Frequency, overlap and origins of palatal sonorants in three Iberian languages

Carlos Silva
CLUP - Centro de Linguística,
University of Porto /
Porto, Portugal
cssilva@letras.up.pt

Luís Trigo
CODA - Center for Digital Culture and
Innovation, University of Porto
CLUP - Centro de Linguística,
University of Porto /
Porto, Portugal
ltrigo@letras.up.pt

Abstract

The frequency distributions of sounds within languages are closely related to how languages arise and develop over time. Palatal consonants did not exist in Latin, but they flourished in the Romance languages, especially in the Iberian Peninsula. Still, they are considered complex or marked segments because they are inherently heavy and restricted in terms of their distribution, in relation to other consonants. This study correlates intra and interlanguage frequency across three Iberian languages, namely Galician, Portuguese, and Spanish based on a Wiktionary sample. Beyond extracting the frequency values, we calculate the overlap of specific lexical items containing these phonemes. Finally, we assess the relevance of the etymological pathways to the frequency observed in each language using a list of aligned cognates. We find that, in spite of some contamination through contact, the frequencies in synchronic and diachronic data of /ʎ/ and /ɲ/ in Galician match those of Portuguese and not Spanish. These results suggest low-frequency consonants are highly relevant to language classification.

1 Introduction

Galician is a Western Romance language with Portuguese as its closest relative (Alkire and Rosen, 2010). However, it has been noted that Galician has been moving closer to Spanish in the last decades due to intensive language contact and it displays now about the same distance regarding its geographical neighbors (Campos, 2020). This approximation makes it harder to automatically distinguish Galician from both Spanish and Portuguese in text corpora.

English is another language that experienced intensive contact, especially during the Norman invasions (11th century). Despite 85% of the Old English lexicon has been lost and replaced by borrowing from other languages (Baugh and Cable, 1993), (Stockwell and Minkova, 2001), the frequencies of English consonants remained largely the same over time (Martin, 2007). Still, frequent consonants tended to get more frequent over time. Rare consonants are preserved to avoid homophony within a language.

Palatal sonorants are known to display low frequencies in Portuguese across dictionary corpora, namely /ɲ/ 1.7% to 2.5% and /ʎ/ 2.3% to 3.1% (Trigo and Silva, 2022). The global low-frequency values of the palatal sonorants are expected in light of their late acquisition in Portuguese (Costa, 2010), and their low-frequency across the world's languages (Moran and McCloy, 2019). What is not clear so far is why /ʎ/ is more frequent than /ɲ/ as it is acquired later and typologically rarer.

The mismatch between cross-linguistic and language-internal frequency can be explained either by contextual biases or historical sound change (Gordon, 2016). These hypotheses were previously put forward (Trigo and Silva, 2022), but they were not statistically tested yet. This study fills in this gap by analyzing the frequency, overlap, and historical origins of /ʎ/ and /ɲ/ in Galician, Portuguese, and Spanish. Although these languages are related, they display differences concerning the phonotactic restrictions and the historical origins of these consonants (Holt, 2003), (Zampaulo, 2019), as we show in Table 1 and Table 2.

Historical sources	Galician	Portuguese	Spanish
Initial /pl, kl, fl/→ʎ	No	No	Yes
Long /lː/→ʎ	No	No	Yes
/l+i/→ʎ	Yes	Yes	Yes
/kl gl/→ʎ	Yes	Yes	No
Long /nː/→ɲ	No	No	Yes
/gn/→ɲ	Yes	Yes	Yes
/n+i/→ɲ	Yes	Yes	Yes

Table 1: Etymological sources of palatal sonorants in Galician, Portuguese, and Spanish.

Phonotactics	Galician	Portuguese	Spanish
Initial ʎ	No	No	Yes
Intervocalic ʎ	Yes	Yes	Yes
Final ʎ	No	No	No
Initial ɲ	No	No	No
Intervocalic ɲ	Yes	Yes	Yes
Final ɲ	No	No	No

Table 2: Contextual differences of palatal sonorants in Galician, Portuguese, and Spanish.

Galician	Portuguese	Spanish	Latin
orella	orelha	oreja	auricula
ollo	olho	ojo	oculus
ventrullo	barriga	barriga	ventriculum
sobrecella	sobrancelha	ceja	supercilium
unlla	unha	uña	ungulam
...	...	...	...
pulso	pulso	muñeca	pulsus
ano	ano	año	annus
tinxir	tingir	teñir	tingere
constrinxir	constringir	constreñir	constringere
estrinxir	obstipar	estreñir	stringere

Table 3: Extract from the palatal sonorants dataset.

