Exploring consonant frequency in Sri Lanka Portuguese

Carlos Silva[1]*[0000−0002−8052−4271] and Luís Trigo[2][0000−0002−3772−7081]

[1] CLUP Centro de Linguística, University of Porto
[2] LIACC Laboratório de Inteligência Artificial e Ciência de Computadores, University of Porto
silvacarlosrogerio@gmail.com, trigoslab@gmail.com

Abstract. Although phoneme selection is a well-studied subject in contact linguistics, phoneme integration is mostly unexplored. This study aims at assessing phoneme integration by measuring consonant frequency in Sri Lanka Portuguese and Portuguese. For that, we select two large lexical corpora and, take several preparation steps to make the data uniform, consistent and reusable. In terms of integration, we find that the more unconstrained a consonant is concerning its phonotactic patterns, the more frequent it is. We also find that being coronal has a positive impact on integration, whereas being palatal has a negative impact. Moreover, we find that in spite of the apparently random changes in the consonant frequency, consonant classes are robustly transmitted from the lexifier to this creole.

Keywords: Phoneme frequency · Sri Lanka Portuguese · Language contact.

1 Introduction

Language contact leads to borrowing events between the languages involved. On phonological grounds, these events often result in the addition of new phonemes to the language's inventory. Over the last decades, many researchers tried to identify which segments were more or less more/less likely to be borrowed either in individual cases (e.g. [9]) or more generally in the world's languages [13]. These studies give us a good idea of which segments are selected and why, but do not report on how well integrated the phonemes are. Creoles make an especially interesting case study, because they go beyond borrowing. They are assumed to be a case of language shift [23].

While social motivations and structural compatibility are known to play a role in phoneme selection during creole formation, the importance of the unintentional psycholinguistic mechanisms is still to be explored. Usage frequency is one of those mechanisms, which is crucial for the development of an individual's linguistic knowledge [20,11]. Such factors are difficult to evaluate, especially when it comes to understudied languages, due to the lack of consistent data or quantitative methods. However, thanks to recent developments, we are finally able to start filling this gap.

This study aims at assessing the viability of this approach in contact linguistics, as suggested by [14], through the exploration of consonant frequency in two lexical corpora in a lexifier-creole pair, i.e., Portuguese and Sri Lanka Portuguese. Our goal is to check the similarity rates between consonant frequency in the lexifier and the creole, and to establish a possible link between the latter and cross-linguistic frequency [19]. Thus, we expect to contribute to one aspect of the "creole debate", by answering the question "Are the creoles similar to their lexifier [8], their substrates [1] or do they reflect universal preferences [3]?".

Sri Lanka Portuguese can be seen as a typical case of "light creole" [6] or a "trade creole" [2]. It was formed during the first half of the 16 th century, after the establishment of a trading post at Colombo in 1517, born from the contact between Portuguese, Sinhalese, and Tamil [22]. Later, it come into contact with Dutch and English, which are considered its adstrates. The speakers were members of the burgher communities, mainly represented by descendants of Portuguese (and Dutch) men who married local women. After the English took power over the island, this creole declined rapidly [10]. Nowadays, this language is severely endangered, and it is spoken just by few members of the Batticaloa and Trincomalee communities in the Tamil dominant east coast [16].

2 Methods

Although measuring consonant frequency is mathematically simple, the preparation steps are often challenging, especially for understudied languages. On the one hand, if a relevant measurement should rely on a large lexical or usage corpora with phonetic transcription, on the other hand, such corpora are rarely available. Therefore, the first challenge is precisely converting orthographic characters into IPA[3] symbols.

As Sri Lanka Portuguese had no available corpus, we based our research on the most extensive dictionary of this language [10], converted it into a consistent format, and created PtLanka [24], an open-access online database that assembles a total of 2522 words of Sri Lanka Portuguese. Then, we used a mapping table to assign phonetic symbols to the orthographic characters, according to PHOIBLE conventions [19]. This step was taken to facilitate the future conversion of the data into CLDF [12]. Finally, we split the phonetic symbols in order to count them and assign them percentage values. Overall, this data set is composed of 9 126 consonants.

[3] IPA - International Phonetic Alphabet

For Portuguese, we extracted the entries of the Portuguese version of Wiktionary [4]. As a multilingual collaborative web-based project with phonetic information, wiktionary [18] seems to be a suitable tool, which does not imply complex preparatory steps. However, as its structure relies on collective user choices, it has no fixed structure. Therefore, before counting and assigning percentage values to the consonants, some preparation was needed. We started by retrieving the IPA entries and cleaning irrelevant characters (e.g. =, ., *). Then, we corrected the phonetic transcription, and convert some misplaced SAMPA symbols into IPA. Finally, we group together some phonetic variants into broader phonetic symbols (e.g. ɫ→l and ʁ→ʀ). This second data set has a total of 49 674 consonants.

The method was developed with the Python programming language. The implementation in Jupyter Notebook is available in the same repository as PtLanka with the aim of enabling transparency and reproducibility.

