Anas Fahad Khan, Ana Salgado, John McCrae, Chamila Liyange, Priya Rani, Rute Costa, Isuri Anuradha, Atul Kr. Ojha, and Francesca Frontini

CULTURAL HERITAGE AND MULTILINGUAL UNDERSTANDING THROUGH LEXICAL ARCHIVES (CHAMUÇA)

Portuguese Borrowings in Contemporary Asian Languages

Abstract CHAMUÇA (Cultural HeritAge and Multilingual Understanding through lexiCal Archives) is a pioneering initiative aimed at exploring the impact of the Portuguese language on Asian languages, rooted in the historical exchanges between Portuguese traders, colonists, and diverse Asian cultures. The impact of these interactions extends beyond historical remnants to the modern-day lexicon of Asian languages, which includes a diverse array of Portuguese borrowings, ranging from general vocabulary units to specialised units. We aim to detail the initiative's current status, its goals, and the methodology it employs. Additionally, it will outline the essential steps required for organising and structuring the knowledge embedded within and associated with the borrowings. CHAMUÇA, an innovative open-source resource designed to document and study these Portuguese linguistic contributions, will augment the pool of structured lexical data and support cross-linguistic analysis, using state-of-the-art frameworks such as OntoLex-Lemon and TEI Lex-0 to structure the lexical data. Following FAIR principles – ensuring data is findable, accessible, interoperable, and reusable – CHAMUÇA is poised to contribute to linguistic borrowings, cultural interchange, and the preservation of linguistic heritage. Furthermore, the project will encourage community involvement and scholarly collaboration to evolve and enrich its contents, leveraging collective expertise to illuminate the nuances of language contact phenomena.

Keywords language contact; lexical resource; lexicon; Ontolex; Portuguese language; South Asian languages; TEI

1. Introduction

The current article discusses a newly commenced project CHAMUÇA (Cultural Heritage and Multilingual Understanding through lexiCal Archives) which aims to bring together lexicographers, linguists and other relevant experts to document and analyse the historical influence of the Portuguese language on the languages of Asia, with an initial focus on South Asia. The project will publish its results as a FAIR lexical resource, also named CHAMUÇA, to be based on data from pre-existing language resources, as well as on original scholarly work and crowdsourcing. This resource, which will be released in several versions each under an open licence, is intended to be an open ended one. At the time of

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

writing, we are working on releasing the first version of CHAMUÇA for a limited number of languages. CHAMUÇA, the language resource, will be made available in two open formats, utilising pre-existing FAIR vocabularies whenever possible. More specifically, it will be published in XML using TEI Lex-0[1], a customisation of the Guidelines for Electronic Text Encoding and Interchange[2] (TEI-XML) and as Linked Open Data (LOD) using the OntoLex-Lemon[3] vocabulary. Each version of CHAMUÇA will offer a comprehensive lexical archive of Portuguese borrowings in each of the target languages that it covers, such as Urdu/Hindi, Sinhala, Tamil, Gujarati, and Bengali.

The rest of the paper is organised as follows. Section 2 gives some historical background on the influence of Portuguese in Asia, with a concentration on the linguistic aspect of this influence. Section 3 shifts to the CHAMUÇA project, detailing its development, focusing on data compilation and the use of conceptual models and guidelines, such as OntoLex-Lemon and TEI Lex-0, for structuring a lexical archive. Finally, Section 4 highlights future work and presents concluding remarks.

2. Legacy of Portuguese in Asian Languages

2.1 Historical Background

The influence of the Portuguese language in Asia began with the so-called Age of Discovery (*Era dos Descobrimentos* in Portuguese) in the 15th and 16th centuries, a period of unprecedented maritime expansion for what was then the Kingdom of Portugal. In particular, during that time several trading posts were established on the Malabar coast, including locations such as Cochin, Cannanore, Quilon, Cranganor, Tanor, and Calicut, giving the Portuguese their first foothold in the continent. Over the next few centuries, and thanks to extensive contact with numerous Asian cultures through trade and colonisation, the Portuguese language was able to leave behind a profound linguistic legacy in the whole of Asia. As Cardoso (2016, p. 71) states, "A língua portuguesa encaixou-se na região asiática ao ponto de se converter em importante língua franca de comérico e diplomacia" [The Portuguese language became deeply rooted in the Asian region to the extent that it evolved into an important lingua franca for commerce and diplomacy]. Despite subsequent, largely successful, moves towards challenging Portuguese dominance in the region by other colonial powers such as the Dutch, French and English, Portugal maintained a presence in colonies such as Goa, Daman, Diu, East Timor, and Macau well into the 20th century.

