Natural Language Engineering **1** (1): 000–000. Printed in the United Kingdom

Measuring diachronic language distance using perplexity. Application to English, Portuguese and Spanish.

José Ramom Pichel[1] , Pablo Gamallo[2] and Iñaki Alegria[3]
[1]***imaxin***|*software, Santiago de Compostela, Galiza*
jramompichel@imaxin.com
[2]*CiTIUS, University of Santiago de Compostela, Galiza*
pablo.gamallo @usc.es
[3]*IXA group, Univ. of the Basque Country (UPV/EHU)*
i.alegria@ehu.eus

(*Received July 30, 2019*)

Abstract

The objective of this work is to set a corpus-driven methodology to quantify automatically diachronic language distance between chronological periods of several languages. We apply a perplexity-based measure to written text representing different historical periods of three languages: European English, European Portuguese and European Spanish. For this purpose, we have built historical corpora for each period, which have been compiled from different open corpus sources containing texts as close as possible to its original spelling. The results of our experiments show that a diachronic language distance based on perplexity detects the linguistic evolution that had already been explained by the historians of the three languages. It is remarkable to underline that it is a unsupervised multilingual method which only needs a raw corpora organized by periods.

1 Introduction

The prevailing view is that distance between two languages or varieties cannot be measured appropriately by using a well-established score because they may differ in many complex linguistic aspects such as phonetics, phonology, lexicography, morphology, syntax, semantics, pragmatics, and so on. In addition, languages change (even their spelling rules) throughout their history (**?**), so it is also difficult to measure the diachronic distance within the same language.

Quantifying all these aspects and reduce them automatically to a single language distance measure between languages or historical periods of the same language is a difficult task which is far from being fulfilled or at least appropriately addressed, perhaps because it has not yet been a priority in natural language processing.

However, there have been different approaches, not always based on corpus linguistics, to obtain language distance measures, namely in phylogenetic studies within historical linguistics (**?**), in dialectology (**?**), in language identification (**?**), and in studies about learning additional languages within the field of second language acquisition (**?**).

Our work falls within the broader scope of understanding language variation. For this we have created a methodology that is corpus-driven and more exploratory in nature in comparison to other (more traditional) approaches. Thus this article proposes a corpus-driven methodology for calculating a diachronic language distance between languages from historical corpora. We consider that the concept of language distance is closely related to the process of language identification (**?**). In fact, the more difficult the identification of differences between two languages or language varieties is, the shorter the distance between them.

In our previous research, perplexity-based measures were used for language identification (**?**), to measure the distance between languages (**?**), and to quantify the diachronic distance in a language (**?**). The results are encouraging because it is an unsupervised method and only raw historical corpora are required.

The objective of the present article is to apply this perplexity-based measure to study and compare the distance among historical periods, performing experiments in three different languages: European English, European Portuguese, and European Spanish, from 12th to 20th century. As a result, two kind of results are

reported: 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. The results of the second experiment show how this orthographic transcription smooths the distance between historical periods of languages.

As the evaluation of the distance is not a trivial task, the objective is to verify whether the distance fits with the opinions of the experts. 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. In addition, previous work **?** can help to compare the historical distance between periods of a language with the current and synchronic distance between languages.

The article is organized as follows: First, some studies on language distance are introduced in Section 2. Then, the experimental method and the language distance measure are described in Section 3, while each one of the historical corpus created *ad hoc* with its main characteristics by language is presented in Section 4. In Section 5, the two above mentioned experiments are described and the results discussed. Finally, a final discussion interpreting the results of the previous experiments and some conclusions are addressed in Sections 6 and 7, respectively.

2 Related Work

Language distance has been measured and defined from different perspectives using different methods. Many of the methods compare lists of words in order to find phylogenetic links or dialectological relations (**?**). In addition, other language identification and language distance approaches have been developed, both working from the comparison of probability distributions using different measures obtained from linguistic corpora. Each of them is described below.

2.1 Language Identification

Language identification is a subfield of Computational linguistics that has been extensively studied. For this purpose language identification has used n-gram language models, word pockets, dictionaries based on word lists and heuristics (spelling, morphology, syntactic characteristics). Among the most relevant studies we can highlight the following: "N-gram-based text categorization" (**?**) which is one of the first papers to use n-grams for Language Identification or “Statistical Identification of Language” (**?**).

