A methodology to measure the diachronic language distance between three languages based on perplexity

José Ramom Pichel [1], Pablo Gamallo [2], Iñaki Alegria [3], Marco Neves [4]

[1]imaxin software, Santiago de Compostela, Galiza; [2] CITIUS, Santiago de Compostela, Galiza; [3] University of Basque Country, Donostia-San Sebastián, Basque Country; [4] Universidade Nova de Lisboa, Lisboa, Portugal

ARTICLE HISTORY
Compiled February 15, 2020

ABSTRACT

The aim of this paper is to apply a corpus-based methodology, based on the measure of perplexity, to automatically calculate the cross-lingual language distance between historical periods of three languages. The three historical corpora have been constructed and collected with the closest spelling to the original on a balanced basis of fiction and nonfiction. This methodology has been applied to measure the historical distance of Galician with respect to Portuguese and Spanish, from the Middle Ages to the end of the 20th century, both in original spelling and automatically transcribed spelling. The quantitative results are contrasted with hypotheses extracted from experts in historical linguistics. Results show that Galician and Portuguese are varieties of the same language in the Middle Ages and that Galician converges and diverges with Portuguese and Spanish since the last period of the 19th century. In this process, orthography plays a relevant role. It should be pointed out that the method is unsupervised and can be applied to other languages.

KEYWORDS
Language Distance; Historical Linguistics; Perplexity

1. Introduction

Throughout history, languages undergo changes in their phonetics, phonology, morphology, lexicon, syntax, semantics, and even pragmatics. In addition, according to Kloss, Heinz (1967), languages can be divided into two categories regarding their relationship with others: languages by distance (called *Abstand*), which are separated by a significant linguistic distance, and languages by elaboration (*Ausbau*), which are so close to each other that an arbitrary boundary is imposed between them. For all these reasons, measuring the synchronic and diachronic language distances are challenging.

Different descriptive, statistical or corpus-driven methodologies have been developed in the fields of dialectology, phylogenetics, sociolinguistics or natural language processing to measure the intralingual and cross-lingual language distance.

In our previous research, we created perplexity-based methodologies to measure the synchronic distance between European *Abstand* and *Ausbau* languages (Gamallo, Pichel, and Alegria, 2017), to quantify the intralingual diachronic language distance between three languages, one *Abstand* (English) in relation to the others, which have an *Ausbau* relationship (Portuguese and Spanish) (Pichel, Gamallo, and Alegria, 2018, 2019b), and finally to measure the cross-lingual diachronic distance between two historical *Ausbau* languages: Portuguese and Spanish (Pichel, Gamallo, and Alegria, 2019a).

As our methodology is able to detect changes in trends in the distance between languages over time, it may serve to measure the distance between very close *Ausbau* languages and to trace the historical development of their conflicting elaboration. By observing the historical elaboration of very close languages, we can confirm consolidated linguistic hypotheses about when they come closer to each other, being perceived as varieties, and when they separate. In addition, our methodology helps to clarify not only consolidated hypotheses but also controversial claims, which may shed more light on the relationship between very close languages in the process of elaboration. Since orthography also plays an important role in the *Ausbau* language development process, we will also measure language distance by taking this variable into account.

The main goal of the present article is to apply the perplexity-based measure methodology to measure the diachronic language distance among historical periods of three related *Ausbau* languages (Portuguese, Galician and Spanish), by focusing on the movements of approximation and separation of the Galician language with respect to the other two languages. For this purpose, two types of diachronic distances will be measured: the intralingual distance between diachronic varieties within the same language, which we abbreviate to *IntraDiaDist*, and the the cross-lingual distance between diachronic varieties of different languages, which we abbreviate to *CrossDiaDist*.

Our corpus-driven methodology is unsupervised and, therefore, only raw historical corpora were required. The texts on which we carried out the experiments regarding linguistic distance preserve the original spelling; we also calculated the distance between those same texts transliterated into an orthography that is common to the three languages. From now on, we will use the acronyms *OS* for original spelling and *TS* for transcribed spelling.

The specific goal of our experiments is to try to confirm empirically consolidated hypotheses (see H1-H8 below) as well as get new observations from data to verify controversial hypotheses (H9-H10). We report the confirmation of consolidated hypotheses in Section 4, while controversial hypotheses are discussed in Section 5. Table 1 shows the citations and quotes that support the following hypotheses:[1]

(1) H1: Galician has two distinct historical periods: the Galician-Portuguese medieval period and the contemporary period.
(2) H2: Portuguese and Spanish have been considered related languages since the Middle Ages.
(3) H3: Portuguese and Spanish experienced periods of convergence and divergence during their history.
(4) H4: Galician and Spanish have been considered as close but distinct languages.
(5) H5: Galician has progressively converged with Spanish since the second half of the 19th century.
(6) H6: Galician and Portuguese in the Middle Ages are considered two variants of the same language, known as the "Galician-Portuguese" period.
(7) H7: Galician and Portuguese have been separated since the 16th century.
(8) H8: Galician has progressively converged with Portuguese since the first half of the 20th century.
(9) H9 (controversial): During the nineteenth century there was an important import

[1] Many of the quotations are originally in Galician, Portuguese or Spanish. To make reading easier, we have translated them all into English. This is not only valid for this table but for the rest of the article.

of materials in Portuguese from Spanish which brought the languages closer together.

(10) H10 (controversial): The only alternative for Galician language is to be Galician-Portuguese or Galician-Spanish.

To summarize, our experimental research tries to verify if the three languages were gradually separated or whether, on the contrary, there was a much more discontinuous evolution, with convergent and divergent periods. In addition, we also try to measure to what extent spelling plays a role in the distance between periods and languages, both in terms of *IntraDiaDist* and *CrossDiaDist*.

The article is organized as follows: First, some studies on language distance from different approaches will be introduced in Section 2 . Then, the corpus and methodology are described in Section 3. Then, Section 4 reports the results and, finally, controversial results are discussed in Section 5.

2. Related work

2.1. Language Distance

Distance between languages has been approached by numerous studies in the field of the automatic detection of languages and variants of the same language (Jauhiainen, Lui, Zampieri, Baldwin, and Lindén, 2019; Molina, AlGhamdi, Ghoneim, Hawwari, Rey-Villamizar, Diab, and Solorio, 2019; Zampieri, Gebre, Costa, and Van Genabith, 2015). The distance between texts has also been quantified from a diachronic perspective, for example for the automatic classification of the users' stance (Lai, Patti, Ruffo, and Rosso, 2018).

Additionally, these measures have been used in more diverse areas, such as economy (Isphording and Otten, 2013), cultural distance (West and Graham, 2004), the dynamics of language survival (interlinguistic similarity) (Mira and Paredes, 2005), mutual intelligibility (Gooskens, Nerbonne, Vaillette, et al., 2007) or areas related to the acquisition of the second language (Chiswick and Miller, 2004).

There are different methods for calculating the distance between languages. Most of them are based either on lexical comparison (mostly phylogenetic linguistics methods), or on corpus-driven methodologies.

2.1.1. Linguistic Phylogenetics methodologies

Languages can be classified by means of trees that encompass different families, subfamilies and individual languages. This classification is carried out by phylogenetics, which is a sub-field of historical and comparative linguistics, and whose aim is to construct a tree that describes the historical evolution of a set of related languages or linguistic variants from a single root.

There are different methods for building these trees in an automated way, such as *lexicostatistics*, based on lists of words between languages (e.g. Swadesh list (Swadesh, 1952)). The most common methods measure the percentage of shared cognates or involve more complex strategies relying on comparing words that have the same historical origin (Bakker, Muller, Velupillai, Wichmann, Brown, Brown, Egorov, Mailhammer, Grant, and Holman, 2009; Barbançon, Evans, Nakhleh, Ringe, and Warnow, 2013; Holman, Wichmann, Brown, Velupillai, Muller, and Bakker, 2008; Kolipakam, Jordan, Dunn, Greenhill, Bouckaert, Gray, and Verkerk, 2018; List, Walworth, Green-

H1	*"Galician unquestionably framed as an Abstand Galician-Portuguese language, is an Ausbau language that has been consolidated since the nineteenth century."* (Paz, 2008, p. 288)
H2	*"Portuguese and Spanish are the closest Romanesque languages"* (Richman, 1970)
H3	The full book by Fernando Corredoira: *"The construction of the Portuguese against the Spanish. The Galician as opposite case"* shows in detail this hypothesis (Corredoira, 1998).
H4	*"Galician is a language both close to Spanish and Portuguese, with important influences of Spanish throughout the last 500 years."* (Pérez-Pereira, 2008) and *"Galician and Spanish are two very close languages"* (Pérez-Pereira, Alegren, Resches, Ezeizabarrena, Díaz, and García, 2007)
H5	*"Galician has a norm that is substantially close to Spanish and that is a break with respect to medieval Galician-Portuguese and current Portuguese in relation to other standards"* (Mato, 2015).
H6	*"Around 1350, when the Galician-Portuguese literary school became extinct, the consequences of the displacement to the South of the center of gravity of the independent kingdom of Portugal came to light. Portuguese, already separated from Galician by a political border, becomes the language of a country whose capital - that is, the city where the king generally resides - is Lisbon. "* (Teyssier, 1982) and *"Here, we have another incontestable fact: in its early days, the Portuguese language existed concomitantly with Galician. Thus, there was relative linguistic unity between Portugal and Galicia"*(Passerini et al., 2019).
H7	*"The first distinction of Galician and Portuguese as two different languages that I am able to point out for now is found in the account of the events organized in 1572 on the occasion of the transfer to Monterrei of the mortal remains of the founder count of the Jesuit school in that locality."* (Paz, 2008, p. 52).
H8	*"Among the writers of the first third of the 20th century it was also common the substitution of legitimate Galician words by sporadic lusisms such as: até, embora, estudo, nervosas, porén, tolice, etc."* (Paz, 2008, p. 467), or *"For many of the protagonists of the Nós generation (same period) the Portuguese functioned little more than as a place to find the voices that the necessary modernization of the Galician lexicon demanded"* (Paz, 2008, p. 468).
H9	*"In the last quarter of the 18th century, in fact, the fight against the influence of the French burst onto the Portuguese scene, a fight that would continue, lit and militant, throughout the 19th century (...) As French materials were soon seen and felt as strangers, and therefore rejectable, the Spanish were absorbed in complete calm."*(Venâncio, 2014).
H10	*"Galician is either Galician-Portuguese or Galician-Spanish. Galician language is either a form of the western system or of the central system. There is no other alternative"*(Carvalho, 1979).