It should be noted that the pathway from Latin to Romance languages follows several steps. For instance, many instances of /kl/, /gl/, or even /gn/ in Proto-Romance result from an earlier vowel syncope, e.g. oculus → *oclus "eye" (Table 3).

2 Methods

We extracted the Galician wiktionary dump (latest on October, 10th) that accounted for 96395 words. From these entries, we selected a sample that included the translation for Portuguese and Spanish as well as the Latin etymology for the Galician word. The resulting subset was composed of 2583 entries. Then we verified that some of the translation and the etymology slots were empty, and there were also some repeated words in the translation entries - i.e. synonyms that would not be helpful for having comparative statistics. Thus, we further filtered the dataset and obtained 2248 entries.

For comparing consonant frequencies, we extracted these phonemes directly from the orthographic entries. This process is straightforward for /ʎ/ (”lh” for Portuguese and ”ll” for Galician and Spanish), /ɲ/ (”nh” for Portuguese and ”ñ” for Galician and Spanish), and /p/ (”p” for all languages). Concerning /m/, we had to use a regex expression to look for 'm' followed by vowels, i.e. string where this phoneme was in the onset position.

In our dataset, 126 entries contained palatal sonorants (Table 3) - 74 for Galician, 77 for Portuguese and 66 for Spanish). Portuguese and Galician share 93% of the Latin cognates while Galician and Spanish share 88%. The missing etymological data was manually filled using data from printed dictionaries. All changes to the original extraction regarding Latin etymology were manually annotated in the dataset.

All the source data and processing python scripts can be found in the repository: https://github.com/Portophon/Gal-palatals.

3 Results

In this section, we perform a visual inspection of the frequency patterns that characterize palatal sonorants in Galician, Portuguese, and Spanish. The relative percentage values are given according to the size of each corpus as described in section 2.

Figure 1 shows the frequency of /ɲ/ and /ʎ/, rare and complex consonants, compared with two near-universal and non-complex consonants /p/ and /m/ in onset position across the three languages. In light of the values obtained for Portuguese in a previous study (Trigo and Silva, 2022), our dataset seems to be representative of the full lexicon of these languages.

We see that the frequency range between Galician and Portuguese is about 0.04 percentage points for /p/ and /m/ and 0.11 percentage points for /ɲ/. Concerning /ʎ/, there is a perfect match between this pair of languages. The difference between Galician and Spanish is not significant regarding /p/ (0.04 percentage points) and /m/ (0.19 percentage points), but it is exacerbated in the palatal sonorants, i.e. for /ʎ/ there is a positive difference of 0.47 percentage points , and for /ɲ/ here is a positive difference of 0.78 percentage points. As a consequence, Spanish becomes an interesting case study as it further increases the mismatch between language-internal and cross-linguistic frequency (Gordon, 2016).

The pink bars of Figure 2 highlight the differences mentioned above. In addition, the green bars show the percentage of correspondence between the specific cognates in the pairs Galician-Portuguese and Galician-Spanish. Overall, we find that there is

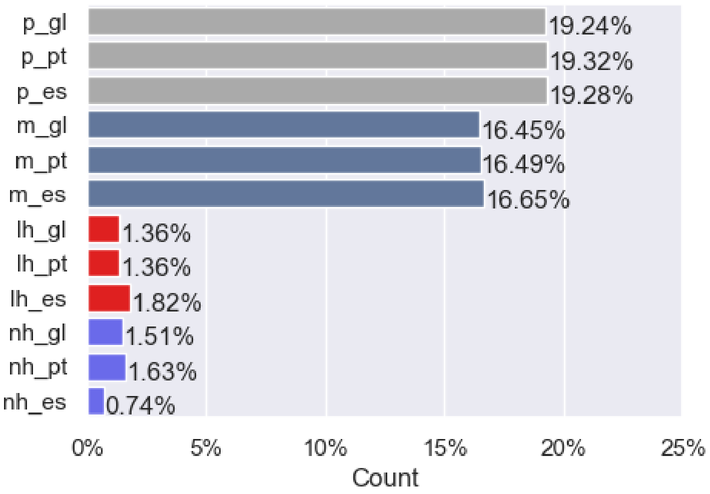
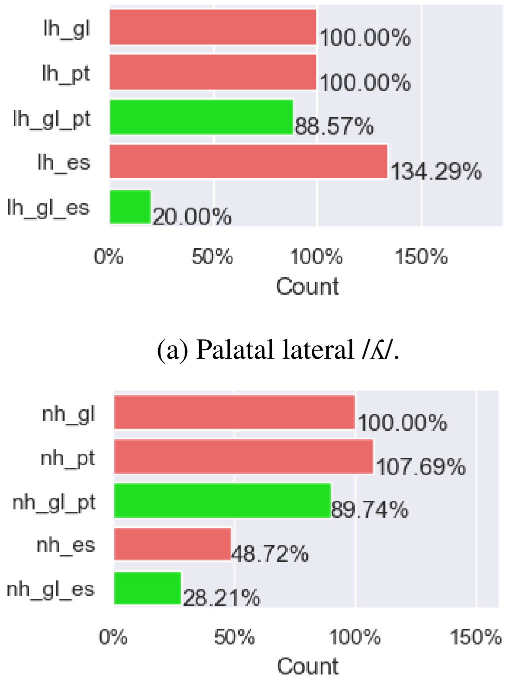


