A Historical Database for the Study of Obstruent-Lateral Palatalization in Ibero-Romance

Andrea García-Covelo
Institute for Phonetics and Speech Processing, LMU Munich
Schellingstr. 3, 80799 Munich
andrea.garcia@phonetik.uni-muenchen.de

Abstract

Studying irregular sound changes requires documenting not only words that underwent the change but also those that did not. Obstruent-lateral (OL) palatalization in Ibero-Romance, i.e., Galician, Portuguese, and Spanish, is one such change, exhibiting three distinctive patterns: unusual distribution (/pl fl kl/ typically palatalized but /bl gl/ rarely did), irregular implementation (not all eligible words underwent palatalization), and variable outcomes (dependent on obstruent voicing and cluster word position). This paper presents a cross-linguistic historical dataset of 659 inherited words from principally Galician, Portuguese, and Spanish, with and without palatalization, traceable to etyma containing OL clusters. The dataset draws on etymological dictionaries, philological works, and historical corpora. A digitalized version of the *Diccionario Crítico Etimológico Castellano e Hispánico* (Corominas and Pascual, 2012) served as the backbone for systematically identifying etyma containing OL clusters. The compiled corpus contains 473 words with certain etymologies and comparable coverage across the three languages. By providing the first comprehensive compilation of both palatal and non-palatal historical evidence, this dataset enables the systematic study of OL palatalization in Ibero-Romance.

Keywords: dataset, Ibero-Romance, sound change, corpus linguistics

1. Introduction

Understanding the mechanisms and regularity of sound change requires comprehensive documentation, not only of words that underwent a given change, but crucially also of those that did not. Without systematic evidence of both outcomes, it is impossible to determine whether a sound change was regular, to identify the conditioning factors that triggered or blocked it, or to quantify its scope. This is especially relevant for sound changes that appear to have applied irregularly, since the absence of evidence for the change may simply reflect gaps in documentation rather than genuine irregularity.

Obstruent-lateral (OL) palatalization in the Ibero-Romance languages, i.e., Galician, Portuguese, and Spanish, is one example.[1] Latin had five main OL clusters - /pl fl bl kl gl/ - , which could be primary or secondary. Primary OL clusters were etymological, while secondary O(V)L clusters originally contained an unstressed vowel between the first and second member of the cluster (i.e., $C_1(V)C_2$) that was often lost through syncope, e.g., Lat. PLŬVĬA 'rain'[2] vs. Late Lat. PŎPLUS (< Lat. PŎPŬLUS) 'people'. These clusters occurred word-initially, postconsonantally and postvocalically.[3] The major sound change affecting these clusters was OL palatalization, which often rendered them to single palatal or post-alveolar sounds or to sequences of obstruent + palatal segment, e.g., Lat. PLŬVĬA > Gal. [ʧ]*uvia*, Sp. [ʎ] or [ʝ]*uvia*, and It. [pj]*ova* (cf. Zampaulo, 2019).[4]

In the Ibero-Romance languages, OL palatalization presents three features that make it a particularly complex and understudied phenomenon: an unusual distribution (only certain clusters palatalized), variable outcomes depending on obstruent voicing and word position, and apparent irregularity (not all eligible words underwent the change). Despite this complexity, previous research has relied on a limited set of examples, often drawn from a single language, hindering the comprehensive study of this process.

This paper addresses these gaps through a cross-linguistic, corpus-based approach. A historical wordlist was compiled containing over 600 inherited words from principally Galician, Portuguese, and Spanish, both with and without

[1] The following abbreviations are used in this paper: Lat. = Latin, Gal. = Galician, Pt. = Portuguese, (O)Sp. = (Old) Spanish, It. = Italian, Rib. Arag. = Ribagorçan Aragonese, GP = Galician-Portuguese, Cat. = Catalan, O = Obstruent, L = Lateral, C = Consonant, V = Vowel, (V) = syncopated vowel, #_ = word-initial, C_ = postconsonantal, V_ = postvocalic.

[2] All Latin meanings and words were taken from Lewis and Short (1879) and Georges (1998) unless otherwise specified.

[3] Postvocalically, OL clusters could have a short or long obstruent, e.g., Lat. DŬPLĀRE 'to double' vs. APPLAUDĔRE 'to clap'.

[4] This process is labelled palatalization, although not all stages of the cluster development involved palatalization and the results include both palatal and post-alveolar outcomes.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 185–193
11–16 May 2026.

OL palatalization, traceable to etyma containing OL clusters. By documenting non-palatal outcomes alongside palatal ones, the dataset enables the systematic analysis of the conditions, outcomes, and irregularity of OL palatalization in Ibero-Romance.