Table 1. Quotations related to the hypotheses (H1-H10) previously mentioned.

hill, Tresoldi, and Forkel, 2018; Nakhleh, Ringe, and Warnow, 2005; Satterthwaite-Phillips, 2011).

There are other methods to create language trees based on Levenshtein distance between words (Petroni and Serva, 2011), with a normalized Levenshtein distance (Yujian and Bo, 2007), in a cross-lingual list (Petroni and Serva, 2010) or a relationship between languages based on renormalized Levenshtein distance (Serva and Petroni, 2008). Müller, Wichmann, Velupillai, Brown, Brown, Sauppe, Holman, Bakker, List, Egorov, et al. (2010) used techniques based on Levenshtein distance and neighbour-joining algorithm: "The tree is generated through use of the neighbour-joining computer algorithm originally designed to depict phylogenetic relationships in biology." (Saitou and Nei, 1987). Levenshtein distance has also been applied to Galician in relation to other Romance languages in Alecha and González (2016).

2.1.2. Corpus-driven methodologies

Corpus-driven methods for calculating the distance between languages have been carried out, starting from large cross-lingual parallel corpora. Methodologies have been developed based on lexical distances, such as Ellison and Kirby (2006); Heeringa, Golubovic, Gooskens, Schüppert, Swarte, and Voigt (2013) and Criscuolo and Aluisio (2017) with convolutional neural networks; phonetic distances between languages, such as those of Nerbonne and Heeringa (1997), Kondrak (2005) or Singh and Surana (2007), in addition to the comparison of phonological forms between languages as in Eden (2018).

There are other methodologies to measure language distance using monolingual corpora based on word co-occurrences (Asgari and Mofrad, 2016; Gao, Liang, Shi, and Huang, 2014; Liu and Cong, 2013), cross-entropy (Rama, Borin, Mikros, and Macutek, 2015; Singh and Surana, 2007), and perplexity (Gamallo et al., 2017; Hinkka et al., 2018).

An important challenge has been the development of methods to measure the distance between very similar languages or variants and for short texts, where more precision is required, such as in Porta and Sancho (2014); Purver (2014) and Goutte, Léger, Malmasi, and Zampieri (2016).

Finally, corpus-driven methodologies have also been carried out for the measurement of the historical distance (diachronic) between texts in the same language as in Zampieri, Malmasi, and Dras (2016), by using entropy to verify diachronic variation in scientific English (Degaetano-Ortlieb, Kermes, Khamis, and Teich, 2016), or using perplexity applied to diachronic texts in English, Portuguese and Spanish (Pichel et al., 2019b). Buckley and Vogel (2019) use character n-grams in order to explore diachronic change in medieval English. Automatic periodization within a language is a related task, and for this aim, Degaetano-Ortlieb and Teich (2018) use relative entropy. For a similar aim, combination of perplexity and Recurrent Neural Networks (RNN) has been used for identifying temporal trends in a corpus of medieval charters (Boldsen, Agirrezabal, and Paggio, 2019).

Perplexity has been used to compute the cross-lingual diachronic distance between two *Ausbau* languages such as Portuguese and Spanish (Pichel et al., 2019a).

2.2. Sociolinguistics

Languages are tools of communication between people and, as such, they are conditioned by the human societies where they are used. These societies are in continuous evolution, which affects their language or languages in different ways. Holmes and Wilson (2017) claim: “Language varies in three major ways which are interestingly interrelated – over time, in physical space and socially. Language change – variation over time – has its origins in spatial (or regional) and social variation”. Sociolinguistics is focused on the relationship between societies and languages.

The distinction between language and variety (or dialect) has always been controversial. Nordhoff and Hammarström (2011) claim the following: “The question of what is a dialect and what is a language is a very old one, and up to now, there are no agreed upon criteria how to resolve it”. The case of Quechua is used as an example: “Some linguists argue for instance that Quechua is a language family comprising 2, 6, or 46 languages, while others argue that Quechua is one language with a certain number of dialects”. There are countless political aspects to what one vision or the other entails. Nordhoff and Hammarström (2011) conclude: “Political considerations also play a role here: a pan-Quechuan identity advocated by the Academia Mayor de la Lengua Quechua is easier to vindicate if they share a common language rather than if they share a common language family”.

For these reasons, sociolinguists have created different concepts to better understand the relationship between politics, society, languages and varieties.

Written and oral standards have developed in historically consolidated languages, based on prestigious variants normally associated to centres of power. Therefore: “a standard variety is generally one which is written, and which has undergone some degree of regularisation or codification (for example, in a grammar and a dictionary); it is recognised as a prestigious variety or code by a community” (Holmes and Wilson, 2017, p. 78). Standards and dialectal variants of a language also change over time: “change is always interesting, but not always predictable” (Holmes and Wilson, 2017, p. 211).

To study the relationship between different languages, sociolinguists have developed concepts such as *Ausbau* languages (languages historically constructed as distinct to close languages), *Abstand* languages (languages intrinsically distant from other languages) (Kloss, Heinz, 1967), and *polycentric* systems: languages with different centres of political and economic power (da Silva, 2018) that create different linguistic standards (Muhr, 2013).

After the definition of these concepts, we find different approaches aimed at distinguishing languages from dialects (Wichmann, 2016), measuring dialect differences (Heeringa, 2004; Kessler, 1995; Nerbonne and Heeringa, 1997; Nerbonne and Hinrichs, 2006) and classifying polycentric language systems (Zampieri and Gebre, 2012). Dubert and Sousa (2016) developed a methodology specific to the Galician language.

The present work is framed within the corpus-driven methodology, using language distance measure based on perplexity. We will apply the measure to historical variants of three very close *Ausbau* languages (Portuguese, Galician, Spanish), where there has always been sociolinguistic controversy over issues related to the perception of language or variant.

3. Materials and Methods

3.1. Corpus

The corpus required for each language must be representative, of sufficient size, split up in different historical periods, and written with the same orthography as (or very close to) the original texts.

According to Biber (1993), a representative corpus must include "a range of text types in a language". According to Rissanen, Kytö, and Palander-Collin (1993), a historical corpus should be split into, at least, three periods: Medieval (12th-15th centuries), Modern Age (16th-18th centuries), and Contemporary Age (19th-20th centuries). Yet, it is important to bear in mind what Klarer (2013) points out: "The convention of periodical classification must not distract from the fact that such criteria are relative and that any attempt to relate divergent texts –with regard to their structure, contents, or date of publication– to a single period of literary history is always problematic".

Concerning size, the authors of the Helsinki Corpus of Historical English (Rissanen et al., 1993) state that: "The first problem to be decided upon in compiling a corpus is its size" and "The size of the basic corpus is c. 1.5 million words".

Taking into account all these issues, we have created a historical corpus which contains balanced fiction and non-fiction texts with a total size of at least 1.5 million words for each historical period and for each language: Galician, Portuguese and Spanish. Furthermore, the texts included in the corpus are in a spelling as close as possible to the original spelling, since the experiments are carried out both in OS and in an automatically TS.

However, although Portuguese and Spanish have a historical corpus of sufficient size for the three main periods mentioned above, this is not the case for Galician. In particular, from the 16th century to the second half of the 19th century, there are not enough written texts for our experiments. For this reason, our historical corpus contains the Medieval period but not the Modern Age. Moreover, Galician developed a standard spelling historically late, namely in 1981, as opposed to Portuguese and Spanish, which have undergone spelling standardization since the end of the 18th century.

In order to measure the distance between the three languages in a more accurate way and only in periods with a sufficient volume of texts, as well as with important orthographic and linguistic changes, we have defined the following periods: the medieval period; the second half of the 19th century; the 20th century, subdivided into two subperiods of 50 years.

As a result, we created the historical corpus *Carvalho*, which contains four diachronic periods for the three languages: Carvalho-GL (for Galician), CarvalhoPT-PT (for Portuguese in Portugal) and Carvalho-ES-ES (for Spanish in Spain). The four periods are: medieval (XII-XV, i.e., 12th-15th centuries), second half of the 19th century (XIX-2), first half of the 20th century (XX-1), and second half of the 20th century (XX-2). Carvalho is freely available, except for Galician due to copyright issues.[2]

Finally, the three corpora and their periods were divided into train and test parts so as to compute the perplexity-based measure. Table 2 shows the size of both Train and Test corpora across the 4 periods of each language.

The next section characterizes the diachronic corpus of Carvalho for each of the languages. We will focus on the different repositories from which all the documents have been extracted and the significant characteristics of each language.