Figure 1: Percentage of palatal sonorants compared to near-universal consonants.

a great correspondence between the lexical items having either /ʎ/ or /ɲ/ in Portuguese and Galician (89%). Comparing Galician and Spanish, the correspondence of items with /ʎ/ is low, even though this consonant is considerably more frequent in Spanish. However, the correspondence of /ɲ/ in the two languages signals some degree of approximation between Galician and Spanish, because the general frequency of items with the nasal palatal is reduced in the latter.

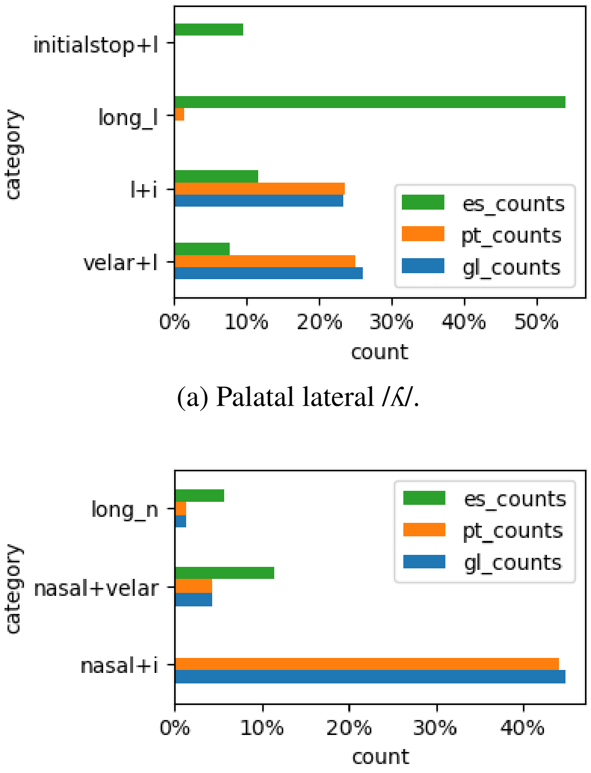
In Figure 3, the values in the x-axis represent the relative percentage of conversion from Latin to Galician, Portuguese, and Spanish with regard to the total number of palatals in each language. Thus, we can visualize the preferred historical pathways for the emergence of each palatal sonorant in the languages of our sample. In line with the literature (Table 1), our data confirms that the palatals of Spanish have different origins from those of Portuguese and Galician, namely initial stop /p k f/ plus /l/ and long /lː/ or /nː/. Latin long /lː/ seems to be a particularly important source of the Spanish /ʎ/. It might explain the greater frequency of this consonant in comparison to Galician and Portuguese which converted the Latin long /lː/ into a plain /l/. The Galician and Portuguese words which have an etymon with a long /lː/ or /nː/ in Latin were likely borrowed through Spanish. Overall, there is symmetry between Portuguese and Galician, and asymmetry between this pair of languages and Spanish.



(a) Palatal lateral /ʎ/.

(b) Palatal nasal /ɲ/.

Figure 2: Relative frequency and overlapping regarding the Galician palatals.



(a) Palatal lateral /ʎ/.

(b) Palatal nasal /ɲ/.

Figure 3: Preferred evolution pathways for palatals.

4 Discussion

The study investigates what drives the frequency of palatal sonorants by analyzing their distribution, overlap, and etymological origin in Galician, Portuguese, and Spanish. In line with previous studies for English (Martin, 2007), we find that the frequency of rare consonants like /ʎ/ and /ɲ/ reflects the phylogenetic signal of each language more faithfully than frequent consonants like /p/ and /m/. This characteristic could be explained by the low borrowability rates of the first pair, i.e. /ɲ/ 1.04 and /ʎ/ 0.99, when compared to the second pair, i.e. /p/ 10.58 and /m/ 3.45 (Grossman et al., 2020). Another explanation, which complements the former, is the preference for highly frequent consonants in new lexicon entering a given language (Stockwell and Minkova, 2001). Consequently, /ʎ/ and /ɲ/ become more associated with the patrimonial lexicon and functional words or morphemes over time.

