, France. European Language Resources Association.

Ana Salgado. 2022. *Terminological Methods in Lexicography: Conceptualising, Organising, and Encoding Terms in General Language Dictionaries*. Ph.D. thesis, Universidade NOVA de Lisboa.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.