3.1.1. Galician Corpus

Regarding Galician, the medieval period (12th-15th centuries) is known as the Galician-Portuguese period: "From the late twelfth century to the early fourteenth,

[2] `https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho`

Carvalho	Train-gl	Test-gl	Train-pt	Test-pt	Train-es	Test-es
XII-XV	1.515M	308K	1.509M	305K	1.317M	314k
XIX-2	1.390M	385K	1.464M	312K	1.315M	257K
XX-1	1.404M	319K	1.325M	336K	1.252M	253K
XX-2	1.504M	398K	1.688M	363K	1.231M	250K

Table 2. Size of Train and Test corpora in four historical periods of Galician, Portuguese and Spanish

Galician-Portuguese, a convenient term limited to the period when the two languages had not yet become clearly differentiated" Azevedo (2005); Robl (1982). There are sufficient texts belonging to the medieval period, which lasted from the 12th to the 15th century.

During the 16th to 18th centuries and the first half of the 19th century (XIX-1) there are not enough texts written in this language for our experiments. However since the second half of the 19th century (XIX-2), from the period called "Rexurdimento" to the present time (Carvalho, 1981; Vilavedra and Fdez, 1999), we do have sufficient documents to be able to apply the methodology described in Section 3.2.

Regarding orthography, from the Middle Ages to the present day, Galician spelling oscillates between proximity to Portuguese orthography (medieval period) and to Spanish spelling (modern and contemporary period).

The Carvalho-GL corpus we have compiled for the medieval period (XII-XV) is part of the TMILG (Galician Language Medieval Treasure) corpus (Moura, López, and Pichel, 2008; Varela Barreiro, 2004). For periods XIX-2, XX-1 and XX-2, we have used texts from the TILG (Galician Language Computerized Treasure) corpus (Santamarina, 2003). The Carvalho-GL corpus cannot be accessed due to copyright law, although its authors can be contacted.

Table 3 shows some relevant information required to build the Carvalho-GL corpus: the historical studies we used to prepare the material, the corpus resources from which the documents in OS were selected, and some samples of fictional and non-fictional documents included in the final corpus.

3.1.2. Portuguese Corpus

Texts in Portuguese, contrarily to Galician and similarly to Spanish, didn't stop being written at the end of the 15th century and continued uninterruptedly until the present day. For this reason, there is a corpus with sufficient size for our experiments, encompassing texts from the 12th century to the end of the 20th century.

From the point of view of standardized orthography, as also happens with Spanish, the Academy of Sciences of Lisbon has promoted different orthographic standards and norms since the year 1779 (e.g.: 1885, 1911, 1945, 1973, 1990), some of them fraught with controversy (e.g., the last reform, known as "Acordo Ortográfico de 90").

For the elaboration of this corpus, we have selected texts with the spelling as close as possible to the original, removing edited texts such as the one we can see in Table 4. Thus in texts of the 19th century and the first period of the 20th century, the spelling "ph" was used for the phoneme /f/ and in many available digital versions the texts were adapted to modern spelling by replacing "ph" with "f". We discarded these versions.

We have already used Carvalho-PT-PT to measure the *IntraDiaDist* of Portuguese

[3] https://ilg.usc.es/tmilg/

[4] https://ilg.usc.es/TILG/

studies	"Historia da Literatura galega contemporánea" (Carvalho, 1981), "Galician and Castilian in contact: historical, social and linguistic aspects" (Monteagudo and Santamarina, 1993), "A construção da língua portuguesa frente ao castelhano: o galego como exemplo a contrario." (Corredoira, 1998), "Historia social da lingua galega: idioma, sociedade e cultura a través do tempo" (Monteagudo and Romero, 1999), "Historia da Literatura galega" (Vilavedra and Fdez, 1999), "Gramática da lingua galega II. Morfosintaxe " (Freixeiro Mato, 2000), "O estudo do mundo lusófono no sistema literário galego. Bases metodológicas para o estudo dos sistemas emergentes e as suas relaçons intersistémicas." (Torres Feijó, 2002) "A fouce, o hórreo eo prelo: Ánxel Casal ou o libro galego moderno" (Vázquez Souza, 2003) "Historia de Galicia"(Villares, 2004) "Historia da lingua galega" (Paz, 2008), "O galego (im)possível" (Rodrigues Fagim, 2001)
sources	TMILG (Tesouro Medieval Informatizado da Lingua Galega) [3], TILG (Tesouro Informatizado da Lingua Galega) [4],
fiction	"Cantigas de Santa Maria" by Alfonso X, "Follas Novas" by Rosalia de Castro, "Queixumes dos Pinos" by Eduardo Pondal, "Da Terra asoballada" by Ramón Cabanillas, "Crónica de nós" by Xosé Luís Méndez Ferrín
non-fiction	"Crónica Geral de Castela", "O Tío Marcos da Portela" by Valentín Lamas Carvajal, "A nosa terra" a galician magazine, "Para un axeitado dereito foral galego" by Carlos Abraira López

Table 3. Metadata on Carvalho-GL corpus: historical studies, corpus resources and an ordered sample from the Middle Age to the 20th century of fictional and non-fictional writings.

in Pichel et al. (2019b) and Pichel et al. (2018). In those articles, we reported studies, sources and examples of fiction and non-fiction texts used to compile the corpus.

OS	**TS**	**Edited**
Deus, a vida, os grandes problemas, não são os philosophos que os resolvem, são os pobres vivendo (...)	deus, a vida, os grandes problemas, não são os filosofos que os resolvem, são os pobres vivendo (...)	Deus, a vida, os grandes problemas, não são os filósofos que os resolvem, são os pobres vivendo (...)

Table 4. Portuguese excerpt in three versions: original spelling (OS), transcribed (TS), and edited text.

3.1.3. Spanish Corpus

Regarding Spanish, there is, as is the case of Portuguese, a corpus with sufficient size in all historical periods, which allowed us to carry out our *IntraDiaDist* and *CrossDiaDist* distance experiments.

Since the time of Alfonso X, in Spain, there was a desire to harmonize spelling and create a single standard. However, only after the creation of the Real Academia Española in 1713 and the orthographic standard in 1741 (Lapesa and Pidal, 1942) a standardized spelling began to spread. The Spanish spelling standard didn't include solutions that are still used in the rest of the Romance languages, such as "ss", "ç" and latinisms (Alatorre, 2002).

We have already used Carvalho-ES-ES to measure the *IntraDiaDist* of Spanish in (Pichel et al., 2019b). In that article, we have reported the studies, sources, and samples of fiction and non-fiction texts used in the elaboration of the corpus.

3.2. Methodology

In previous work, our methodology has been used to measure the *IntraDiaDist* in three different languages: Portuguese, Spanish and English (Pichel et al., 2019b). It has also been applied to measure the *CrossDiaDist* between two closely related languages, such as Portuguese and Spanish (Pichel et al., 2019a).

Now we will improve this methodology to calculate the *CrossDiaDist* between three languages. In our case it will be applied to a language (Galician) that historically has a very close *Ausbau* relationship with two other also related *Ausbau* languages.

This methodology is unsupervised as no annotated text is required.

In the following section, we will describe the corpus-based measurement and the different steps of the method.

3.2.1. Perplexity-Based Measurement

Perplexity is frequently used as a quality measure for language models built with n-grams extracted from text corpora (Chen and Goodman, 1996; Dieguez-Tirado, Garcia-Mateo, Docio-Fernandez, and Cardenal-Lopez, 2005; Sennrich, 2012). It has also been used in very specific tasks, such as for classifying formal and colloquial tweets (González, 2015), and for identifying closely related languages (Gamallo, Alegria, Pichel, and Agirrezabal, 2016).

This is a metric about how well a language model is able to fit a text sample. A low perplexity indicates the language model is good at predicting the sample. On the contrary, a high perplexity shows the language model is not good at predicting the given sample. It turns out that we could use perplexity to compare the quality of language models in relation to specific textual tests.

More formally, the perplexity (called PP for short) of a language model on a textual test is the inverse probability of the test. For a test of sequences of characters $CH = ch_1, ch_2, ..., ch_n$ and a language model LM with n-gram probabilities $P(\cdot)$ estimated on a training set, the perplexity PP of CH given a character-based n-gram model LM is computed as follows:

$$PP(CH, LM) = \sqrt[n]{\prod_{i}^{n} \frac{1}{P(ch_i|ch_1^{i-1})}} \tag{1}$$

where n-gram probabilities $P(\cdot)$ are defined in this way:

$$P(ch_n|ch_1^{n-1}) = \frac{C(ch_1^{n-1}ch_n)}{C(ch_1^{n-1})} \tag{2}$$

Equation 2 estimates the n-gram probability by dividing the observed frequency

(C) of a particular sequence of characters by the observed frequency of the prefix, where the prefix stands for the same sequence without the last character. To take into account unseen n-grams, we use a smoothing technique based on linear interpolation.

Our perplexity-based language distance, called PLD, is defined as follows:

$$PLD(L1, L2) = \frac{(PP(CH_{L2}, LM_{L1}) + PP(CH_{L1}, LM_{L2}))}{2} \tag{3}$$

The lower the perplexity of both CH_{L2} given LM_{L1} and CH_{L1} given LM_{L2}, the lower the distance between languages (or language periods) $L1$ and $L2$. Notice that PLD is the symmetric mean derived from two asymmetric divergences: $PP(CH_{L2}, LM_{L1})$ and $PP(CH_{L1}, LM_{L2})$.

In the current work, our aim is to apply Equation 3 to measure *IntraDiaDist* and *CrossDiaDist* for three different languages in the same historical periods. In order to be able to compare the perplexity distances we have obtained with those reported in Gamallo et al. (2017), we use the same PLD configuration: namely, 7-gram language models, a smoothing technique based on linear interpolation, and train/test corpora with 1.25M/250K words, respectively.

In order to allow researchers to measure PLD distances between periods of any language, we have developed a pipeline architecture in Perl, which is freely available.[5].