Language identification was one of the first natural language processing problems for which a statistical and corpus-based approach was used. The best language identification systems are based on n-gram models of characters extracted from textual corpora (**?**). As a result, 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 (>=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 space) 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. ión) has been removed to simplify language encoding) is a noun suffix and *de* a very frequent preposition (*of* in English), introducing prepositional phrases.

However, there are still big challenges such as classifying some close-related varieties of the same language (e.g. Nicaraguan Spanish and Salvadoran Spanish) and Ausbau languages (**?**) (e.g. Czech and Slovak), or languages by development, which are languages that can be constructed at different historical moments to relate to or to separate. Thus, there have been remarkable works to discriminate among these two kind of languages (**???**), and also for language detection on noisy short texts such as tweets (**??**)

In recent years reasonable results have been achieved even for very closely related varieties using corpus-based strategies. For instance, **?** reported an approach using a log-likelihood estimation method for language models built on orthographical (character n-grams), lexical (word unigrams) and lexico-syntactic (word bigrams) features. As a result, they reported an extremely high accuracy of 0.998 for distinguishing between European Portuguese and Brazilian Portuguese, and 0.990 for Mexican and Argentinian Spanish.

To conclude, the VarDial workshop has become the reference in this area in recent years (**?**). In the German Dialect Identification task in 2016 the best language identification systems were based on n-gram models (**?**). Finally, in 2018 the two best systems are using n-gram models (character 4-gram in the first ranked system).

2.2 Linguistic Phylogenetics

According to **?**, genetic linguistics (also known as "phylogenetics" or "comparative-historical linguistics") and dialectology are the most popular fields dealing with language distance. This author claimed that "traditionally, dialectological investigations have focused mainly on vocabulary and pronunciation, whereas comparative-historical linguists put much stock in grammatical features". However, "we would expect the same kind of [language distance] methods to be useful in both cases" (**?**, p. 7).

The objective of linguistic phylogenetics, a sub-field of historical and comparative linguistics, is to classify the languages building a rooted tree describing the evolutionary history of a set of related languages or varieties.

In order to automatically build phylogenetic trees, many researchers made use of a specific technique called *lexicostatistics*, which is an approach of comparative linguistics that involves quantitative comparison of lexical cognates, which are words with a common historical origin (**?????**). More precisely, lexicostatistics is based on cross-lingual word lists (e.g. Swadesh list (**?**) or ASJP database (**?**)) to automatically measure distances using the percentage of shared cognates. Among these studies, **?**, **?** and **?** can be highlighted. Levenshtein distance among words (**?**) in a cross-lingual list is one the most common metrics used in this field (**?**). **?** present a method, called PHILOLOGICON, to build language taxonomies comparing lexical forms. The method only compares words language-internally and never cross-linguistically. Finally, **?** and **?** test four techniques to construct phylogenetic trees from corpora: cross–entropy, cognate coverage distance, phonetic distance of cognates and feature N-Gram. They conclude that these measures can be very useful for languages which do not have linguistically hand-crafted lists.

Finally, using perplexity-based distance we built a tree which represents the current map of similarities and divergences among the main languages of Europe (**?**).

2.3 Language distance

To measure language distances, there were first approaches such as those of **?** and **?** from cross-lingual comparison of phonetic forms, "but some researchers have argued against the possibility of obtaining meaningful results from crosslingual comparison of phonetic forms" (**?**).

More complex language models have been built from large cross-lingual and parallel corpora to obtain language distances automatically. In these works, models are mainly built with distributional information on words, i.e., they are based on co-occurrences of words, and therefore languages are compared by computing cross-lingual similarity on the basis of word co-occurrences (**???**).

Recently, **?** have presented an information-theoretic approach based on entropy to investigate diachronic change in scientific English, **?** have used cross-entropy to measure distances and **?** have used relative entropy for detection and analysis of periods of diachronic linguistic change.

Works that address other computational linguistics tasks from a diachronic perspective (e.g. stance evolution reported in **?**) can also be cited.

3 Methodology

The proposed method consists of applying a distance measure on different periods of a historical corpus. In the following, we define the distance measure (Subsection 3.1) and how it is used in a historical corpus (Subsection 3.2).

3.1 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. It has also been used in very specific tasks, such as to classify between formal and colloquial tweets (**?**), to identify varieties of very related languages (**?**), or to measure distances among languages (**?**).