Today, traces of this influence can still be seen in religious practices and various local cuisines as well as in linguistic borrowings, including toponyms, and, more broadly speaking, the linguistic identities of numerous different Asian communities.

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

[2] https://tei-c.org/guidelines/

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

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The Portuguese language, once an instrument of trade and empire-building, has inadvertently sown seeds in the rich linguistic soil of Asia, cultivating a lexical heritage that continues to exert a strong influence on modern vernaculars. This influence can be profound, often surfacing in unexpected cultural contexts and leaving behind a rich, indelible legacy that the CHAMUÇA project aims to explore and document. By identifying words of Portuguese origin in Asian languages, the scope of the CHAMUÇA project, we not only document the spread of the Portuguese language in a more global context but also facilitate the study of how these languages adapt and incorporate foreign elements over time.

2.2 Review of Past Literature and Contemporary Scholarship

Cardoso (2016) offers an essential entry point into the scholarly literature on this topic, one that has historical roots stretching back centuries. As well as containing an excellent overview of previous studies on the diffusion of Portuguese throughout Asia and the wider Pacific region from the 16th century onwards, it also details the development of Asian Portuguese creoles. Indeed, the article covers the whole spectrum of linguistic integration, ranging from languages with a modest array of Portuguese-derived words to those which are richly studded with borrowings from Portuguese, to the aforementioned creoles. Cardoso's study also highlights the significance of early Portuguese contact on the Malabar coast and the subsequent linguistic impacts this has had. In terms of historical scholarship, Sebastião Rodolfo Dalgado's pioneering work (Dalgado, 1913) continues to be a cornerstone in this field. In particular, his lexicon of Portuguese borrowings in Asian languages remains an essential lexicographic and scholarly reference which continues to inform and inspire ongoing research, including the CHAMUÇA project.

2.3 Overview of Portuguese Borrowings in South Asian Languages

In this first, preparatory, phase of CHAMUÇA, and given the expertise and interests of the initial participants to the project, we decided to focus on the Portuguese influence in South Asia and South-Eastern Asia, an area covering modern-day countries such as Bangladesh, India, Indonesia, Malaysia, Nepal, Pakistan, and Sri Lanka. This part of the world is rich in linguistic diversity and home to hundreds of languages belonging to the Indo-European, Dravidian, Sino-Tibetan, and Austroasiatic language families. Utilising the framework provided by Cardoso (2016) (which in turn references Dalgado's (1913) study) we list the initial languages covered[4] by the CHAMUÇA project below (Table 1) with their respective language codes and the number of borrowings for each language listed as given in Dalgado's 1913 lexicon.[5] In CHAMUÇA we would like to create a lexical resource which makes past research into these borrowings accessible by publishing them in the form of a collection of digital lexicons.

[4] Note that these are only a subset of the languages covered in Dalgado's original lexicon.

[5] The current number will presumably be less, although there is the possibility that some newer borrowings such as 'piri-piri' or 'samba' have since then entered these languages, especially via English.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Table 1: Portuguese linguistic borrowings

Languages	**Language Code (ISO 639)**	**Number of Borrowings (Dalgado 1913; Cardoso 2016)**
Bengali	bn	173
Hindi/Urdu	hi/ur	53/107
Gujarati	gu	105
Konkani	kok	1768
Malay	ms	431
Marathi	mr	116
Malayalam	ml	127
Sinhala	si	208
Tamil	ta	171
Telugu	te	83