1.1. OL palatalization in Romance

The distribution of OL palatalization, i.e., which clusters palatalized, follows three main patterns in Romance (Tuttle, 1975; Repetti and Tuttle, 1987; Zampaulo, 2019):

1. All OL clusters were affected, like in Tuscan Italian, Ribagorçan Aragonese, and Franco-Provençal varieties.

2. Only OL clusters with a velar C_1 (i.e., /kl gl/) were targeted, like in Romanian, Northern Abruzzo Italian, and in several Gallo-Romance and French dialects.

3. Only secondary postvocalic O(V)L clusters (e.g., /k(V)l g(V)l/) palatalized, like in Catalan and Standard French.

In addition, the phonological outcomes are usually uniform across word positions in these languages, i.e., the outcome of a given cluster is the same whether it occurs word-initially, postconsonantally, or postvocalically,[5] e.g., Lat. PLANTA 'plant' > Rib. Arag. [pʎ]*anta* and It. [pj]*anta*, Lat. COMPLĒRE 'to fill up,' > Rib. Arag. *cum*[pʎ]*ir* and It. *com*[pj]*ere*, and Lat. DŬPLĀRE 'to double' > Rib. Arag. *do*[bʎ]*ar* and It. *do*[p:j]*are* (Tuttle, 1975; Viudas Camarasa, 1979; Repetti and Tuttle, 1987). While exceptions to OL palatalization exist in these varieties, non-palatalized words can be explained as (late) borrowings.

1.2. OL palatalization in Ibero-Romance

In contrast to the patterns described for Romance, the Ibero-Romance languages show an unusual development of OL palatalization in three respects: distribution, outcomes, and regularity.

In Galician, Portuguese, and Spanish, OL clusters with voiceless obstruents (/pl fl kl/) underwent palatalization, while those with voiced obstruents (/bl gl/) rarely did. Moreover, the outcomes of /pl fl kl/ merged. Table 1 illustrates both properties (de Andrés Díaz, 2013: 235; González González, 2012; Real Academia Española, 2014; Houaiss et al., 2004).

[5] Secondary postvocalic O(V)L clusters can be an exception, i.e., Lat. DŬPLĀRE > Rib. Arag. *do*[bʎ]*ar* but Lat. ŏvĭcŭla 'little sheep' > Rib. Arag. *ue*[ʎ]*a* (Dirección General de Política Lingüística de Aragón, 2025).

Lat.	Gal.	Pt.	Sp.	Palat.
PLŌRĀRE 'to weep'	[ʧ]*orar*	[ʃ]*orar*	[ʎ]*orar*	Yes
FLAMMA 'flame'	[ʧ]*ama*	[ʃ]*ama*	[ʎ]*ama*	Yes
CLĀVIS 'key'	[ʧ]*ave*	[ʃ]*ave*	[ʎ]*ave*	Yes
BLANDUS 'flattering'	[br]*ando*	[br]*ando*	[bl]*ando*	No
GLANS, -DIS 'acorn'	[l]*andra*	[l]*ande*	[l]*andre*	No

Table 1: Distribution of OL palatalization in Ibero-Romance (word-initial clusters)

Unlike the uniform cross-positional outcomes found in other Romance varieties, the outcomes of OL palatalization in Ibero-Romance vary by word position, especially in Spanish: word-initially [ʎ] (Lat. CLĀMĀRE 'to call, cry out' > Sp. [ʎ]*ama*), postconsonantally [ʧ] (Lat. ĭNFLĀRE 'to blow into' > Sp. *hin*[ʧ]*ar*), and postvocalically [x] (Lat. ŏcŭlus 'eye' > Sp. *o*[x]*o*, OSp. *o*[ʒ]*o*) ([ʎ] if the cluster contains a geminate obstruent, e.g., Lat. APPLĬCĀRE 'to join, to attach' > Sp. *a*[ʎ]*egar*).

Lastly, the regularity of OL palatalization in Ibero-Romance has been questioned, not only due to its unusual distribution but also due to its absence in many inherited words. While some non-palatal results can be explained as borrowings, many cannot. Table 2 shows pairs of etyma with the same cluster where palatalization applied in one case but not in the other.

Lat.	Gal.	Pt.	Sp.	Palat.
CLĀVIS 'key'	[ʧ]*ave*	[ʃ]*ave*	[ʎ]*ave*	Yes
CLĀVUS 'nail'	[kr]*avo*	[kr]*avo*	[kl]*avo*	No
CUNĬCŬLUS 'rabbit'	*coe*[ʎ]*o*	*coe*[ʎ]*o*	*cone*[x]*o*	Yes
PĔRĪCŬLUM 'danger'	*peri*[g]*o*	*peri*[g]*o*	*peli*[gr]*o*	No

Table 2: Irregular application of OL palatalization in Ibero-Romance

1.3. The Documentation Gap in OL Palatalization Research

A thorough historical foundation is necessary for the quantitative and qualitative study of OL palatalization in Ibero-Romance, including its unusual distribution, irregularity, and outcome diversity. However, previous research has relied on a limited set of examples, often from a single language (cf. Repetti and Tuttle, 1987; Wireback, 1997; Mariño Paz, 2017; Zampaulo, 2019). Moreover, ex-

isting studies have focused almost exclusively on words that did undergo palatalization, while non-palatal outcomes, essential for assessing the regularity and conditioning of the process, have received only superficial treatment and have rarely been analysed alongside palatal results.