3 Results & Discussion

According to Bybee [4,5], the mental representations of language are internal representations of an individual's experience. Then, if we can measure the amount of an individual's or a community's experience with a given phoneme, we can assess how well it is integrated in a given phonological system.

Table 1 shows the raw counts and the percentage values of each consonant in Sri Lanka Portuguese (table 1a) and its lexifier (table 1b).

Regarding individual consonants, we can inspect the results in two dimensions: integration and transmission. Integration corresponds to the degree to which a consonant is frequent and, therefore, more strongly represented in each individual language. Transmission correlates with the extent to which frequency values are kept when comparing the creole and its lexifier.

If we look at PtLanka, we notice that there is no particular natural class on the top, that is, /ɾ s d n/ do not share the same manner of articulation. However, /ɾ s n/ are consonants that are allowed to occupy several syllable positions, such as the onset (simple or complex) and the coda. Consequently, our data suggests that the constrains on phonotactics influence frequency [17] and, therefore, phonological integration. Furthermore, it is worth remarking that the seven more frequent consonants are all coronal consonants. This finding corroborates Carvalho's proposal [7] which says that [coronal] is the unmarked point of articulation within the oral cavity. On the contrary, palatal consonants /d̠ʒ t̠ʃ ɲ ʎ j ʃ l^j n^j/ are grouped on the bottom of the table 1a. This observation can be explained, by their low frequency in the lexifier [25], on the one hand, and, on the other hand, it confirms its high complexity [27,26,21].

When looking at both tables, we find no clues for a robust transmission at a first glance. From those consonants with more statistical relevance, only /b g z f/ have a range of less than 1% between the creole and the lexifier's lexicon.

[4] https://pt.wiktionary.org

Table 1: Consonant frequency in Sri Lanka creole and Portuguese

(a) PtLanka

IPA	count	%
ɾ	1273	13,95
s	1055	11,56
d	905	9,92
n	874	9,58
t	847	9,28
m	679	7,44
p	545	5,97
k	530	5,81
l	461	5,05
ʋ	310	3,4
b	308	3,37
z	305	3,34
f	246	2,7
g	224	2,45
r	160	1,75
d͜ʒ	158	1,73
t͜ʃ	69	0,76
ɲ	54	0,59
ʎ	42	0,46
j	35	0,38
tː	16	0,18
ʃ	15	0,16
l^j	13	0,14
n^j	2	0,02

(b) Wikcionario

IPA	count	%
ɾ	8132	16,37
t	5546	11,16
d	3741	7,53
l	3580	7,21
s	3537	7,12
k	3525	7,1
p	2525	5,08
m	2500	5,03
ʃ	2321	4,67
n	1924	3,87
v	1486	2,99
b	1479	2,98
j	1364	2,75
w	1316	2,65
f	1315	2,65
z	1224	2,46
ʒ	1199	2,41
g	1193	2,4
ʀ	1121	2,26
ʎ	298	0,6
ɲ	208	0,42
k^w	140	0,28

Nevertheless, if we group the consonants into natural classes (e.g. stops, fricatives, etc), following [15], we recognize a robust transmission for most cases. For instance, stops represent 36.5% of the phonemes in Sri Lanka Portuguese and 36% in the lexifier language. The main positive difference is in the nasals, whose usage is increased by 5.3% in the creole. On the opposite side, rothics usage decreases about 3 percentual points. In light of these results, we conclude that, whereas individual consonants may not show strong correlates between the lexifier and the creoles, consonant classes do. The language contact itself and historical change both in the Portuguese and Sri Lanka Portuguese may have affected the phonetic shape of the segments but it didn't have major effects on phonological classes as a whole. Thus, consonants in creole are not simplified, they are adapted.

Although we believe to have reached some interesting results, this study is far from complete. Firstly, we would like to condition the frequency of the consonants and measure it in particular syllable positions (e.g. onset only). On the one hand, this would make the consonants more comparable between them and, on the other hand, it would serve as a test for the hypothesis above, i.e., phonotactics affects consonant frequency. In the second place, a comparison with cross-linguistic frequency and cross-linguistic borrowability rates would shed light on the influence of universal tendencies on phonological integration in creoles. Finally, it would also be worth looking into other creoles which have different languages as substrates and different contact situations, which would complete our view on phonological integration and also bring some valuable outcomes for historical linguistics and second language acquisition.

Acknowledgments This research was conducted within the doctoral program of Languages Sciences (Faculty of Arts-University of Porto), was funded by the Portuguese Foundation for Science and Technology (FCT MCTES) through the PhD grant SFRH/BD/2020.07466.BD and supported by the Center of Linguistics of the University of Porto (FCT-UIDB/00022/2020).