We can make a number of initial observations about the data in the table. Firstly, Konkani (an Indo-Aryan language primarily spoken in the Konkan region, along the western coast of India) shows the highest influence from Portuguese with 1,768 borrowings listed in Dalgado's work, a figure which is indicative of the deep and prolonged contact with Portuguese culture and language which took place in Konkan. Malay (an Austronesian language spoken in Malaysia, Indonesia, Singapore, Brunei and Thailand) follows with 431 borrowings, evidence of strong maritime trade links and colonial interactions with the Kingdom of Portugal. Hindi/Urdu (widely spoken across the Indian subcontinent) demonstrates a moderate level of Portuguese influence, with 53 and 107 borrowings respectively. Other languages like Bengali (an Indo-Aryan language native to the Bengal region of South Asia), Gujarati (an Indo-Aryan language native to the Indian state of Gujarat), Marathi (an Indo-Aryan language predominantly spoken in the Indian state of Maharashtra), Malayalam (a Dravidian language spoken in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry), Sinhala (an Indo-Aryan language primarily spoken in Sri Lanka), Tamil (a Dravidian language natively spoken by the Tamil people of South Asia), and Telugu (a Dravidian language native to the Indian states of Andhra Pradesh and Telangana) also display a significant number of borrowings, ranging from 83 to 208. This data provides insight into the historical depth of the Portuguese linguistic legacy and helps identify areas for further research in the CHAMUÇA project.

There are many interesting questions to be asked about these borrowings, in particular those looking into the interplay of phonological, morphological, semantic and cultural

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

factors behind each borrowing. The purpose of CHAMUÇA is to help answer them via a single, easily accessible, machine-actionable language resource. It might be pointed out that a lot of lexical information relevant to this subject – and which could serve to answer the questions alluded to above – is already widely and freely available on sites like *Wiktionary* and *Wikipedia*. However, this information can often be incomplete (e.g., lacking data on corpus frequencies, something which we plan to add in our work), is not always completely reliable, and nor is it often available as structured data (e.g., DBnary, the linked data version of Wiktionary, does not contain structured etymologies and so we cannot extract such borrowings via a SPARQL query). As part of our background work on this project, we tried numerous ChatGPT prompts to ask some of the research questions which we would like to ask of our resource (e.g., 'list Hindi words which derive from Portuguese') and as expected the results were not very reliable. On the other hand, a great deal of the scholarly work which exists in this field (as in many others) is distributed across large numbers of articles, scholarly and lexicographic works, and in different languages (largely Portuguese) which are often closed or behind a paywall/unavailable electronically, and rarely as structured language resources.

3. CHAMUÇA as Language Resource

Having been motivated by the considerations described above, we decided that we would make CHAMUÇA an open-source lexical resource published using two existing standards, TEI Lex-0[6] and OntoLex,[7] and made available via a Creative Commons Attribution (CC-BY) licence. All of this will help us to ensure that our resource is FAIR by design (although by itself it won't by itself be sufficient to ensure this, we also intend e.g., to add extensive metadata and publish our resource in a repository). In addition, the use of both TEI Lex-0 and OntoLex means that our resource will be machine-actionable. The project also aims to use crowdsourcing methods through platforms like GitHub to enrich the database with additional lexical units, and detailed lexical information. A dedicated team of linguists, and lexicographers will carefully review and curate all contributions to ensure our resource consistently remains of the highest quality. Native speakers will also contribute to ensure language accuracy. Our vision is for CHAMUÇA to serve as a dynamic, modern counterpart to Dalgado's 1913 lexicon, and to therefore help in capturing the evolving nature of Portuguese borrowings in Asian languages. This project is not just about preserving knowledge but enhancing how we understand and interact with linguistic heritage in the digital era, potentially making it an important tool for researchers and the general public in navigating cultural and linguistic history and fostering an appreciation of linguistic diversity.

3.1 Building the Lexical Archive

The first step in creating the CHAMUÇA lexical resource is to develop lexicons for the languages listed in Table 1 along with an index lexicon for Portuguese. These

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

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

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

languages were chosen (from among the many we could have worked with) both because they aligned with the linguistic expertise of the researchers participating in the CHAMUÇA project and because each of these languages demonstrates what we consider to be a significant enough level of borrowing from Portuguese (at least in terms of the frequency and common use of the borrowed lexical units if not always in terms of number of borrowings).

This preparatory work for each Asian language lexicon begins by sourcing lexical information from Wiktionary, using a list of candidate lexical units from Dalgado's lexicon (1913) in combination with other lexicographic and scholarly sources (and eventually corpora information). In most cases Wiktionary provides us with basic kinds of linguistic and lexicographic information which we can subsequently modify and enrich – in particular we intend to enrich our lexical entries with relevant cultural, historical, and geographic data where available, leveraging the possibilities offered by linked data to create a rich, multidimensional resource.