As a result, the documentation of OL palatalization is uneven across cluster types and positions. Whereas certain clusters exhibit clear and well-attested diachronic pathways for palatalization, e.g., primary word-initial and postconsonantal /pl kl fl/ and secondary postvocalic /k(V)l/, others show no documented traces of this sound change, such as primary postvocalic /pl fl kl bl gl/ or postconsonantal /bl gl/. Similarly, OL palatalization is attested in primary postvocalic /p:l f:l/, but not in postvocalic /k:l/. Part of this gap is due to the fact that some clusters were less frequent than others in Latin and are consequently poorly documented.

The absence of historical evidence hinders the thorough study of the development of this palatalization process. Two questions remain open: first, whether OL palatalization is genuinely unattested in certain clusters or simply not yet identified; and second, whether the process as a whole was truly irregular in Ibero-Romance.

This cross-linguistic, corpus-based study addresses these gaps by collecting inherited words, with and without OL palatalization, in principally Galician, Spanish, and Portuguese that may stem from etyma containing OL clusters. By including both palatal and non-palatal outcomes, the dataset provides the first comprehensive source of historical evidence to survey the conditions and outcomes of OL palatalization in Ibero-Romance, quantify its irregularity, and identify the factors or sound changes that might have interfered with this process.

2. Methodology

2.1. Primary Source: DCECH

Many of the sources consulted, especially dictionaries, are not accessible in digital form, and even when a digital version is available, the search engine may be too limited to search for OL clusters in all word positions. A systematic method to identify potential etyma containing these clusters was therefore necessary. For this purpose, the *Diccionario Crítico Etimológico Castellano e Hispánico* or *DCECH* (Corominas and Pascual, 2012) was selected as the primary source of this dataset for two reasons: it is the most comprehensive etymological dictionary covering the Ibero-Romance territory, providing etymological and historical information not only on Spanish but also on Galician and Portuguese; and it had an electronic version, making its digitalization feasible.

While the DCECH served as the main guide for compiling a list of candidate etyma, the etymological information it provided was verified and complemented with other sources (see following section), and each etymon was systematically searched in medieval Galician and Portuguese resources. Since many of the available historical sources, particularly corpora and databases, document the medieval Galician-Portuguese period (see Section 2.4), these resources provide evidence relevant to both Galician and Portuguese. This approach is the main reason why the number of tokens for the three languages is comparable (cf. Section 3). To optimise the search process and improve reproducibility, the DCECH was converted into a searchable database.

The *DCECH* already has an electronic version, but its search engine only allows character searches at the start and end of words, making it impossible to search for postconsonantal or postvocalic clusters, e.g., Lat. DŬPLĀRE or Lat. COMPLĒRE. Moreover, each OL cluster would need to be searched individually and manually. To optimise the search process, the content of the electronic dictionary, stored as rtf files (one per entry), was extracted through a Python script and structured into categories: word, etymology, first documentation, general information, derivational forms, compound forms, and notes. The structured information was transformed into a data frame (.csv), where the etymology column was searched for words containing primary or secondary OL clusters, indicated by the characters <ptkbdgf(u)l>. Words without etymological information, e.g., verbal forms, were excluded, and the remaining results were manually reviewed to discard borrowings and verified against other sources. The criteria for both the OL cluster search and included words are outlined in Section 2.3.

2.2. Further Consulted Sources

Multiple sources from different languages and media types were employed for the corpus construction. The principal materials consulted include:

- Etymological dictionaries, e.g., Meyer-Lübke (1935), Corominas and Pascual (2012), García de Diego and García de Diego (1989), Roberts (2014), Rivas Quintas (2015), and Machado Filho (2013).
- Monolingual dictionaries, e.g., Real Academia Española (2014), Houaiss et al. (2004), Lewis and Short (1879), Georges (1998).
- Philological works and databases, e.g., Pensado Ruiz (1984), Mariño Paz (2017), de An-

drés Díaz (2013), Zampaulo (2019), Tuttle (1975), González Seoane et al. (2006), Álvarez (2014), and Real Academia Española (2021).

- Online historical corpora, e.g., Varela Barreiro (2004), Real Academia Española, Carracedo Fraga (2024).