3.2.2. Task Description

Our method is tailored to measure *CrossDiaDist* between three languages and is divided into the following sequential tasks:

(1) To define common historical periods for all languages.
(2) To obtain corpora of sufficient size in OS for all languages in those periods. Excerpts in any other language (e.g., Latin) are removed.
(3) To set up a balanced corpus structure divided into train and test for each period. Texts are balanced between fiction and non-fiction in both train and test partitions at approximately 50%. Each train partition contains at least 1.25M words per period, while test partitions have at least 20% of the size of the train partition, i.e. between 250K and 350K words.
(4) To compute the *IntraDiaDist* between periods of each of the languages PLD(L1), PLD(L2) and PLD(L3), by applying PLD to texts in OS.
(5) To compute the *IntraDiaDist* of texts in TS. Before that, a spelling normalization is applied on all the texts and a transcribed version is obtained for each corpus and partition. For this purpose, we have implemented a transcriber whose alphabet consists of 34 symbols, representing 10 vowels (including accents) and 24 consonants, designed to cover most of the commonly occurring sounds, including several consonant palatalizations. The encoding is thus close to a phonological one and makes it possible to simplify and homogenize cases in which similar sounds (generally palatalizations) are transcribed differently in different languages. For instance, the palatalized nasal sound is transcribed by our normalizer as “ny”, thus unifying the Portuguese spelling “nh” and Galician and Spanish spelling “ñ”. Similarly, the palatalized lateral is transcribed as “ly”, unifying the two different spellings: “lh” in Portuguese and “ll” in Galician and

[5] `https://github.com/gamallo/Perplexity`

Spanish. The palatal affricate sound in Galician and Spanish, as well as in Portuguese, represented by the spelling "ch", is transcribed into "ĉ".

(6) To verify that the *IntraDiaDist* of PLD(L1), PLD(L2) and PLD(L3) gives expected results both in OS and TS by considering the studies of the community of historians of each language. If results are not consistent, we check whether there is noise in the corpus (mainly caused by the presence of other languages, encoding problems, repetitions, etc.), and then we go back to task 2 of the method.

(7) To compute the *CrossDiaDist* between periods of each of the language pairs PLD(L1, L2), PLD(L1, L3) and PLD(L2,L3) in OS and TS. The results will be evaluated and analyzed later. With this implementation, we have built train partitions giving rise to six different 7-gram diachronic language models per language. Then, we have analyzed all test documents so as to generate six 7-gram files per language.

4. Results

We carried out several experiments applying our methodology from task 1 to 7 (see Section 3.2), so as to measure several language distances between Spanish, Galician and Portuguese. To this end, Carvalho-GL, Carvalho-PT-PT and Carvalho-ES-ES were used considering all the requirements pointed out in the described methodology.

Regarding the validation task (6), it is worth noting that we have already done and validated the *IntraDiaDist* for Portuguese and Spanish in a previous work (Pichel et al., 2018). So, in this section, we only compute *IntraDiaDist* for Galician language.

Having verified that all *IntraDiaDist* are accurate, we compute all the possible *CrossDiaDist* as described in task 7 for all possible combinations: Portuguese-Spanish, Galician-Portuguese, and Galician-Spanish. We will analyze the results by highlighting the observations that allow us to confirm the eight consolidated hypotheses reported in the Introduction. Later, in Section 5, we will try to shed light on the two remaining controversial hypotheses (9 and 10).

4.1. Intralingual Diachronic Distance for Galician

Table 5 shows the results of calculating the PLD in OS between all periods of Galician using the Carvalho-GL corpus. On the other hand, Table 6 shows the results of performing the same experiment after transcribing all periods into the same spelling (TS). In Figure 1(a) we can see the evolution of distance across all periods in OS, while Figure 1(b) presents the same evolution, but using TS.

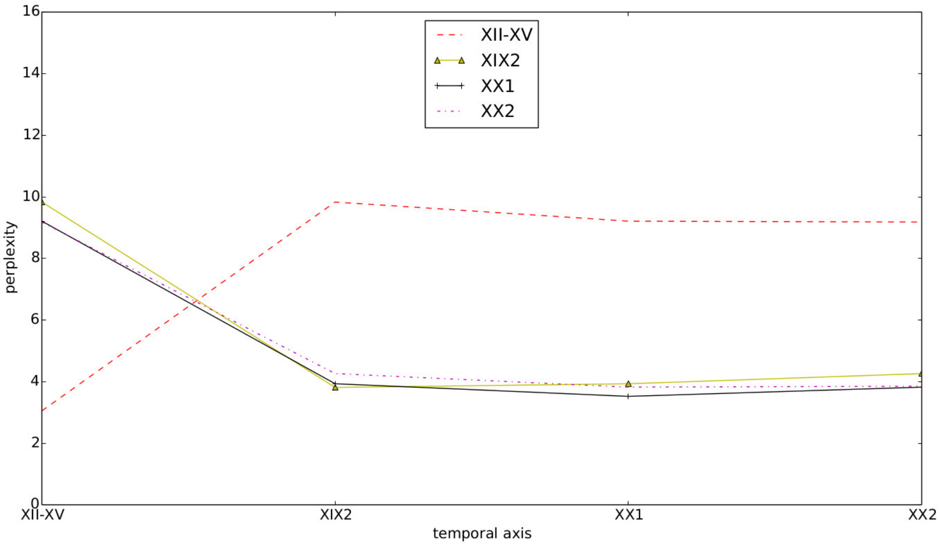
The PLD values in both OS and TS show that Galician in the Middle Ages (XII-XV) shows a significant distance from the period when the Galician language started being written again in an extensive way (XIX-2). This distance decreases progressively in the following subperiods of the 20th century (XX-1 and XX-2).

Regarding the results in OS, we can observe that the medieval period (XII-XV) is distant from the XIX-2 period, with a PLD of 9.83 (the most significant distance). This may be due to the fact that, as Areán-García (2011) said: "The Galician language, after its medieval splendour and development as a cultured language, went through a period of strong decadence, known as the Dark Ages, from the end of the Middle Ages to the beginning of the 19th century, and only had its first grammar published at the end of the 19th century."

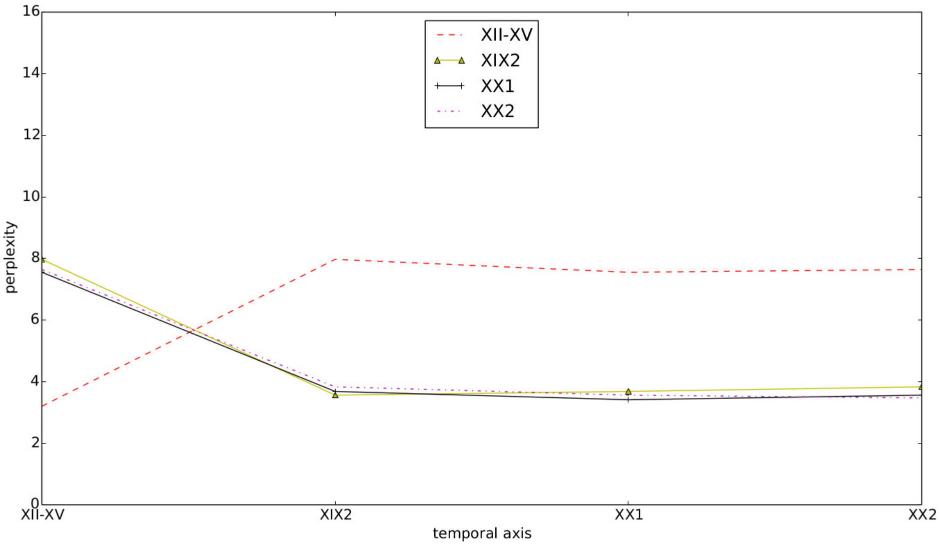
Then, the PLD distance between XII-XV and XX closes a little (9.21 and 9.18 in

	XII-XV	XIX-2	XX-1	XX-2
XII-XV	3.05	9.83	9.21	9.18
XIX-2	9.83	3.81	3.93	4.26
XX-1	9.21	3.93	3.52	3.82
XX-2	9.18	4.26	3.82	3.85

Table 5. PLD diachronic measurement in OS (Carvalho-GL corpus)



(a) Original spelling



(b) Transcribed spelling

Figure 1. In (a) we compare the Galician PLD distances between XII-XV and XX-2 across all periods (except XVI-XVIII and XIX-1) in OS. In (b) the same comparison using a TS.

	XII-XV	XIX-2	XX-1	XX-2
XII-XV	3.2	7.97	7.55	7.64
XIX-2	7.97	3.56	3.68	3.83
XX-1	7.55	3.68	3.41	3.56
XX-2	7.64	3.83	3.56	3.47

Table 6. PLD diachronic measurement in TS (Carvalho-GL corpus)

XX-1 and XX-2, respectively). The reason for this may be the setting of an academic standard for Galician, cleansed of dialectalisms and vulgarisms, the creation in 1905 of the Real Academia Galega (RAG) with the aim of creating an official Galician dictionary and a grammar, "although these ambitious projects were only partially accomplished from the 1980s onwards" (Ramallo and Rei-Doval, 2015), and "the discovery of the ancient (medieval) tradition, which in any case did not translate into proposals for the adoption of its graphic conventions." (Gulías, 1992; Paz, 2008; Seoane, 1992).

Concerning the results in TS, we see that the distance between the medieval period (XII-XV) and all other periods is less significant than in OS: PLD 7.97 in XIX-2, PLD 7.55 in XX-1 and PLD 7.64 in XX-2. This may because Spanish served as the basis of the orthographic model for Galician in this period: "Of course, Spanish was a model they could not ignore as it was the language they had learned to write in." (Ramallo and Rei-Doval, 2015).