More formally, the perplexity, PP, 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.

A Perplexity-based distance between two languages or two periods of the same language is defined by comparing the n-grams of a text in one language or period of language with the n-gram model trained for the other language or period of language. This comparison must be made in the two directions as PP is a divergence with asymetric values. Then, the perplexity of the test text ***CH*** in language $L2$, given the language model ***LM*** of language $L1$, as well as the perplexity of the test text in $L1$, given the language model of $L2$, are used to define the perplexity-based language distance, PLD, between $L1$ and $L2$ as follows:

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

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 $PP(CH_{L1}, LM_{L2})$.

PLD distance has been firstly defined in **?**. In order to have comparable results, we configured the PLD distance and the corpora with the same hiper-parameters than those used in that work to measure contemporary European languages. So, PLD has been configured with 7-grams and train/test corpora with 1,25M/250K words, respectively.

3.2 Task Description

Our methodology is based on the application of PLD measure to a language historical corpora (also called "diachronic corpora"), in order to obtain a diachronic language distance between periods both in original spelling and transcribed spelling. In the experiments reported later, it will be applied to three international languages in their European variety: English (United Kingdom), Portuguese (Portugal) and Spanish (Spain). For this purpose, a representative and balanced historical corpus is required for each language.

The corpora are divided into two parts: train and test subcorpora. Also, train and test must be divided into different language periods, which have been previously defined according to historical linguistics criteria. Taking into account PLD measure and perplexity requirements, the test corpus should contain roughly 20% number of words with regard to the train corpus. It is worth mentioning that the train partitions are not manually annotated as our method is fully unsupervised. Finally, we must emphasize that no test partition is included in the train, being a different corpus.

More precisely, to apply PLD on diachronic corpora for computing the distance between periods, our method is divided into the following specific tasks:

To obtain diachronic corpora in original spelling: First, we need to obtain text sources to create our di-

achronic corpora with a spelling as close as possible to the original for each language. It is important to check first if these corpora already exist as open access and if they are total or partially in orthography as close as possible to the original. Once the textual sources have been selected, we must eliminate noise from the documents, specially texts in other languages.

To define historical periods for diachronic corpora: Attending to **?**: "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". These periods of linguistic change and lexical and grammatical features contributing to change could be detected automatically for each language using the method of **?** or the method of identification of stages done by **?**. Because we want to compare the historical change in the three languages, a matter that will be explained in (6), we have chosen to define common periods for the three languages manually. Thus, we have chosen to use broader historical periods: medieval period (XII-XV), modern age (XVI-XVIII) and contemporary age (XIX and XX). As in our case we have carried out experiments for English, Portuguese and Spanish, and the latter have undergone different orthographic changes since the end of the 18th century, we have divided the contemporary age into two subperiods per century.

To select representative/balanced diachronic corpora: We must select representative and balanced historical corpora. In order to design a corpus that is representative according to **?**: "variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistics distributions in a language." For this purpose, texts from several genres and topics must be retrieved. For our corpus, texts from both non-fiction and fiction for each period have been collected, including fiction subgenres such as: narrative, poetry, theater, religious texts for the medieval period, etc., whereas for the non-fiction essays were mostly used. In addition to the size of the corpus, we have opted for the same size as the Helsinki Corpus of Historical English (**?**): "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".

To set Train and Test subcorpora in original spelling: Once the textual sources of our corpora have been selected and the periods have been established, two subcorpora are created for each period: one for the train and the other for the test. In the train, we include for each period texts in original spelling in fiction and non-fiction. In total there must be at least 1,250,000 words per period. In the test we do the same, obtaining per period original spelling texts in fiction and non-fiction with a number of words of at least 20% of the train, i.e. between 250,000 and 350,000 words. 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 per period should be balanced at approximately 50% (the test and train texts are distinct sets).

To set Train/Test subcorpora in transcribed spelling: A spelling normalization is applied on all the texts and a transcribed version is obtained for each corpus. The final 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 spellings "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. The palatal affricate sound in English, represented by the spelling "ch", is transcribed into "ĉ", as well as in Spanish and Portuguese.

To compute PLD: Finally, we perform the PLD calculations between all the different periods in the two spellings: original and transcribed texts.

This strategy was applied to a specific historical corpus and the results are evaluated and analyzed in the next section.