An automated protocol (Python) was also developed to systematically search and extract data from the TMILG (Varela Barreiro, 2004). While this protocol was tested and is functional, it was not employed in the construction of the present dataset. Both the protocol and the DCECH digitalization scripts are available at https://github.com/agcovelo/linguistic-resource-processing.

2.3. Word selection

2.3.1. Primary and Secondary OL Clusters

Etyma containing primary OL clusters were directly included in the wordlist. In contrast, secondary O(V)L clusters originally had an unstressed vowel between C_1 and C_2 that was later lost through syncope, e.g., Lat. ˈ**ŏ**.cŭ.lus > oclus > Sp. *ojo*, Gal. *ollo*.[6] Due to time constraints, only secondary O(V)L clusters meeting the conditions most favourable for syncope were included.

Based on previous descriptions of syncope (Loporcaro, 2010; Mariño Paz, 2017; de Andrés Díaz, 2013; Penny, 2002; Williams, 1962), only secondary O(V)L clusters where the vowel between C_1 and C_2 met the following conditions were included:

- It was unstressed since stressed vowels resisted deletion.
- It was in an open syllable, i.e., a syllable with a short vowel nucleus and without a syllable coda. If the lateral had been followed by another consonant, i.e., -O(V)LC-, syncope would have produced clusters violating Latin phonotactic constraints (Lloyd, 1970; Weiss, 2009; Loporcaro, 2010).
- It was in an intertonic position, i.e., a word-medial syllable between the initial and stressed syllable or between the stressed and final syllable (Penny, 2002: 46; Williams, 1962: 51; Loporcaro, 2010: 59; and Mariño Paz, 2017: 195), where syncope was most common.[7]
- It was /u/, which was the most frequent vowel in O(V)L clusters due to the productivity of Latin suffixes such as *-cul-* and *-bul-*.
- Additionally, C_1 was an obstruent (/p t k b g d f/), since combinations of /n m r/ + /l/ would not have triggered OL palatalization.

Words were excluded if the secondary O(V)L cluster was followed by a palatal glide (e.g., /k(V)lj/), because both Lat. /lj/ and Lat. /k(V)l g(V)l/ yielded the same outcomes, i.e., OSp. [ʎ] > Sp. /x/ and GP [ʎ] (Zampaulo, 2019; de Andrés Díaz, 2013). Consequently, it is not possible to determine whether a palatal result stemmed from OL palatalization or from /lj/.

2.3.2. Borrowings

Many words lacking OL palatalization are Latin borrowings, i.e., words introduced into Spanish, Galician, or Portuguese through written or learned transmission rather than oral inheritance (Penny, 2002: 59). Borrowings did not undergo the regular sound changes that affected inherited vocabulary, such as glide metathesis, stop voicing and deletion, and different vowel changes (cf. de Andrés Díaz, 2013; Mariño Paz, 2017; Penny, 2002; Pensado Ruiz, 1984; Lloyd, 1987). The absence of these early sound changes therefore serves as a diagnostic for identifying learned words, e.g., Lat. applaudĕre 'to clap' > Sp. Gal. Pt. *aplaudir*.

The lack of OL palatalization in a given word can have two explanations: either the word was borrowed after OL palatalization had ceased to be productive, or it entered the language as a literary form with restricted usage during or before the period of OL palatalization. In both cases, borrowings were generally excluded from the wordlist. Words were classified as borrowings based on the absence of early sound changes and on documentation in the consulted sources. As a general threshold, words first attested after the 15th century and lacking early sound changes were treated as borrowings, since OL palatalization was widely attested in Galician-Portuguese and Old Spanish from the 8th–13th centuries (cf. Mariño Paz, 2017; Torreblanca, 1990). Older literary words, e.g., documented in the 13th century and exhibiting some of those early sound changes but not others, were included for reference and comparative purposes. In ambiguous cases, the wordlist marks the etymology as uncertain or flags the word as a possible borrowing rather than making a definitive classification.

Inter-Romance borrowings, i.e., words borrowed from neighbouring Romance varieties, were generally excluded for two reasons. First, if an inherited word shows OL palatalization, this change

[6] The stressed syllable is given in bold (also marked by ˈ), periods indicate syllable boundaries.

[7] In contrast, word-initial vowels were highly resistant to deletion (Loporcaro, 2010: 59; Mariño Paz, 2017: 195).

may have taken place in the source variety prior to borrowing; consequently, the palatalization is not evidence of the process in the borrowing language.[8] Second, inter-Romance loans were typically introduced too late to undergo OL palatalization in the target language. However, early inter-Romance loans were included if they exhibited sound changes contemporaneous with OL palatalization, such as lenition. The first literary attestations of distinct Romance varieties appear in the 9th century (cf. Weiss, 2009: 504-8), when both lenition and OL palatalization were already underway (Mariño Paz, 2017: 358-78; Pensado Ruiz, 1984: 177-223; Lloyd, 1987: 228; Menéndez Pidal, 1964: 274-6; Mariño Paz, 2017: 330-54; García-Covelo, 2025). Therefore, if an inter-Romance loan shows evidence of lenition, it likely entered the language early enough to have potentially undergone OL palatalization as well.