The architecture of CHAMUÇA consists of a single lexical resource, which acts as a container for the separate lexicons[8] for each Asian language covered by the project, all of which are linked to a 'central' Portuguese lexicon. This latter acts as an index for these other lexicons; the architecture of our resource, then, mirrors the structure of Dalgado's 1913 lexicon. We note that the graph-based RDF model of linked data (see Arenas et al., 2009) has proven to be a natural fit for CHAMUÇA, particularly when paired with the graph traversal expressivity of the SPARQL query language.

We have tried to ensure CHAMUÇA's usefulness and relevance by guiding its development via 'competency questions' – specific inquiries that users might potentially ask of our resource via SPARQL queries. These questions range from questions about the domains Portuguese borrowings belong to, to others which concern the specific historical and cultural circumstances of contact for each language. In addition to these kinds of questions, we want our resource to be able to respond to questions respecting the preservation and adaptation of phonological, grammatical, and semantic features, such as gender, across different languages. At a minimum, for each lexical entry in each Asian language lexicon, we capture the following:

- Standard information associated with lexical entries: including the lemma of the entry, its part of speech, transliteration, pronunciation, gender, sense definition, links to related resources including Wiktionary and Wordnet (for the relevant language where available), different morphological variants of the form and other grammatical information
- Etymological information including an original hypothesised Portuguese etymon or etymons and notes
- In the case of Hindi, the equivalent (it exists) in Urdu and vice versa

[8] In this we were inspired by the definition of the class Lexical Resource in the LMF standard (ISO 24613-1:2019).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

- A flag if the word is found in Dalgado's original 1913 lexicon
- One or more domain label tags, see the next section.

At the time of writing, as a test, we have completed the first iteration of our lexicons for Portuguese, Hindi, and Urdu. More information on how these lexicons are structured and their modelling in RDF using OntoLex-Lemon can be found in (Khan et. al, 2024).

3.2 Knowledge Organisation

A pivotal aspect of the construction of the CHAMUÇA lexical resource is the organisation of knowledge to capture the complexity and richness of the original linguistic phenomena being described while ensuring the resource's accessibility and practical utility. To this end, we have adopted domain labelling to categorise and delineate the contexts in which Portuguese borrowings can be found in different Asian languages. Domain labelling foresees the classification of each lexical borrowing into specific categories based on usage and semantic considerations. Note that, by a domain label, we mean a 'marker which identifies the specialised field of knowledge in which a lexical unit is mainly used' (Salgado, Costa, and Tasovac, 2019). A manual approach is chosen to ensure precise and context-aware categorisation, capturing subtle nuances that automated systems might overlook. This organisation method facilitates a user's navigation through the resource, helping them understand the integration of Portuguese lexical items into the relevant cultural and functional contexts. For example, the Hindi and Urdu words for 'church' are both borrowed from the Portuguese *igreja*, **गिरिजा** and اجڑگِ (both with the pronunciation girajā), respectively and are categorised under Religion, underscoring the specific cultural significance of this borrowing. The influence of the Portuguese language is manifest in its introduction of new lexical units across multiple domains as noted by Dalgado (1913, pp. XXVII-XXIX). These impacts are organised into several principal domain labels such as the following:

Religion. This domain label includes units related to Christian practices and concepts introduced by Portuguese missionaries. For instance, the Hindi and Urdu words for 'cross', (**क्रूस** Hi., سورک Ur.) are derived from *cruz*, and those for '(Christian) priest' (**पादरी** Hi., یادرپا Ur.) from *padre*. This labelling not only aids in the study of linguistic influence in the context of religious conversion but also facilitates the tracing of how these terms were adopted and adapted by other cultures.