With these results in both OS and TS, we can verify that the medieval period (XII-XV) is considerably distant from all other periods, especially in OS. The hypothesis (H1), which states that Galician has two distinct historical periods (XII-XIV and XIX-2/XX-1/XX-2), is thus confirmed.

Finally, other observations related to these results will be discussed in Section 5.

4.2. Cross-lingual Diachronic Distance

We will now apply the described methodology to measure the distance between three languages across the same historical periods. Thus, we performed PLD calculations for each language pair combination: Portuguese-Spanish, Galician-Spanish and Galician-Portuguese. The experiments were carried out with both OS and TS. Our aim is to verify whether our results correlate with the consolidated hypotheses reported in the Introduction.

4.2.1. Portuguese-Spanish

Table 7 shows the results of applying PLD to OS and TS versions of the Portuguese and Spanish corpora (Carvalho-PT-PT and Carvalho-ES-ES), period by period. In Figure 2, we can see all the information in a plot so as to better observe how the two languages behave in relation to each other through the time axis (except 16th-18th and 19th-1 periods).

We can observe how the PLD distance (in OS and TS) decreases from the medieval period to the second half of the 19th century, where it reaches the minimum PLD score: 9.78 (OS) and 7.49 (TS). The influence of French on the Romance languages during this period may be the cause of this approximation (Curell, 2006), although it may be also due to an huge import of linguistic materials from Spanish into Portuguese between the 15th and 18th centuries (Venâncio, 2014).

Periods	PLD (OS)	PLD (TS)
XII-XV	11.48	8.9
XIX-2	9.78	7.49
XX-1	13.20	9.34
XX-2	11.99	9.04

Table 7. Cross-lingual diachronic distance (PLD) between Spanish and Portuguese across four historical periods in original spelling (OS) and transcribed (OS).

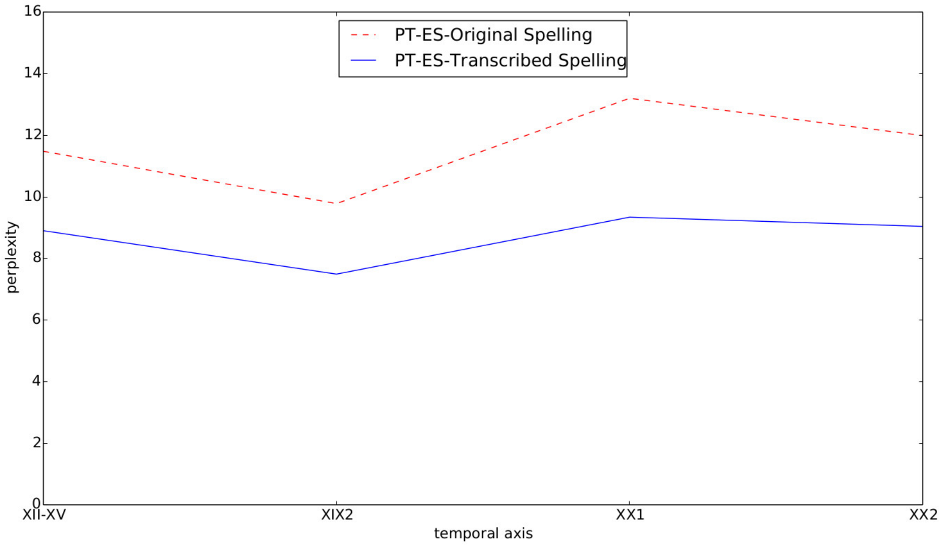


Figure 2. Cross-lingual diachronic distance between Portuguese and Spanish through time axis in OS and TS.

Then, in a short period, the distance increases again, peaking in the first half of the 20th century: 13.20 (OS) and 9.34 (TS). This greater distance could be partially explained by the new orthographic rules applied to the Portuguese standard during the period of the Republic, the strengthening of the nation-state concept, involving compulsory schooling, and the new importance given to ending spelling variations in official publications (dos Reis Aguiar, 2007).

Later on, in the second half of the 20th century (XX-2), the two languages approach each other again: we find a PLD of 11.99 in OS and 9.04 in TS. The latter PLD value is very similar to the distance between present-day Spanish and present-day Catalan: a PLD of 8.63 in TS.[6]

Concerning the relationship between Portuguese and Spanish since medieval ages, we see that the closest and the furthest distance are equivalent to the current distance between Spanish and Catalan. So, we can confirm the hypothesis (H2) stating that since the Middle Ages Portuguese and Spanish are close languages.

Furthermore, the relationship between these two languages goes through different periods of convergence and divergence. As we have explained above, there may

[6] This PLD value was computed by making use of the distance search engine (`https://gramatica.usc.es/~gamallo/php/distance/`) which was the result of the work described in Gamallo et al. (2017).

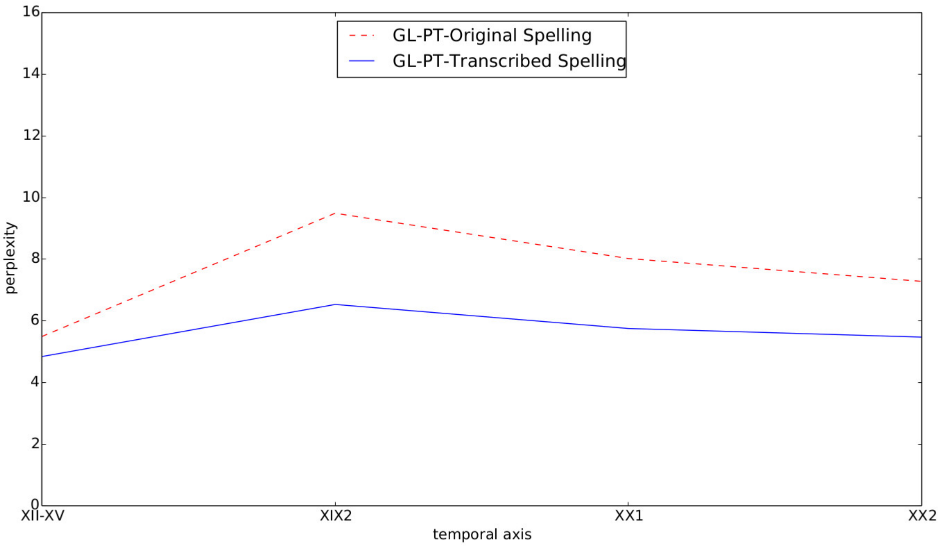


Figure 3. Cross-lingual diachronic distance between Galician and Portuguese through time axis in OS and TS.

be socio-political reasons that explain the sequence of periods of closeness/distance between these two languages separated by elaboration (*Ausbau*). This confirms the hypothesis (H3), which states that both languages experienced periods of convergence and divergence during their history.

Finally, other observations related to these results will be discussed in Section 5.

4.2.2. Galician-Portuguese

Table 8 shows the results of applying PLD to OS and TS versions of the Galician and Portuguese corpora (Carvalho-GL and Carvalho-PT-PT), period by period. In Figure 3, we can see the same information in a plot so as to better observe how the two languages behave in relation to each other throughout history (except the 16th-18th and 19th-1 periods).

Periods	**PLD (OS)**	**PLD (TS)**
XII-XV	5.49	4.84
XIX-2	9.49	6.53
XX-1	8.02	5.75
XX-2	7.28	5.47

Table 8. Cross-lingual diachronic distance (PLD) between Galician and Portuguese across four historical periods in OS and TS.

The PLD values in both OS and TS show that, in the Middle Ages (XII-XV period), Galician and Portuguese were very close, but they moved away considerably in the 19th century, especially in OS. Later, in the two sub-periods of the twentieth century (XX-1 and XX-2), they move closer to each other again.

The minimum *CrossDiaDist* between Galician and Portuguese is found in XII-XV: 5.49 PLD in OS, being even closer in TS: 4.84 PLD. Notice that this last value is equivalent to that of two very close diachronic varieties of Spanish: the XII-XV variety and the XVI-XVIII variety, which present a PLD of 4.95 in TS. This confirms the hypothesis (H6), which states that Galician and Portuguese, in the medieval period (known as Galician-Portuguese period (da Silva, 2018; Diez, 2008)), are considered as two historical varieties, and not as two close but distinct languages.

The largest distance between Galician and Portuguese, after the medieval period, is the one found in the XIX-2 period: 9.49 in OS and 6.53 in TS. This seems to confirm the hypothesis (H7) that Galician and Portuguese have undergone a process of separation until the end of the nineteenth century, when they start to be considered as two close but different languages.

Later, starting from the first half of the 20th century, the *CrossDiaDist* between Galician and Portuguese progressively decreases, presenting a PLD of 8.02 (OS) and 5.75 (TS) in the first half of the 20th century and 7.28 (OS) and 5.47 (TS) in the second half of the 20th century. As we have reported in (Pichel et al., 2019b), this distance is equivalent to historical variants close in time, for instance the *IntraDiaDist* between the 16th-18th and 19th-1 periods in Spanish: 5.57 (TS). Furthermore, a similar value is also found between very close languages/varieties, such as Bosnian and Croatian, with a distance of 5.90.[7] These low values seem to confirm the hypothesis (H8) that Galician gradually converges with Portuguese starting from the first half of the 20th century.

Finally, other observations related to these results will be discussed in Section 5.

4.2.3. Galician-Spanish

Lastly, Table 9 shows the results of applying PLD to OS and TS versions of the Galician and Spanish corpora (Carvalho-GL and Carvalho-ES-ES), period by period. Figure 4 allows us to better visualize all the data.

Periods	**PLD (OS)**	**PLD (TS)**
XII-XV	8.18	7.35
XIX-2	7.40	6.04
XX-1	7.32	6.01
XX-2	7.08	5.81

Table 9. Cross-lingual diachronic distance (PLD) between Galician and Spanish across four historical periods in OS and TS.