2.4. Linguistic Considerations

The goal of this dataset is to document specific sound changes within linguistic territories; therefore, the precise linguistic classification of a given variety is not critical. Several conventions were adopted accordingly.

Galician and Portuguese share a common linguistic stage known as Galician-Portuguese, spoken approximately between the 9th and 15th centuries (Mariño Paz, 2008; Maia, 1986). Since inherited words and their earliest attestations often coincide for both varieties, words attested before 1500 are treated as Galician-Portuguese and counted as entries for both languages to avoid redundancy. Attestations of Castilian before 1500 are labelled Old Spanish, and those from 1500 onward are labelled Spanish.

Texts written in "Romance Latin", i.e., the variety of Latin used in a particular region and influenced by contemporary vernacular languages, are not differentiated from Galician-Portuguese or Old Spanish. Similarly, the terms "Vulgar Latin" and "Late Latin" are not distinguished in this dataset and are uniformly referred to as "Late Latin."

Inherited words from Ribagorçan Aragonese were occasionally included because this variety exhibits regular OL palatalization with distinct outcomes, making it valuable for reconstructing the evolutionary stages of the process. Words from Asturleonese varieties, which may also have distinct outcomes, were included when they provided the only available evidence of palatalization for a given etymon. These and other regional or geographical varieties are given their own labels following the conventions used in the DCECH (Corominas and Pascual, 2012) and the *Romanisches etymologisches Wörterbuch* (Meyer-Lübke, 1935).

2.5. Phonological Derivation

After reviewing the words originally containing OL clusters extracted from DCECH and other sources, no evidence of OL palatalization was found in some clusters, e.g., /bl/, /gl/, or postvocalic /kl/. To search for possible unattested outcomes, phonological derivation was employed: the systematic reconstruction of plausible inherited forms in descendant languages by applying well-documented sound changes to Latin etyma.

To make this process manageable, Latin etyma were grouped into word families sharing the same root or stem. Only the root of each family was derived, rather than every individual etymon, reducing the risk of producing inaccurate forms. This is an important consideration given that many sound changes, such as syncope, vowel raising or voiced stop deletion, applied irregularly across the Ibero-Romance lexicon (cf. Penny, 2002; Mariño Paz, 2017; de Andrés Díaz, 2013). The derivations were then searched in several historical corpora and databases[9] using simple regular expressions, mainly character sequences with quantifiers such as * and ?.

For example, Lat. CLAMĀRE and its derivational forms, such as CLĀMĬTĀTĬO, ACCLĀMARE, CONCLĀMĀTUS, RĒCLĀMO (cf. Lewis and Short, 1879), share the root *-clam-*. The derivation of this root was guided by attested outcomes: /kl/ predominantly became /ʧ/ in Galician-Portuguese regardless of word position, while Old Spanish shows /ʎ/ word-initially and with geminates, and /ʧ/ postconsonantally (cf. Mariño Paz, 2017: 329-31; de Andrés Díaz, 2013), e.g., Lat. CLĀVIS > Gal. Pt. *chave* and Sp. *llave* (cf. Table 1), and Lat. MASCŬLUS 'male, masculine' > Gal. Pt. Sp. *macho*. Therefore, *-clam-* was derived as *-/ʧ/am-* for Galician-Portuguese and as either */ʎ/am-* or *-/ʧ/am-* for Old Spanish, depending on position. For clusters where OL palatalization is unattested, e.g., postvocalic /k:l/ or /kl/, derivations were based on outcomes from related clusters or from other palatalization processes, yielding additional candidates such as /ʎ/, /ʒ/, or /Oʎ/ (for further information, see García-Covelo, 2025: 20-2).

Once derived, the reconstructed roots were matched to their probable orthographic represen-

[8] For example, Lat. SPECULARIA 'window panes' > Cat. *espillera* -> Sp. *aspillera* (Corominas and Pascual, 2012: aspillera), where the palatal outcome originated in Catalan, not in Spanish.