- **Botany.** The domain covers units referring to plant species introduced across the Asian continent which retained their original Portuguese names. This is the case for the Hindu and Urdu words for pineapple, (**अनन्नास** Hi., اننانس Ur.), which both derive from *ananás*, and those for cabbage (**गोभी Hi.**, یهبوگ Ur.), from *couve,* at least according to one hypothesis. This focus allows for a detailed study of how these plant names propagated across different regions along with the plants themselves.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

- **Culinary.** The significant impact of Portuguese influence on Indian cuisine is evident in a number of culinary terms. It was responsible for introducing new ingredients and products, including chilli peppers, which revolutionised Indian cooking. An exemplary term is *balti*, originally from the Portuguese word *balde* 'bucket', now synonymous with a style of Punjabi cuisine popular in the UK. This reflects not only the culinary practices but also the journey of a word across cultures.
- **Clothing.** Lexical borrowings demonstrate how Portuguese fashion influenced local attire, filling gaps and adding diversity to the clothing vocabulary in Asian languages.
- **Cultural**: This domain indicates items of everyday use and material culture, such as the words for 'cabinet' in Hindi and Urdu which are borrowed from *armário* (**अलमारी** Hi., الماری Ur.), the word for 'bucket' from *balde* (**बाल्टी Hi**., بالٹی Ur.)[9], and pistol from *pistola* (**पस्तौल Hi**., پستول Ur.). Analysing these lexical units helps in understanding how the Portuguese language influenced daily life and material culture through trade and colonial interactions.

Through these organised domains, CHAMUÇA helps to contribute towards a general understanding of cultural linguistics.

4. Final Considerations and Future Work

The CHAMUÇA project aims to describe and assist into research on the impact of Portuguese borrowings into Asian languages, helping to provide data on the complex dynamics of language contact and cultural exchange. By employing advanced lexical frameworks and adhering to FAIR principles, we aim to ensure that the data is not only accessible but also interoperable across different platforms.

While the initial phase focused on South Asian (in addition to one Southeast Asian language), future expansions will include East and Southeast Asian languages, broadening the scope of our research and making the archive more inclusive. We aim to collaborate with linguists, historians, and cultural scholars to explore the broader implications of linguistic borrowings. These collaborations will help contextualise the lexical data within larger narratives of historical interactions and cultural transformations.

Ensuring the long-term sustainability of CHAMUÇA is a priority. We will seek additional funding sources and partnerships to maintain and expand the archive.

References

Arenas, M., Gutierrez, C., & Pérez, J. (2009). Foundations of RDF databases. In *Reasoning Web International Summer School* (pp. 158–204). Springer Berlin Heidelberg.

[9] Note that in cases where a borrowing has more than one sense we can either tag at the sense level or the level of the whole entry. In the case of the borrowing of *balde* for instance we will tag at the sense level.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Cardoso, H. (2016). O português em contacto na Ásia e no Pacífico. In A. M. Martins, & E. Carrilho (Eds.), *Manual de Linguística Portuguesa* (pp. 68–97). Berlin, Mouton de Gruyter.

Cimiano, P, McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report, Technical Report.* W3C Ontology-Lexicon Community Group. https://www.w3.org/2016/05/ontolex/

Dalgado, S. R. (1913). *Influência do vocabulário português em línguas asiáticas (abrangendo cerca de cinquenta idiomas).* Imprensa da Universidade.

Khan, F., Salgado, A., Anuradha, I., Costa, R., Liyanage, C., McCrae, J. P., Ojha, A. K., Rani, P., & Frontini, F. (2024). CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024* (pp. 44–48). ELRA and ICCL.

Salgado, A., Costa, R., & Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography. *Journal of ASIALEX, 6*(2), 133–156. doi:10.1007/s40607-019-00061-x

Tasovac, T., Romary, L. et al. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1.* DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html, last accessed 26 January 2024.

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 Research Centre for Data Analytics.

Contact information

Anas Fahad Khan
CNR-ILC, Italy
fahad.khan@ilc.cnr.it

Ana Salgado
NOVA University Lisbon, Portugal
anasalgado@fcsh.unl.pt

John McCrae
University of Galway, Ireland
john.mccrae@universityofgalway.ie

Chamila Liyange
University of Colombo, Sri Lanka
cml@ucsc.cmb.ac.lk

Priya Rani
University of Galway, Ireland
priya.rani@insight-centre.org

Rute Costa
NOVA University Lisbon, Portugal
rute.costa@fcsh.unl.pt

Isuri Anuradha
University of Wolverhampton, UK
Isuri.Anuradha@wlv.ac.uk

Atul Kr. Ojha
University of Galway, Ireland
atulkumar@insight-centre.org

Francesca Frontini
CNR-ILC, Italy
francesca.frontini@ilc.cnr.it

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.