A methodology to measure the diachronic cross-lingual distance of one language in the midst of two close-related languages based on perplexity

José Ramón Piché Campos¹, Pablo Gamallo Otero², Iñaki Alegria Loínaz³

¹imáxin software, Santiago de Compostela, Galiza; ²CITIUS, Santiago de Compostela, Galiza; ³University of Basque Country, Donostia-San Sebastián, Basque Country

ARTICLE HISTORY
Compiled September 12, 2019

ABSTRACT
The aim of this paper is to apply a corpus-based methodology, based on the measure of perplexity, to automatically calculate the linguistic distance between historical periods of a language and two closely related to it. The three historical corpus has been constructed and collected with the closest spelling to the original on a balanced basis 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 results of our experiments show that Galician and Portuguese are varieties of the same language in the Middle Ages and converges and diverges with Portuguese and Spanish since the last period 19th century. In this process, orthography plays a relevant role. It is remarkable to underline that it is an unsupervised method that we can apply to any threesome of languages.

KEYWORDS
Sections; lists; figures; tables; mathematics; fonts; references; appendices

1. Introduction

The distance between languages is a complex challenge since languages diverge from each other in multiple levels such as phonetics, phonology, lexicography, morphology, syntax, semantics, pragmatics, and so on.

Moreover, if we want to measure the historical distance within a language, between two or more languages or between diatopic variants of the same language and reduce it to a single number, the matter is more complex, since historically languages evolve not only internally but also in relation to others.

The ability to reduce the calculation of this distance to a single robust measure is a task for which historically there have been different approaches from phylogenetics, automatic language identification, dialectology, glottochronology or metrics that calculate distance from corpus mediante métricas basadas en n-grams.

It is this last one that interests us, since character n-grams not only encode lexical and morphological information but also phonological features since phonographic written systems are related to the way languages were pronounced in the past. In addition,
long n-grams (\(n=5\)grams) also encode syntactic and syntagmatic relations as they may represent the end of a word and the beginning of the next one in a sequence. For instance, the 7-gram ion#de# (where '#' represents a blank is a frequent sequence of letters shared by several Romance languages (e.g. French, Spanish, or Galician). This 7-gram might be considered as an instance of the generic pattern “noun-prep-noun” since ion (The stress accent (e.g. i´ on) has been removed to simplify language encoding) is a noun suffix and de a very frequent preposition (of in Galician), introducing prepositional phrases.

En nuestro caso realizaremos el cálculo de las distancias aplicando la metodología de distancias sobre corpus basada en perplexity que ya se ha aplicado para la identificación de lenguas, medir la distancia entre lenguas, el cálculo de las distancias diacrónicas intralingüísticas y las distancias diacrónicas interlingüísticas.

In our case, we have carried out two experiments measuring first the historical intralinguistic distance for Galician, Portuguese and Spanish, and then the historical interlinguistic distance between one language (Galician) and two very close languages (Portuguese and Spanish). To this end, we have applied a methodology for intralinguistic and interlinguistic distance based on perplexity for one or two languages. The objective is to measure the evolution of historical convergence and divergence between the three.

As a result, two kinds of experiments have been issued: the first one uses our perplexity-based method in historical corpora with an orthography closely related to that of the original texts; the second experiment was conducted using transliterated corpora in order to use the same transcribed orthography for all varieties and languages.

We have observed in our experiments the hypotheses of the experts which are as follows:

1. In medieval times until the 14th century GL and PT used similar orthographies and can be considered indistinguishable. Castilian, on the other hand, is already a Romanesque variant close to but distant from Galician-Portuguese.
2. From the 15th century onwards, the Galician-Portuguese language ceased to be written in Galicia and continued to be written in Portugal. Castilian was at its peak.
3. When Galician is rewritten from the second half of the 19th century onwards, the spelling of Galician is indistinguishable from that of Castilian except for a few small strokes.
4. The discovery of medieval songs at the end of the 19th century began in the 20th century to converge in lexicon and morphology with Portuguese, although the use of an orthography indistinguishable from Castilian continues.
5. More specifically, this research tries to observe if the three languages evolved in the same way or whether, on the contrary, there are periods of a language with more changes and to what extent spelling plays a role in that distance.