The PLD values show that Galician and Spanish reached the maximum distance (8.18 in OS and 7.35 in TS) in the Middle Ages (XII-XV period) and move progressively closer from then on, reaching the minimum distance (7.08 in OS and 5.81 in TS) in the last sub-period (XX-2).

The approximation between Galician and Spanish in the second half of the 19th century may be due to the fact that the Galician authors recovered the literary and educated usage of Galician after the so-called “dark centuries” (16th to 18th century) without being aware of the medieval tradition; therefore, they mostly reproduced, in their writings, the oral varieties that were obviously influenced by the Spanish

[7] PLD value computed by making use of the search engine `https://gramatica.usc.es/~gamallo/php/distance/`.

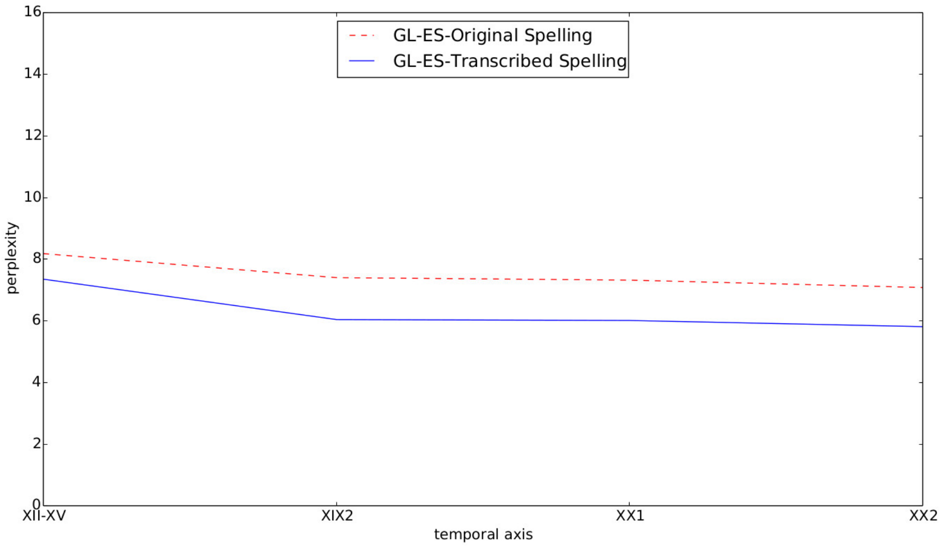


Figure 4. Cross-lingual diachronic distance between Galician and Spanish through time axis in OS and TS.

language, as we mentioned in Section 4.1.

From the XX-1 period on, the two languages continued to get closer to each other both in OS and TS. This progressive approximation between Galician and Spanish can be explained by the attempt to create a standard for Galician that refuses popular forms, as in the previous period, and the effects of the creation of a standard by the RAG, also commented on previously in Section 4.1.

This progressive approach confirms the hypothesis (H5), which claims that Galician has progressively converged with Spanish since the second half of the 19th century.

It is worth noting that the furthest distance between Galician and Spanish, reached in the 12th-15th centuries period, is similar to the perplexity distance between two distinct (but close) languages such as Czech and Slovak: 8.1 in TS.[8] This seems to confirm the hypothesis (H4), which states that Galician and Spanish have been considered close but distinct languages since the Middle Ages.

Further observations related to these results will be discussed in Section 5.

5. Discussion

In the previous section, we verified that results obtained by our method correlate with the eight consolidated hypotheses. Therefore, since the measurement of perplexity allows us to independently detect trends and patterns previously described by specialists, we may conclude that the proposed method is solid and can be used to find new patterns and to support or reject controversial hypotheses.

In this section, we emphasize new observations drawn from the results reported in the previous section. We focus on trends and patterns that were not discussed in the previous section, as they are not related with the consolidated hypotheses, but rather

[8] Extracted from the search engine `https://gramatica.usc.es/~gamallo/php/distance/`

	PT-ES(OS)	GL-PT(OS)	GL-ES(OS)	PT-ES(TS)	GL-PT(TS)	GL-ES(TS)
XII-XV	11.48	5.49	8.18	8.9	4.84	7.35
XIX-2	9.83	9.49	7.40	7.52	6.53	6.04
XX-1	13.2	8.02	7.32	9.34	5.75	6.01
XX-2	11.99	7.28	7.08	9.04	5.47	5.81

Table 10. Cross-Lingual Diachronic Distance in OS and TS of the three compared pairs: pt-es, gl-pt, and gl-es.

with controversial ones. In addition, we also discuss other assumptions that were not mentioned until now.

We start with the *IntraDiaDist* concerning the Galician language. Then, regarding *CrossDiaDist*, we describe the relationship of the three languages as a group and discuss some new observations made on the basis of the three language pairs: Portuguese-Spanish, Galician-Portuguese, Galician-Spanish. Table 10 and Figure 5 do not introduce new data. They synthesize the results of the three compared pairs, allowing us to better visualize the new trends and patterns.

5.1. Final Discussion

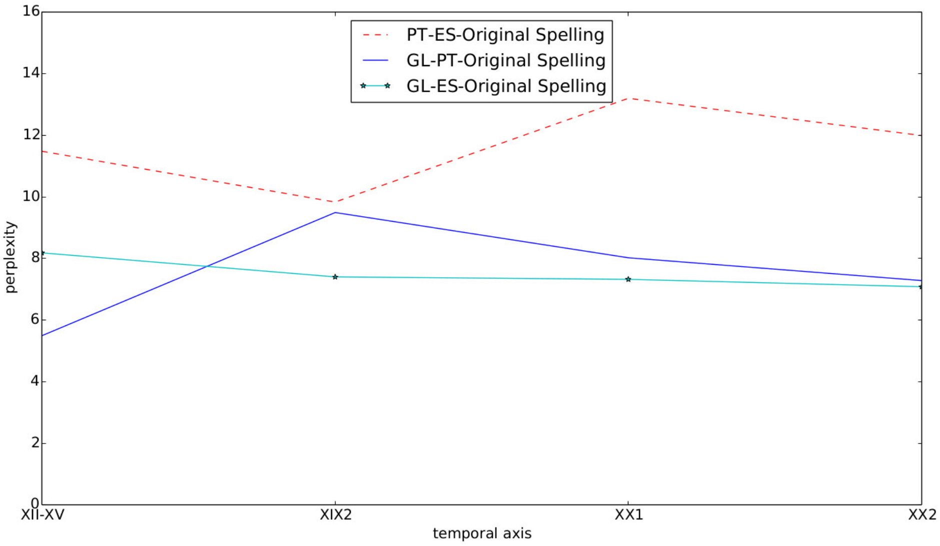
Regarding the *IntraDiaDist* in Galician, we observe the two following facts:

(1) Firstly, the distance in OS and TS between the medieval period (XII-XV) and the second half of the nineteenth century (XIX-2) is greater than that occurring in Portuguese between the same periods (Pichel et al., 2019b). This observation does not seem to be in accordance with the assumption claimed by Monteagudo (2017): "Galician was not unilaterally split from an original and common Galician-Portuguese trunk that would be better represented by European Portuguese; in fact, in a series of aspects Galician is closer to the medieval linguistic stage, while in others it is Portuguese that is closer to it".

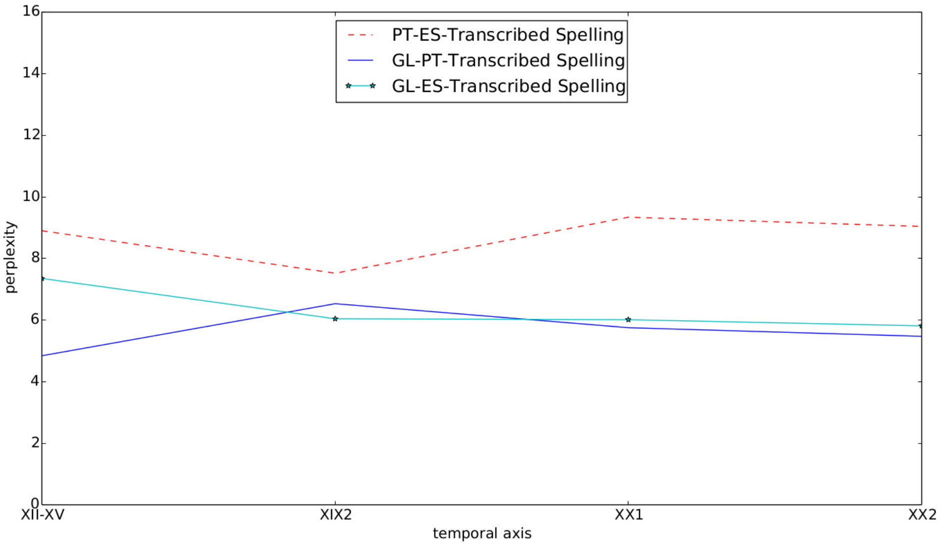
(2) Secondly, we observe that all recent periods (XIX-2, XX-1 and XX-2) are close to each other. In fact, their PLD values are lower than equivalent PLD values found when measuring the distance between different periods both in Portuguese and in Spanish (Pichel et al., 2019b). This observation seems to contradict the generalized idea (an intuition or prejudice) that Galician is always changing as opposed to more stable languages such as Spanish and Portuguese. Our data show that Galician is more stable than expected.