When considering the palatal sonorants as a whole, they seem to be more complex or marked than other consonants like /p/ and /m/, because they are about ten times less frequent. However, when we observe them individually, we notice that their language-internal frequency does not mirror their cross-linguistic frequency, i.e. /ʎ/ 5% and /ɲ/ 42% (Moran and McCloy, 2019), against what phonological theory predicts (Clements, 2003), (Clements, 2009). In all languages of our sample, /ʎ/ is more frequent than /ɲ/. This difference is exacerbated in Spanish. At first sight, we could propose that /ʎ/ is more frequent than /ɲ/, because /ɲ/ is more restricted in terms of its phonotactics than /ʎ/. However, this explanation would only work for Spanish and it would not explain why this happens in Portuguese and Galician where the same restrictions apply to both /ɲ/ and /ʎ/. Moreover, the initial context of Latin /pl, kl, fl/ does not seem particularly fruitful in the emergence of the Spanish /ʎ/.

Thus, the answer to the question: “What drives the divergence between cross-linguistic and language-internal frequency?” does not lie in contextual biases, but rather in the historical sources of sound change (Table 1). In other words, our data suggests that not only the number of possible pathways but also the frequency of each pathway in the source language (Latin) play a role in boosting (or reducing) the frequency of the palatal sonorants. For instance, the high frequency of long /lː/ in Latin motivates directly the high frequency of /ʎ/ in Spanish, whereas the low frequency of long /nː/ in Latin results in a lower frequency of its nasal counterpart.

The overlap of the lexical items that have a particular consonant (Figure 2) and of the historical pathways (Figure 3) showcases how misleading orthography can be in language detection and classification. Portuguese represents /ʎ/ as <lh> and /ɲ/ as <nh>, while the symbols <ll> and <ñ> as used in Spanish and Galician. Nevertheless, Galician <ll> is closer to Portuguese <lh> on all accounts, i.e. frequency, overlap, and historical origin.

Further investigation should measure the frequency of the historical sources of palatals in the Latin lexicon to have more representative data, and confirm the hypotheses put forward based on Figure 3. Moreover, the measurement of the lexical overlap in more Iberian languages would bring new light to change that is not originated by etymological, but rather through language contact.

Acknowledgements

Carlos Silva is funded by Portuguese Foundation for Science and Technology (FCT) (SFRH/BD/2020.07466.BD) and supported by the Center of Linguistics of the University of Porto (CLUP) (FCT-UIDB/00022/2020).

Luís Trigo is supported by Centre for Digital Culture and Innovation (CODA) funded by FCT, under the CEECINST/00050/2021 contract programme, and also supported by CLUP (FCT-UIDB/00022/2020).

References

Ti Alkire and Carol Rosen. 2010. *Romance Languages: A Historical Introduction*. Cambridge University Press, New York.

Albert Baugh and Thomas Cable. 1993. *A History of the English Language*. Routledge, London.

José Ramom Campos. 2020. *Medidas de distância entre línguas baseadas em corpus: Aplicação à linguística histórica do galego, português, espanhol e inglês*. Ph.D. thesis, Universidad del País Vasco.

George Nick Clements. 2003. Feature economy in sound systems. *Phonology*, 20:287–333.

George Nick Clements. 2009. The role of features in speech sound inventories. In Eric Raimy and Charles Cairns, editors, *Contemporary Views on Architecture and Representations in Phonology*, page 19–68. MIT Press, Cambridge, MA.

Teresa Costa. 2010. *The acquisition of the consonantal system in European Portuguese: focus on place and manner features*. Ph.D. thesis, Universidade de Lisboa.

Matthew Gordon. 2016. *Phonological Typology*. Oxford University Press.

Eitan Grossman, Elad Eisen, Dmitry Nikolaev, and Steven Moran. 2020. 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)*.

Eric Holt. 2003. The emergence of palatal sonorants and alternating diphthongs in Old Spanish. In Eric Holt, editor, *Optimality Theory and Language Change*, pages 285–305. Springer.

Andy Martin. 2007. *The evolving lexicon*. Ph.D. thesis, University of California, Los Angeles.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Robert Stockwell and Donka Minkova. 2001. *English Words: History and Structure*. Cambridge University Press, Cambridge.

Luís Trigo and Carlos Silva. 2022. Comparing lexical and usage frequencies of palatal segments in Portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 353–362, Berlin. Springer.

André Zampaulo. 2019. The historical emergence of Spanish palatal consonants. In Sonia Colina and Fernando Martínez-Gil, editors, *The Routledge Handbook of Spanish Phonology*. Routledge.