We have also obtained the following observations from our experiments:

1. Spanish and Portuguese were separated mainly in the XX-1 century.
2. Galician has an equidistance since the 20th century between Portuguese and Spanish when the original orthography is used.
3. Galician is closer to Portuguese when it uses a transcribed orthography common to all three languages.
4. The article is organized as follows: First, introduce some studies on language...
distance Section 2 will be introduced.

Then, the language distance measure is described in Section 3. In addition, in Section 4, the experimental method is introduce and in Section 5 each one of the historical corpus created ad hoc with its main characteristics by language is presented. In Section 6, the two above mentioned experiments are described and the results discussed. Finally, conclusions are addressed in Section 8.

2. Related work

(?) (Kloss, Heinz, 1967) (?)

2.1. Sociolinguistics

2.2. Language Identification

2.3. Linguistic Phylogenetics

2.4. Language Distance

3. Materials and Methods

3.1. Corpus

We have created a historical corpus called Carvalho for each of the languages (Galician, Portuguese and Spanish) which contains balanced fiction and non-fiction texts. The texts which are included in the corpus are in the orthography closest to the original since the experiments that we will carry out will be developed both in original orthography and in an automatically transcribed orthography.

Also to design the historical corpus we have taken into account the guidelines for the creation of historical periods by Corpus Helsinki (Rissanen et al., 1993) dividing into medieval period (XII-XV), modern age (XVI-XVIII), and contemporary age (XIX-XX).

However, although Portuguese and Castilian have sufficient historical corpus for our experiments for these three main periods, this is not the case for Galician. Thus, from the sixteenth century to the second half of the nineteenth century, there is no enough written corpus. This would force us to compare corpus of Galician of the second half of the 19th century with corpus of Portuguese and Spanish in part of the first half of the 19th century. Also, since the late eighteenth century different changes in spelling and standardization are done both Portuguese and Spanish and Galician has set his standard and his orthography in the year 1981.

For these reasons we have subdivided the 19th and 20th centuries into two subperiods of 50 years each, in order to measure the distance between Galician and Portuguese and Spanish since the second half of the 19th century, and to measure the distance between the three languages in a more detailed way, just in periods of greater volume of texts and orthographic and linguistic changes.

As a result we hold the historical corpus Carvalho which is freely available (except for Galician due to copyright) and contains the diachronic corpus for the three languages: Carvalho-GL (for Galician), CarvalhoPT-PT (for Portuguese in Portugal) and Carvalho-ES-ES (for Spanish in Spain).
The next section describes the characteristics of 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

On the one hand, in the case of the Galician, there is sufficient corpus for the medieval period (Galician-Portuguese period) corresponding to the XII-XV centuries. On the other hand, from the beginning of the 16th century until the second half of the 19th century, in the period known as the "Dark Centuries" and the beginning of the so-called "Literary Resurgence", there is not enough written corpus for our experiments. From there, there is enough corpus to be able to calculate both the intralinguistic distance and the cross-lingual distance with the other two languages.

Concerning orthography, we can observe that from the Middle Ages to the present day Galician orthographies oscillate between those very close to Portuguese (medieval period) and those very close to Castilian (modern and contemporary period).

The corpus Carvalho-GL that we have compiled for the medieval period (XII-XV) is part of the corpus TMILG (Galician Language Medieval Treasure) (Varela Barreiro, 2004). For the periods XIX-2, XX-1 and XX-2 we have used texts from the corpus TILG (Galician Language Computerized Treasure). This corpus cannot be made available online to replicate the experiments due to the copyright of both corpus. The Carvalho-GL corpus can not be accessed due to copyright law although its authors can be contacted.

Table 2 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 original spelling were selected, and some samples of fictional and non-fictional documents taking part in the final corpus. As it has been mentioned in the methodology section, we extracted 1.25/1.5M words for the train partitions, and 250/350K words (between 20

3.1.2. Portuguese Corpus

There are sufficient corpus for all historical periods in Portuguese. With respect to orthography since the creation of the Academia das Ciências de Lisboa (Lisbon Academy of Sciences), one of the bodies that regulate the standardization of European Portuguese language created in 1779 in Lisbon, many proposals of different orthographic standards have taken place, e.g.: 1885, 1911, 1945, 1973, 1990, with a great deal of controversy over its partial or total compliance.

It is worth noting that documents that are not in original orthography have been carefully removed; even some modern ones from the early twentieth century. For example, the spelling "ph" was used for the phoneme /f/ in texts of the XIX and XXth centuries, and in many available digital versions the texts were adapted to modern spelling by replacing "ph" with "f". But we discarded these versions. Once all the tests have been obtained, we have divided them into two groups, train and test. In Table 5 we show the number of words in the train and test per period, which are similar to the numbers of the Galician version. Balanced train-test pairs allows us to compute PLD measure without bias.