In relation to *CrossDiaDist*, we observe two other facts related to the three languages under study:

(1) Portuguese and Spanish were coming closer to each other from the Middle Ages (Pichel et al., 2019a) until the second half of the 19th century, when they reached the shortest distance. This may be due to the fact that Portuguese has imported an enormous amount of linguistic material from Spanish, which seems to be aligned with the controversial hypothesis (H9) by Venâncio (2014), who claims : "In the last quarter of the 18th century, in fact, the fight against the influence of the French burst onto the Portuguese scene, a fight that would continue, strong and militant, throughout the 19th century [...] As French materials were soon seen and felt as foreign, and therefore to be rejected, Spanish materials were calmly absorbed". Our experiments support this hypothesis.



(a) Original spelling



(b) Transcribed spelling

Figure 5. In (a) we compare the Portuguese-Galician-Spanish PLD distances between XII-XV and XX-2 across all periods (except XVI-XVIII and XIX-1) in OS. In (b), the same comparison using TS.

(2) Galician has shown a strong relationship with both Portuguese and Spanish since the XIX-2 period, in which orthography has played a fundamental role.

Regarding the second observation, we can present the following details:

- Galician, Portuguese and Spanish are closer if they use a common orthography (TS).
- Ortography is not a relevant factor to separate Galician and Spanish since the XIX-2 period. This may be due to the fact that Spanish significantly modified its orthography with respect to other Romance languages around the end of the 18th century and the early 19th century (Villa, 2013), and this had an impact on Galician writing since the XIX-2 period as Monteagudo and Santamarina (1993) claims: "in the early day of the *Rexurdimento*, written Galician ignored medieval and Portuguese spelling conventions, making use of Spanish orthography, which was familiar to Galician writers".
- Ortography is a relevant factor to analyse the distance between Galician and Portuguese since the XIX-2 period, but not in the medieval period. In fact, Galician and Portuguese between the 12th and 15th centuries have a similar distance in OS to that which exists between both languages in the XX-2 period in TS. The relevant issue is that, in the medieval period, Galician and Portuguese were written with similar spellings, while, in the second half of the 20th century, they used different ones. This is in accordance with the claim made by Jones and Mooney (2017): "the use of Spanish orthographic conventions may help to distinguish Galician from Portuguese, to which it is linguistically more similar".
- Galician comes closer to both Spanish and Portuguese since the 20th century. This may be due to the fact that, since the XX-1 period, Galician has had a tendency to construct "a standard with characteristics similar to those of the Spanish and Portuguese, assuming the hierarchization that standardization brings with it" (Álvarez and Monteagudo, 2005). The standardization of Galician makes it closer to Spanish and Portuguese at the same time.
- Galician comes closer to Spanish in OS and to Portuguese in TS, in the 20th century. This may be due to the fact that Galician seems to behave as an *Ausbau* language in which orthography is relevant to establish its relationship with Portuguese and Spanish. This is consistent with the claim by Kloss, Heinz (1967): "The process of *ausbau*, and the creation of *abstand*, involves establishing linguistic autonomy from related languages by *reshaping* the visual representation of the language while the linguistic structure of the language(s) remains, in principle, unchanged".

Finallly, bearing in mind the last observation and considering that the distance between Galician and the other two languages in TS in the XX-2 period is equivalent to the distance between Bosnian and Croatian (Gamallo et al., 2017), Galician can be seen either as Galician-Spanish in OS or as Galician-Portuguese in TS. This is in accordance, in fact, with the controversial hypothesis (H10) stated by Carvalho (1979): "Galician is either Galician-Portuguese or Galician-Spanish. Galician language is either a form of the western system or of the central system. There is no other alternative".

5.2. Further work

Based on these results, we would like to apply PLD to measure the distance in polycentric languages such as Portuguese (European and Brazilian Portuguese) and Spanish

(European and Latin American Spanish). It is also our aim to measure distances in a diachronic perspective (i.e. did the distance between Argentinean Spanish and European Spanish increase or decrease during their history?, Is the distance between Brazilian Portuguese and European Portuguese greater, lesser, or equal to the distance between Argentinean Spanish and European Spanish?).

Our aim is also to use PLD with different language models: e.g. n-grams calculated from relevant linguistic words, more complex phonological rules modifying the spelling, (contextualized) word embeddings, etc.

Finally we would like to investigate the relationship between language distance using PLD and Machine Translation Quality estimation (Han, Lu, Wong, Chao, He, and Xing, 2013; Specia, Scarton, and Paetzold, 2018).

Acknowledgement(s)

We are very grateful to Xavier Varela and Antón Santamarina from ILG (Galician Language Institute) by lending the original Galician corpus of TMILG and TILG in order to carry out these experiments. Special acknowledgment is due to Xavier Varela, José António Souto Cabo of the Universidade de Santiago de Compostela for his expertise in medieval historical linguistics of Galician-Portuguese, Antón Santamarina for his expertise in modern and contemporary Galician language history, Fernando Venâncio of the University of Amsterdam, Marco Neves of the Universidade Nova de Lisboa and Carlos Quiroga from Universidade de Santiago de Compostela for their expertise in the history of Portuguese, Maria Isabel Fernández Domínguez for her expertise in the history of Spanish, Teresa Moure Pereiro of the University of Santiago de Compostela for her contributions from linguistics and Alfonso Barata Villapol for the bibliographical contributions on the history of the English language. Finally we would like to thank to Marcos Garcia of the Universidade da Coruña for his contributions to the development of the experiments.

Funding

This work has received financial support from DOMINO project (PGC2018-102041-B-I00 , MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08, Consolidation and structuring of Groups with Growth Potential: ED431B 2017/39) and the European Regional Development Fund (ERDF).

References

Alatorre, Antonio. 2002. *Los 1001 años de la lengua española*, vol. 3. Fondo de Cultura Económica.

Alecha, Esteve Valls and Manuel González González. 2016. Variación e distancia lingüística na romania antiqua: unha contribución dialectométrica ao debate sobre o grao de individuación da lingua galega. *Estudos de Lingüística Galega* 8:229–246.

Álvarez, Rosario and Henrique Monteagudo. 2005. *Norma lingüística e variación: unha perspectiva desde o idioma galego*. Inst. da lingua Galega.

Areán-García, Nilsa. 2011. A divisão do galego-português em português e galego, duas línguas com a mesma origem. *Revista philologus* 49:1–14.

Asgari, Ehsaneddin and Mohammad R. K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 65–74. San Diego, California.

Azevedo, Milton M. 2005. *Portuguese: A linguistic introduction*. Cambridge University Press.

Bakker, Dik, Andre Muller, Viveka Velupillai, Soren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology* 13(1):169–181.

Barbançon, F., S. Evans, L. Nakhleh, D. Ringe, and T. Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30:143–170.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and linguistic computing* 8(4):243–257.

Boldsen, Sidsel, Manex Agirrezabal, and Patrizia Paggio. 2019. Identifying temporal trends based on perplexity and clustering: Are we looking at language change? In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 86–91.

Buckley, Kevin and Carl Vogel. 2019. Using character n-grams to explorediachronic change in medieval english. *Folia Linguistica* 40(2):249–299.

Carvalho, R. 1979. Sobre a nosa lingua. *Grial* 17(64):140–152.

Carvalho, Ricardo. 1981. *Historia da literatura galega contemporánea: 1808-1936*. Editorial Galaxia.

Chen, Stanley F. and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chiswick, B.R. and P.W. Miller. 2004. *Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages*. Discussion papers. IZA.

Corredoira, Fernando Vázquez. 1998. *A construção da língua portuguesa frente ao castelhano: o galego como exemplo a contrario*. Edicións Laiovento.

Criscuolo, Marcelo and Sandra Maria Aluisio. 2017. Discriminating between similar languages with word-level convolutional neural networks. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 124–130.

Curell, Clara. 2006. La influencia del francés en el español contemporáneo. In *La cultura del otro: español en Francia, francés en España*, pages 785–792. Universidad de Sevilla.

da Silva, Augusto Soares. 2018. Variação linguística e pluricentrismo: novos conceitos e descrições1. In *Actas do XIII Congreso Internacional de Lingüística Xeral: Vigo, 13-15 de xuño de 2018*, pages 838–845. Universidade de Vigo.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2016. An information-theoretic approach to modeling diachronic change in scientific english. *Selected Papers from Varieng-From Data to Evidence (d2e)* .

Degaetano-Ortlieb, Stefania and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33.

Dieguez-Tirado, Javier, Carmen Garcia-Mateo, Laura Docio-Fernandez, and Antonio Cardenal-Lopez. 2005. Adaptation strategies for the acoustic and language models in bilingual speech transcription. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, pages I–833. IEEE.

Diez, Xoán Carlos Lagares. 2008. Sobre a noção de galego-português. *Cadernos de Letras da UFF–Dossiê: Patrimônio cultural e latinidade* 35:61–82.

dos Reis Aguiar, Monalisa. 2007. As reformas ortográficas da língua portuguesa: uma análise histórica, lingüística e ideológica. *Filologia e Linguística Portuguesa* 9:11–26.

Dubert, Francisco and Xulio Sousa. 2016. On quantitative geolinguistics: an illustration from galician dialectology. *Dialectologia: revista electrònica* pages 191–221.

Eden, S Elizabeth. 2018. *Measuring language distance through phonology*. Ph.D. dissertation. UCL.

Ellison, T Mark and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 273–280. Association for Computational Linguistics.

Freixeiro Mato, Xosé Ramón. 2000. Gramática da lingua galega ii. morfosintaxe. *Vigo: A Nosa Terra* .

Gamallo, Pablo, Inaki Alegria, José Ramom Pichel, and Manex Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177.

Gamallo, Pablo, José Ramom Pichel, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications* 484:152–162.

Gao, Yuyang, Wei Liang, Yuming Shi, and Qiuling Huang. 2014. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications* 393(C):579–589.

González, Meritxell. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7.

Gooskens, Charlotte, John Nerbonne, Nathan Vaillette, et al. 2007. Conditional entropy measures intelligibility among related languages. *LOT Occasional Series* 7:51–66.