[9] The main corpora consulted are Varela Barreiro (2004), González Seoane et al. (2006), Carracedo Fraga (2024), Santamarina et al. (2018), and Real Academia Española.

tations in medieval manuscripts, e.g., /ʧ/ as <ch>, /ʎ/ as <ll, lj, l, lh>, /ʒ/ as <j, g> (Mariño Paz, 2017; Maia, 1986; Echenique Elizondo and Martínez Alcalde, 2011: 81-4), and searched in historical corpora using regular expressions. Table 3 illustrates this process for /gl/ in Galician-Portuguese by word position,[10] showing each derived outcome, its possible graphemic representations, and the corresponding search regular expressions.[11] The

Outcome	Position	Graphemes	RegEx
/ʧ/	#_, V_	<ch>	*ch*
/ʎ/	#_, V_	<li ly lj ll lh>	*l?*
/ɲ ɲʎ/	C_	<nll nlh nli nlj nly nn nh ñ>	*nl?*, *ñ*, *n?*
/ʒ/	#_, V_	<j y i g>	*?*, *j*, *i*, *y*, *g*

Table 3: Phonological derivation of the outcomes for /gl/ in Galician-Portuguese

phonological derivation and corpus search was time-intensive; consequently, it was prioritized for the clusters with the least existing evidence of OL palatalization, namely /bl/, /gl/, and postvocalic /k:l kl/.

2.6. Dataset Structure

The dataset resulting from this study contains etymological information for inherited words in mainly Galician, Portuguese, and Spanish that stem from etyma with primary or secondary OL clusters.

The current dataset design includes the following columns: the etymon, its corresponding root or stem, and its language of origin. For the OL cluster itself, there are columns indicating whether it is primary or secondary, the specific cluster, and its position within the word. For the inherited forms, the dataset includes columns for the specific lexical items, secondary forms or spelling variants, the language variety of origin, the date of first attestation, the phonetic outcome, the sound changes that affected the original OL cluster, and whether the form exhibits OL palatalization. Additionally, there are columns documenting the status of the etymology - specifically, whether the etymology is certain and whether the inherited word is a borrowing - as well as the sources consulted and existing scholarly debate. The dataset is available online at Zenodo (García-Covelo, 2026).

3. Results

The compiled corpus has a total of 659 inherited words, of which 578 are Galician, Portuguese (or Galician-Portuguese) and (Old) Spanish.[12] This number becomes 473 after excluding inherited words with unknown or uncertain origins, (possible) borrowings and words that did not come from OL clusters. The dataset includes a comparable number of Galician (n = 227), Portuguese (n = 219)[13] and Spanish (n = 223) lexical items and a comparable number of primary OL clusters (n = 219) and secondary O(V)L clusters (n = 254).

Table 4 shows the number of inherited words for the five main clusters, grouped by cluster type (primary or secondary) and the presence or absence of palatalization. Tokens with an uncertain palatalization status (n = 5) are excluded. Each row in the table includes all possible cluster configurations, i.e., the tokens listed under /kl/ represent /kl k:l/ and those under /k(V)l/ represent /k(V)l k:(V)l/. The number of instances of a particular OL cluster varies considerably depending on its frequency in the original language: for instance, inherited words with secondary /k(V)l/ are far more numerous than those with primary /kl/, and there are no lexical items from etyma with /f(V)l/.

Cluster	Palatal.	Non-palatal.	Total
/pl/	55	27	83
/p(V)l/	3	11	14
/fl/	16	16	32
/f(V)l/	0	0	0
/kl/	27	23	52
/k(V)l/	121	21	142
/bl/	1	25	27
/b(V)l/	5	27	33
/gl/	1	22	23
/g(V)l/	27	4	31

Table 4: Number of inherited words with (palatal.) and without (non-palatal.) OL palatalization per cluster grouped by primary and secondary type

The distribution of palatalized and non-palatalized words is consistent with the patterns described for Galician, Portuguese, and Spanish: /pl fl kl/ usually palatalized while /bl gl/ rarely did. For the etyma with /pl fl kl/, between 50% and 76% underwent palatalization, compared to only

[10] word-initial (#_), postconsonantal (C_), and postvocalic (V_).

[11] The required regular expressions depended on the resource used: for instance, the search engine is more flexible in TMILG (Varela Barreiro, 2004) than in CORDE (Real Academia Española) because TMILG allows for grapheme equivalences (<i> = <i j y h>) and CORDE does not.

[12] The analysis code for this section is available at https://github.com/agcovelo/lrec-2026-historical-database.

[13] Galician-Portuguese words were counted as both Galician and Portuguese.

10% and 52% for /bl gl/. Moreover, palatalization affected primary and secondary clusters differently, illustrating the unusual development of this sound change in Ibero-Romance: primary /pl fl/ typically palatalized while secondary /p(V)l f(V)l/ did not; conversely, /g(V)l/ usually palatalized but /gl/ did not.

The dataset does not include all inherited words from etyma containing OL clusters, even though most primary clusters and the most common inherited words are included. However, as Table 4 shows, the number of tokens is sufficient to investigate the implementation of palatalization in OL clusters and, therefore, sheds new light on possible factors that blocked this change. A full analysis of this dataset, including newly identified evidence and a discussion of the factors conditioning OL palatalization, is presented in García-Covelo (2025).