Table 4 shows the historical work, resources and samples of fictional and non-fictional documents taking part in the final Carvalho-PT-PT corpus. In Spanish there are different well-known historical corpus such as (Corpus do português), but they are
not usually open since they only allow online access to the texts. Furthermore, the texts do not necessarily have to be in spellings close to the original as they may be edited or adapted. This is one of the reasons why we have chosen to create our own diachronic corpus of Spanish that have been obtained mainly from the following online repositories: Corpus Informatizado do Português Medieval, Project Gutenberg, Wiki source, OpenLibrary, Tycho Brahe corpus, Domínio Público, Arquivo Pessoa, Linguateca, Corpus de Textos antigos, Tycho Brahe and Colonia corpus. The Carvalho-PT-PT subcorpus can be accessed at the address:

3.1.3. Spanish Corpus

In the case of Castilian there is also sufficient corpus for all historical periods, as in the case of Portuguese. With respect to orthography, there is also variability up to appearance of the Real Academia Española in 1713 that a standard was proposed, but it took time to consolidate Lapresa and Pidal (1942)

Thus, since medieval times, there has been a will to standardize the Castilian language, starting with Alfonso X in the 13th century (Del Valle, 2013). However, none of the varied orthographies used until the 18th century crystallized. It was only after the reforms of the Royal Academy (RAE) in 1741 that the process of standardization of the written system was actually consolidated as a result of the removal by the RAE of common spelling with other Romance languages such as "ss", "ç" and latinisms (Alatorre, 2002). Thus, a medieval text can be written like this “dios llamo a moysen dela tienda del paramjento y dixole fabla con los fijos de israel y diles todo onbre de vos que diere ofrenda a dios de ganados esto es de buyes o de ovejas o fazer sacrificios” in Biblia Prealfonsi and a nineteenth-century text, is written as follows: “Se embozó en su capa, y se puso a dar paseos. Entonces vio al alemán sentado en un banco, y mirando al mar”, with the same spelling as the current one. Table 7 show the quantitative data of both train and test partitions.

Table 6 shows the historical work, resources and samples of fictional and non-fictional documents taking part in the final Carvalho-ES-ES corpus. In Spanish there are different well-known historical corpus such as (CORDE19 , ADMYTE20 , Corpus del español 21 ), but they are not usually open since they only allow online access to the texts. Furthermore, the texts do not necessarily have to be in spellings close to the original as they may be edited or adapted. This is one of the reasons why we have chosen to create our own diachronic corpus of Spanish that have been obtained mainly from the following online repositories: Project Gutenberg, Wikisource and Open Archive. The Carvalho-ES-ES subcorpus can be accessed at the address:

3.2. Methodology

Hemos probado nuestra metodología anteriormente para medir la intralingual diachronic distance en tres idiomas, dos próximos y uno distante respecto a ellos: Portuguese, Spanish and English. También hemos aplicado esta metodología para medir la distancia histórica entre dos lenguas próximas como es el caso del portugués y el español. Ahora queremos aplicar esta misma metodología para medir la cross-lingual diachronic distance entre una lengua (gallego) y dos idiomas muy próximos (Spanish and Portuguese) con ella.

Para poder realizar el cálculo de la cross-lingual necesitamos previamente comprobar que el intralingual es correcto. Si esto no es correcto la cross-lingual nos va a dar datos incorrectos.
3.2.1. Perplexity-Based Measure

The distance measure of our method is based on perplexity, which is a widely-used evaluation metric for language models. It has been used as a quality measure for language models built with $n$-grams extracted from text corpora Chen and Goodman (1996); Sennrich (2012). It has also been used in very specific tasks, such as to classify formal and colloquial tweets González (2015), and to identify close-related languages Gamallo, Alegria, Pichel, and Agirrezabal (2016). In Gamallo, Pichel, and Alegria (2017), a specific perplexity-based distance, called PLD, has been defined and applied to compute the distance of different European languages. In a previous work Pichel, Gamallo, and Alegria (2018), we applied PLD to measure the diachronic distance between different historical periods of the same language. In the current work, our aim is to apply PLD to measure cross-lingual diachronic distance between two 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, smoothing technique based on linear interpolation, and train/test corpora with 1,25M/250K words, respectively.