Goutte, Cyril, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. *arXiv preprint arXiv:1610.00031* .

Gulías, Carme Hermida. 1992. *Os precursores da normalización: defensa e reivindicación da lingua galega no Rexurdimento (1840-1891)*. Ed. Xerais de Galicia.

Han, Aaron Li-Feng, Yi Lu, Derek F Wong, Lidia S Chao, Liangye He, and Junwen Xing. 2013. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 365–372.

Heeringa, Wilbert, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2013. Lexical and orthographic distances between germanic, romance and slavic languages and their relationship to geographic distance. *Phonetics in Europe: Perception and Production* pages 99–137.

Heeringa, Wilbert Jan. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, Citeseer.

Hinkka, Atte et al. 2018. *Data-driven Language Typology*. Master's thesis. University of Helsinki.

Holman, E.W., S. Wichmann, C.H. Brown, V. Velupillai, A. Muller, and D. Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42(2):331–354.

Holmes, Janet and Nick Wilson. 2017. *An introduction to sociolinguistics*. Routledge.

Isphording, Ingo Eduard and Sebastian Otten. 2013. The costs of b abylon—linguistic distance in applied economics. *Review of International Economics* 21(2):354–369.

Jauhiainen, Tommi Sakari, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research* 65:675–782.

Jones, Mari C and Damien Mooney. 2017. *Creating orthographies for endangered languages*. Cambridge University Press.

Kessler, Brett. 1995. Computational dialectology in irish gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 60–66. Morgan Kaufmann Publishers Inc.

Klarer, Mario. 2013. *An introduction to literary studies*. Routledge.

Kloss, Heinz. 1967. Abstand languages and Ausbau languages. *Anthropological linguistics* pages 29–41.

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science* 5(3):171504.

Kondrak, Grzegorz. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Lai, Mirko, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer.

Lapesa, Rafael and Ramón Menéndez Pidal. 1942. *Historia de la lengua española*. Escelicer.

List, Johann-Mattis, Mary Walworth, Simon J Greenhill, Tiago Tresoldi, and Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2):130–144.

Liu, HaiTao and Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin* 58(10):1139–1144.

Mato, Xosé Ramón Freixeiro. 2015. Novas perspetivas sobre o papel do português na revitalização do galego nos primórdios do século xxi. In *Estudos da AIL em Ciências da Linguagem: língua, linguística, didáctica*, pages 217–226. Associação Internacional de Lusitanistas.

Mira, Jorge and Ángel Paredes. 2005. Interlinguistic similarity and language death dynamics. *EPL (Europhysics Letters)* 69(6):1031.

Molina, Giovanni, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2019. Overview for the second shared task on language identification in code-switched data. *arXiv preprint arXiv:1909.13016* .

Monteagudo, Henrique. 2017. A lingua no tempo, os tempos da lingua. o galego, entre o portugués eo castelán. *en M. Negro Romero, R. Álvarez Blanco e E. Moscoso Mato (eds.), Gallæcia. Estudos de linguistica portuguesa e galega. Santiago de Compostela: Universidade de Santiago de Compostela* pages 17–60.

Monteagudo, Henrique and Henrique Monteagudo Romero. 1999. *Historia social da lingua galega: idioma, sociedade e cultura a través do tempo*, vol. 1. Editorial Galaxia.

Monteagudo, Henrique and Antón Santamarina. 1993. Galician and castilian in contact: historical, social and linguistic aspects. *Trends in Romance linguistics and philology* 5:117–173.

Moura, António de Carlos, Angel López, and José Ramom Pichel. 2008. Tmilg (tesouro medieval informatizado da lingua galega). *Procesamiento del lenguaje Natural* 41:303–304.

Muhr, Rudolf. 2013. Codifying linguistic standards innon-dominant varieties of pluricentric languages-adopting dominant or native norms? In *Exploring linguistic standards in non-dominant varieties of pluricentric languages*, pages 11–44. Peter Lang.

Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H Brown, Pamela Brown, Sebastian Sauppe, Eric W Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, et al. 2010. Asjp world language tree of lexical similarity: Version 3 (july 2010). *Retrieved* 10(19):2015.

Nakhleh, Luay, Donald A Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2):382–420.

Nerbonne, John and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*, pages 11–18.

Nerbonne, John and Erhard Hinrichs. 2006. Linguistic distances. In *Proceedings of the workshop on linguistic distances*, pages 1–6. Association for Computational Linguistics.

Nordhoff, Sebastian and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Passerini, Thiago Zilio et al. 2019. *Ocultação de paternidade ou filiação ilegítima? O lugar do galego na origem da língua portuguesa em textos dos séculos XVI e XIX*. Pontifícia Universidade Católica de São Paulo.

Paz, Ramón Mariño. 2008. *Historia de la lengua gallega*. Lincom Europa.

Pérez-Pereira, MIGUEL. 2008. Early galician/spanish bilingualism: contrasts with monolingualism. *A portrait of the young in the new multilingual Spain* pages 39–62.

Pérez-Pereira, Miguel, Margareta Alegren, Mariela Resches, Maria Jose Ezeizabarrena, Carmen Díaz, and Inaki García. 2007. Cross-linguistic comparisons between basque and galician. In *Proceedings from the first European network meeting on communicative development inventories*.

Petroni, Filippo and Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389(11):2280–2283.

Petroni, Filippo and Maurizio Serva. 2011. Automated word stability and language phylogeny. *Journal of Quantitative Linguistics* 18(1):53–62.

Pichel, José Ramom, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to european portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155.

Pichel, José Ramom, Pablo Gamallo, and Iñaki Alegria. 2019a. Cross-lingual diachronic distance: Application to portuguese and spanish. *Procesamiento del Lenguaje Natural* 63:77–84.

Pichel, José Ramom, Pablo Gamallo, and Iñaki Alegria. 2019b. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering* pages 1–22.

Porta, Jordi and José-Luis Sancho. 2014. Using maximum entropy models to discriminate between similar languages and varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 120–128.

Purver, Matthew. 2014. A simple baseline for discriminating similar languages. In *Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 155–160. Association for Computational Linguistics.

Rama, Taraka, Lars Borin, GK Mikros, and J Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification.

Ramallo, Fernando and Gabriel Rei-Doval. 2015. The standardization of galician. *Sociolinguistica* 29(1):61–82.

Richman, Stephen H. 1970. Spanish-portuguese agreement in affixed words. *Studia Neophilologica* 42(1):174–179.

Rissanen, Matti, Merja Kytö, and Minna Palander-Collin. 1993. *Early English in the computer age: Explorations through the Helsinki Corpus*. Walter de Gruyter.

Robl, Affonso. 1982. O galego-português. *Revista Letras* 31.

Rodrigues Fagim, Valentim. 2001. O galego (im) possível. *Santiago: Laiovento* .

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4(4):406–425.

Santamarina, Antón. 2003. Tesouro informatizado da lingua galega. *Santiago de Compostela: Instituto da Lingua Galega. http://ilg. usc. es/TILG/[Consultado: 10/01/2016]* .

Satterthwaite-Phillips, Damian. 2011. *Phylogenetic Inference of the Tibeto-Burman Languages Or on the Usefulness of Lexicostatistics (and" megalo"-comparison) for the Subgrouping of Tibeto-Burman*. Stanford University.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 978-1-937284-19-0.

Seoane, Ernesto Xosé González. 1992. *A ortografía ea gramática do galego nos estudios gramaticais do século XIX e primeiros anos do XX*. Ph.D. thesis, Universidade de Santiago de Compostela.

Serva, Maurizio and Filippo Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)* 81(6):68005.

Singh, Anil Kumar and Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, pages 40–47. Association for Computational

Linguistics.

Specia, Lucia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies* 11(1):1–162.

Swadesh, M. 1952. Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society 96*, pages 452–463.

Teyssier, Paul. 1982. *História da língua portuguesa*. Digital Source.

Torres Feijó, Elias J. 2002. O estudo do mundo lusófono no sistema literário galego. bases metodológicas para o estudo dos sistemas emergentes e as suas relaçons intersistémicas. In *Actas do VII Congresso da Associaçom Internacional de Lusitanistas*, pages 527–539.

Varela Barreiro, Xavier. 2004. Tesouro medieval informatizado da lingua galega. *Santiago de Compostela: Instituto da Lingua Galega [http://ilg. usc. es/tmilg](01/09/13-09/10/13)* .

Vázquez Souza, Ernesto. 2003. A fouce, o hórreo eo prelo: Ánxel casal ou o libro galego moderno. *Sada, A Coruña: Ediciós do Castro* .

Venâncio, Fernando. 2014. *O castelhano como vernáculo do português*. Universidad de Extremadura, Servicio de Publicaciones.

Vilavedra, Dolores and Vilavedra Fernández Vilavedra Fdez. 1999. *Historia da literatura galega*, vol. 2. Editorial Galaxia.

Villa, Laura. 2013. *The officialization of Spanish in mid-nineteenth-century Spain: the Academy's authority*. Cambridge University Press.

Villares, Ramón. 2004. *Historia de Galicia*, vol. 6. Editorial Galaxia.

West, Joel and John L Graham. 2004. A linguistic-based measure of cultural distance and its relationship to managerial values. *Management International Review* 44(3):239–260.

Wichmann, Søren. 2016. How to distinguish languages and dialects. *Computational Linguistics* 1(1).

Yujian, Li and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29(6):1091–1095.

Zampieri, Marcos and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI).

Zampieri, Marcos, Binyam Gebrekidan Gebre, Hernani Costa, and Josef Van Genabith. 2015. Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 66–72.

Zampieri, Marcos, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: the case of portuguese. *arXiv preprint arXiv:1610.00030* .