4. Limitations and Future Directions

Limitations of this historical dataset include: etymological subjectivity, limited language sample, research-specific application, and database format.

Etymological research often involves debate over competing reconstruction hypotheses. Since this dataset investigates the development between an etymon and its descendant, the etymological relation cannot be omitted. However, the corpus follows established sources and current consensus, taking a conservative approach by marking etymologies as uncertain when strong debate exists. Information on contested or tentative connections is provided in the dataset.

The etymologies belong to lexical items in three main languages: Galician, Portuguese, and Spanish (and their older stages). While the dataset occasionally includes inherited words from smaller linguistic varieties, e.g., Asturian, Leonese, and Aragonese, these represent a minority. Therefore, this dataset allows for both the analysis of general Ibero-Romance patterns and thorough study of OL palatalization in the three main languages. However, accurate description of this process in other linguistic varieties cannot be achieved without expanding the corpus.

In this regard, the dataset was constructed specifically to investigate OL palatalization, which heavily influenced selection criteria for the lexical items and etyma included and the structure of the dataset. Therefore, this corpus may be inadequate for the study of other sound changes.

Lastly, the wordlist was created as a spreadsheet containing data redundancy, e.g., repetition of language labels and OL cluster categorizations. While functional for the immediate research purposes, this format is not optimal for other applications or for updating and maintaining the dataset.

Efforts are currently underway to address several of these limitations. The dataset can be effectively filtered and analysed in R; however, the high level of data redundancy requires database normalization. This restructuring will enable more efficient addition and modification of information while reducing errors in the wordlist.

5. Ethical Considerations

This work should present minimal ethical concerns as it involves historical linguistic data from documented languages. The corpus is available at Zenodo (García-Covelo, 2026) and relevant scripts are at `https://github.com/agcovelo`.

6. Acknowledgments

The research leading to the construction of this database was funded by the German Academic Scholarship Foundation and by the Elite Network of Bavaria's Marianne-Plehn-Program. This paper is based on work carried out as part of a doctoral thesis (García-Covelo, 2025).

7. Bibliographical References

Ramón de Andrés Díaz. 2013. *Gramática comparada de las lenguas ibéricas*. Biblioteconomía y administración cultural. Trea, Gijón.

María Teresa Echenique Elizondo and María José Martínez Alcalde. 2011. *Diacronía y gramática histórica de la lengua española*, 4a ed. rev. y act edition. Prosopopeya. Tirant Humanidades, Valencia.

Andrea García-Covelo. 2025. *The development of obstruent plus lateral clusters in Ibero-Romance: a historical-phonetic approach to cluster palatalization*. Ph.D. thesis, LMU München - UPPA Pau.

Paul M. Lloyd. 1970. A note on Latin syllable structure. *Classical Philology*, 65(1):41–42.

Paul M. Lloyd. 1987. *From Latin to Spanish*. Number v. 173 in Memoirs of the American Philosophical Society. American Philosophical Society, Philadelphia, Pa.

Michele Loporcaro. 2010. Syllable, segment and prosody. In Martin Maiden, John Charles Smith, and Adam Ledgeway, editors, *The Cambridge History of the Romance Languages*, 1 edition, pages 50–108. Cambridge University Press.

Clarinda de Azevedo (ed.) Maia. 1986. *História do galego-português: estado linguístico da Galiza e do Noroeste de Portugal desde o século XIII ao século XVI: (com referência à situação do galego moderno)*. I.N.I.C, Coimbra.

Ramón Mariño Paz. 2008. *Historia de la lengua gallega*. Number 58 in LINCOM studies in Romance linguistics. Lincom Europa, Muenchen.

Ramón Mariño Paz. 2017. *Fonética e fonoloxía históricas da lingua galega*, 1a. edición edition. Manuais de lingua galega. Edicións Xerais de Galicia, Vigo [España].

Ramón Menéndez Pidal. 1964. *Orígenes del español: Estado lingüístico de la Península Ibérica hasta el siglo XI*, 5 edition. Espasa-Calpe, Madrid. Según la 3. muy corregida y adicionada.

Ralph J Penny. 2002. *A history of the Spanish language*. Cambridge University Press, Cambridge.

Carmen Pensado Ruiz. 1984. *Cronología relativa del castellano*. Number 158 in Acta Salmanticensia. Ediciones Universidad de Salamanca, Salamanca.

Lori Repetti and Edward F. Tuttle. 1987. The evolution of Latin pl, bl, fl and cl, gl in Western Romance. *Studi Mediolatini e Volgari*, 33:53–115.

Máximo Torreblanca. 1990. La evolución /kl-, pl-, fl- / > *ll* en español. *Revista de Filología Española*, 70(3/4):317–327.