3.2.2. Task Description

Our methodology requires a representative and balanced historical corpus for each language. The corpus, divided into different historical periods, consists of two versions: texts with original spelling (or as close as possible to the original), and texts automatically transcribed to a common orthography that phonetically approximates the compared languages. In the current work, we will apply this methodology to two close-related languages: Portuguese (Portugal) and Spanish (Spain). Our method is divided into the following specific sub-tasks:

1. First, we search for textual sources to create our diachronic corpus containing texts with a spelling as close as possible to the original for each language. Once the textual sources have been selected, we eliminate noise from the documents, specially excerpts in other languages.

2. Second, we define linguistic and literary equivalent periods for each language. In the definition of periods, we take into account dates of orthographic changes to better observe the possible variations concerning the distance between languages through the time axis. In the current experiments, we have selected six historical periods for the two compared languages.

3. Third, once we have decided on the common historical periods for all languages, we select a representative and balanced historical corpus with an acceptable size for each language. We try to design a corpus that is representative according to Biber’s criteria Biber (1993): For this purpose, texts from several genres and topics were retrieved. Both non-fiction and fiction texts for each period have been collected, including fiction subgenres such as narrative, poetry, theater, religious texts for the medieval period, etc. Concerning non-fiction texts, essays were mostly used.

4. Once the textual sources of our corpus have been selected and the periods have been established, two subcorpora are created for each period: train and test. In the train partition, we include for each period texts in original spelling in fiction and non-fiction. In order to facilitate a better representation of the language for each period, the fiction and non-fiction texts in both the train and the test were balanced at approximately 50% (the test and train texts are distinct sets). It is worth mentioning that the train and test partitions are not manually annotated as our method is fully unsupervised.
5. A spelling normalization is applied to all the texts and a transcribed version is obtained for each corpus. The common 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 and a variety of vowel articulation. The encoding is thus close to a phonological one and, then, 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 the Spanish “ñ”. Similarly, the palatalized lateral is transcribed as “ly”, simplifying the two different spellings “lh” in Portuguese and “ll” in Spanish.

6. Finally, we perform the PLD calculations between pairs of cross-lingual diachronic periods in both original spelling and in automatic transcription, so as to obtain the corresponding distances. The results are evaluated and analyzed later.

In order to allow researchers to apply the methodology to any language, we have developed a pipeline architecture in Perl, which is freely available.\footnote{https://github.com/gamallo/Perplexity} 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.

### 3.3. Results

#### 3.3.1. Intralingual Diachronic Distance

To ensure that tables are correctly numbered automatically, the \texttt{\textbackslash label} command should be included just before \texttt{\textbackslash end{table}}

#### 3.3.2. Cross-lingual Diachronic Distance

\begin{table}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Type & Class & One & Two & Three & Four & Five & Six \\
\hline
Alpha & A1 & A2 & A3 & A4 & A5 & A6 \\
Beta & B2 & B2 & B3 & B4 & B5 & B6 \\
Gamma & C2 & C2 & C3 & C4 & C5 & C6 \\
\hline
\end{tabular}
\end{table}

\footnote{This footnote shows how to include footnotes to a table if required.}
3.4. Discussion

Acknowledgement(s)

An unnumbered section, e.g. \texttt{section*{Acknowledgements}}, may be used for thanks, etc. if required and included \textit{in the non-anonymous version} before any Notes or References.

Disclosure statement

An unnumbered section, e.g. \texttt{section*{Disclosure statement}}, may be used to declare any potential conflict of interest and included \textit{in the non-anonymous version} before any Notes or References, after any Acknowledgements and before any Funding information.

Funding

An unnumbered section, e.g. \texttt{section*{Funding}}, may be used for grant details, etc. if required and included \textit{in the non-anonymous version} before any Notes or References.

Notes on contributor(s)

An unnumbered section, e.g. \texttt{section*{Notes on contributors}}, may be included \textit{in the non-anonymous version} if required. A photograph may be added if requested.

Nomenclature/Notation

An unnumbered section, e.g. \texttt{section*{Nomenclature}} (or \texttt{section*{Notation}}), may be included if required, before any Notes or References.
Notes

An unnumbered ‘Notes’ section may be included before the References (if using the endnotes package, use the command \theendnotes where the notes are to appear, instead of creating a \section*).

References


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