References

1. Alleyne, M.: Comparative Afro-American: an historical-comparative study of English-based Afro-American dialects of the New World. Karoma, Ann Arbor (1980)
2. Bakker, P., Daval-Markussen, A., Plag, I., Parkvall, M.: Creoles are typologically distinct from non-creoles. Journal of Pidgin and Creole Languages **26**(1), 5–42 (2011). https://doi.org/10.1075/jpcl.26.1.02bak
3. Bickerton, D.: Roots of Language. Karoma (1981)
4. Bybee, J.: Phonology and Language Use. Cambridge Studies in Linguistics, Cambridge University Press, Cambridge (2001). https://doi.org/10.1017/CBO9780511612886
5. Bybee, J.: Language, Usage and Cognition. Cambridge University Press, Cambridge (2010). https://doi.org/10.1017/CBO9780511750526
6. Carvalho, A., Lucchesi, D.: Portuguese in contact. In: The Handbook of Portuguese Linguistics, pp. 41–55. Wiley Blackwell (apr 2016). https://doi.org/10.1002/9781118791844.ch3
7. Carvalho, J.B.D.: Why there is no backness: the case for dismissing both [coronal] and [dorsal]. In: Naïm, J.L.L..S. (ed.) Backness and backing, pp. 45–58. Lincom (2013), https://halshs.archives-ouvertes.fr/halshs-01116259
8. Chaudenson, R.: Des îles, des hommes, des langues: essai sur la créolisation linguistique et culturelle. L'Harmattan (1992)
9. Clements, J.: The status of portuguese/spanish /r/ and /r/ in some iberian-based creole languages. PAPIA **24**(2), 343–356 (2014), http://revistas.fflch.usp.br/papia/article/view/2201
10. Dalgado, S.R.: Dialecto indo-português de Ceylão. Imprensa Nacional, Lisboa (1900)
11. Edwards, J., Beckman, M., Munson, B.: Frequency effects in phonological acquisition. Journal of child language **42**, 306–11 (03 2015). https://doi.org/10.1017/S0305000914000634
12. Forkel, R., List, J.M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarstrom, H., Haspelmath, M., Kaiping, G., Gray, R.: Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. Scientific Data **5** (2018). https://doi.org/10.1038/sdata.2018.205, https://doi.org/10.1038/sdata.2018.20

13. Grossman, E., Eisen, E., Nikolaev, D., Moran, S.: Segbo: A database of borrowed sounds in the world's languages. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020) (2020), http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.654.pdf
14. Hakimov, N., Backus, A.: Usage-based contact linguistics: Effects of frequency and similarity in language contact. Journal of Language Contact **13**(3), 459 – 481 (2021). https://doi.org/https://doi.org/10.1163/19552629-13030009
15. Ladefoged, P., Maddieson, I.: The Sounds of the World's Languages. Wiley (1996)
16. Lee, N.H.: The status of endangered contact languages of the world. Annual Review of Linguistics **6**(1), 301–318 (2020). https://doi.org/10.1146/annurev-linguistics-011619-030427, https://doi.org/10.1146/annurev-linguistics-011619-030427
17. Macklin-Cordes, J., Round, E.: Re-evaluating phoneme frequencies. Frontiers in Psychology **11** (2020). https://doi.org/10.3389/fpsyg.2020.570895, https://www.frontiersin.org/article/10.3389/fpsyg.2020.570895
18. Meyer, C.M., Gurevych, I.: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. na (2012)
19. Moran, S., McCloy, D. (eds.): PHOIBLE 2.0. Max Planck Institute for the Science of Human History, Jena (2019), https://phoible.org/
20. Munson, B.: Phonological pattern frequency and speech production in adults and children. Journal of speech, language, and hearing research **44**, 778–92 (09 2001). https://doi.org/10.1044/1092-4388(2001/061)
21. Silva, C.: The representation of portuguese palatal sonorants through the eyes of portuguese-based creoles. In: Proceedings of the 8th School-Conference Language issues: a young scholars' perspective. Institute of Linguistics of the Russian Academy of Sciences, Moscow (forc)
22. Smith, I.R.: Sri lanka portuguese structure dataset. In: Michaelis, S.M., Maurer, P., Haspelmath, M., Huber, M. (eds.) Atlas of Pidgin and Creole Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig (2013), https://apics-online.info/contributions/41
23. Thomason, S.G., Kaufman, T.: Language contact, creolization, and genetic linguistics. University of California Press (1988)
24. Trigo, L., Silva, C.: Ptlanka: an online corpus of sri lanka portuguese lexicon and phonology. In: OpenCor 2021 (2021)
25. Trigo, L., Silva, C.: Comparing lexical and usage frequencies of palatal segments in portuguese. In: Proceedings of the 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022). Springer, Fortaleza (forc)
26. Veloso, J.: Complex segments in portuguese: The unbearable heaviness of being palatal. In: Zendoia, I.E., Nazabal, O.J. (eds.) Bihotz ahots. M. L. Oñederra irakaslearen omenez, pp. 513–526. Euskal Herriko Unibertsitatea, The address of the publisher (2019)
27. Wetzels, W.L.: Consoantes palatais como geminadas fonológicas no português brasileiro. Revista de Estudos da Linguagem **9**(2), 5–15 (2000). https://doi.org/10.17851/2237-2083.9.2.5-15, http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/2323