Edward F. Tuttle. 1975. The Development of PL, BL, and FL in Italo-Romance: Distinctive Features and Geolinguistic Patterns. *Revue de Linguistique Romane*, 39:400–431.

Antonio Viudas Camarasa. 1979. Sobre la evolución de 'pl-' a 'pll-' y 'cl-' a 'cll-' en aragonés antiguo. *Anuario de estudios filológicos*, 2:355–375.

Michael L. Weiss. 2009. *Outline of the historical and comparative grammar of Latin*. Beech Stave Press, Ann Arbor.

Edwin B Williams. 1962. *From Latin to Portuguese: historical phonology and morphology of the Portuguese language*. University of Pennsylvania Press.

Kenneth J. Wireback. 1997. *The Role of Phonological Structure in Sound Change from Latin to Spanish and Portuguese*. Number v. 215 in American university studies. P. Lang, New York.

André Zampaulo. 2019. *Palatal sound change in the Romance languages: diachronic and synchronic perspectives*, 1 edition. Oxford University Press.

8. Language Resource References

Carracedo Fraga, Xosé. 2024. *Corpus Documentale Latinum Gallaeciae CODOLGA*. Centro Ramón Piñeiro para a Investigación en Humanidades, version 21.

Corominas, Joan and Pascual, José Antonio. 2012. *Diccionario crítico etimológico castellano e hispánico*. Gredos, CD-ROM.

Dirección General de Política Lingüística de Aragón. 2025. *Aragonario. Diccionario castellano/aragonés, aragonés/castellano*. Gobierno de Aragón, version 7. Online dictionary, continuously updated. Part of the POCTEFA-LINGUATEC project.

García-Covelo, Andrea. 2026. *Historico-etymological dataset on the development of obstruent-lateral clusters in Ibero-Romance*. Zenodo, version 1.0.

García de Diego, Vicente and García de Diego, Carmen. 1989. *Diccionario etimológico español e hispánico*. Espasa-Calpe, 3rd ed.

Georges, Karl Ernst. 1998. *Ausführliches lateinisch-deutsches Handwörterbuch: aus den Quellen zusammengetragen und mit besonderer Bezugnahme auf Synonymik und Antiquitäten unter Berücksichtigung der besten Hilfsmittel*. Wissenschaftliche Buchgesellschaft, 8th ed. Henricus - Edition Deutsche Klassik GmbH.

González González, Manuel. 2012. *Dicionario da Real Academia Galega*. Real Academia Galega, online version. Continuously updated.

González Seoane, Ernesto and Álvarez de la Granja, María and Boullón Agrelo, Ana I. 2006. *Dicionario de dicionarios do galego medieval. Corpus lexicográfico medieval da lingua galega*. Seminario de Lingüística Informática - TALG/Instituto da Lingua Galega.

Houaiss, Antônio and Villar, Mauro and Franco, Francisco Manoel de Mello. 2004. *Dicionário eletronico Houaiss da língua portuguesa*. Ed. Objetiva.

Lewis, Charlton T. and Short, Charles. 1879. *A Latin dictionary. Founded on Andrews' edition of Freund's Latin dictionary. revised, enlarged, and*

in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short, LL.D. Clarendon Press, Perseus Digital Library.

Machado Filho, Américo Venâncio Lopes. 2013. *Dicionário etimológico do português arcaico*. Edufba.

Meyer-Lübke, Wilhelm. 1935. *Romanisches etymologisches Wörterbuch*. Winter, number 3 in Sammlung romanischer Elementar- und Handbücher Reihe 3, Wörterbücher, 3rd ed.

Real Academia Española. *Banco de datos (CORDE) (en línea). Corpus diacrónico del español*. Real Academia Española, online version.

Real Academia Española. 2014. *Diccionario de la lengua española*. Real Academia Española, 23rd ed., version 23.8 online. Continuously updated online edition.

Real Academia Española. 2021. *Tesoro de los diccionarios históricos de la lengua española*. Real Academia Española.

Rivas Quintas, Eligio. 2015. *Dicionario etimolóxico da lingua galega*. Tórculo.

Roberts, Edward A. 2014. *A comprehensive etymological dictionary of the Spanish language: with families of words based on Indo-European roots*. Xlibris.

Santamarina, Antón and González Seoane, Ernesto and Álvarez de la Granja, María. 2018. *Tesouro informatizado da lingua galega*. Instituto da Lingua Galega, Universidade de Santiago de Compostela, version 4.1.

Varela Barreiro, Xavier. 2004. *Tesouro medieval informatizado da lingua galega*. Instituto da Lingua Galega, Universidade de Santiago de Compostela.

Álvarez, Rosario. 2014. *Tesouro do léxico patrimonial galego e portugués*. Instituto da Lingua Galega.