Wikisource[2]	*WHEN that Aprilis, with his showers swoot*, *sweet The drought of March hath pierced to the root, And bathed every vein in such licour, Of which virtue engender'd is the flower*
Corpus Prose and Verse[3]	*WHan that Apprille / with his shouris soote the drought of Marche / hath pershid to the roote and bathed euery veyne in swich licoure of which vertue / engendrid is the floure*

Table 1. *The same excerpt from the medieval book The Canterbury Tales by George Chaucer. The first row, extracted from Wikisource, is edited while the second one, from Corpus of Middle English Prose and Verse, is in original spelling*

4 Corpus

The Corpus that we have used for our experiments, called Carvalho, is freely available[1] and contains the diachronic corpus for the three languages: Carvalho-EN-UK (for English in the United Kingdom), Carvalho-PT-PT (for Portuguese in Portugal) and Carvalho-ES-ES (for Spanish in Spain).

Initially, our intention was to classify the historical periods in three fundamental stages: medieval period (XII-XV), modern age (XVI-XVIII), and contemporary age (XIX-XX), following the classification for English provided by Corpus Helsinki (**?**). However, as we have previously explained in stage 2 of our methodology ("Define historical periods for diachronic corpora") the six historical periods used to divide temporal axis of the three target languages are: XII-XV, XVI-XVIII, XIX-1, XIX-2, XX-1, XX-2.

One of the main problems in the process of selecting texts from different historical periods is that, on many occasions, the same text can appear in original spelling in one source but also edited or adapted in another one. For example, Table 1 shows the same English medieval excerpt extracted from two different sources: one version has been edited and adapted (first row), and the other version is close to the original (second row). Given that our experiments will be carried out on texts written in original spelling or automatically transcribed from the original spelling, we have decided to create a historical corpus whose spelling has never been edited or modified, being as close as possible to the original. Bearing this aim in mind, adapted or edited versions have been ruled out.

In the following section we will outline the characteristics of the diachronic corpus that we have created for each language. We will focus on the resources used to extract all the texts of our corpus, their distribution in fiction and non-fiction, as well as the size of the different partitions. In addition, some historical studies are cited for each language. These references were used to identify the periods of each language, situate the texts in their corresponding period and classify them by genre (fiction / non-fiction). They were also useful to learn how to distinguish between original and adapted spelling.

4.1 English Corpus

Table 2 shows some relevant information required to build the Carvalho-EN-UK 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 par-

[1] https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho
[2] https://en.wikisource.org/wiki/The_Canterbury_Tales/General_Prologue
[3] https://quod.lib.umich.edu/c/cme/AGZ8235.0001.001/1:3.1?rgn=div2;view=fulltext
[4] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html
[5] http://www.helsinki.fi/varieng/CoRD/corpora/ZEN/
[6] http://www.gutenberg.org/catalog/
[7] https://openlibrary.org/
[8] https://en.wikisource.org/wiki/Main_Page

studies	"A history of the English language" (**?**), "The Short Oxford History of English Literature" (**?**), "The Story of English: How the English Language conquered the World" (**?**), "The history of English" (**?**), "The historical development of the English spelling system" (**?**), "An Historical Study of English Function, form and change" (**?**)
sources	The Helsinki Corpus of English Texts[4], Zurich English newspaper corpus (ZEN)[5], Project Gutenberg[6], OpenLibrary[7], Wikisource[8].
fiction	"Canterbury Tales" by George Chaucer, "The Complete works" by Shakespeare, "Dracula" by Bram Stoker, "The fifth child" by Doris Lessing
non-fiction	"Þe Story of Englande als Robert Mannyng", "Theological Tracts" by Bacon, "The blind watchmaker" by Richard Dawkins

Table 2. *Qualitative data on Carvalho-EN-UK corpus: historical studies, corpus resources and an ordered sample from the Middle Age to the 20th century of fictional and non-fictional writings.*

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
Train corpus (Words)	1,480,573	1,611,503	1,468,379	1,341,374	1,526,614	1,531,837
Test corpus (Words)	354,056	344,389	342,543	336,240	354,071	360,394
Proportion (Test/Train)	24.11%	21.37%	23.32%	25.06%	23.19%	23.52%

Table 3. *Size of Train and Test partitions in Carvalho-EN-UK.*

titions, and 250/350K words (between 20% and 25% of the train) for the test ones. Table 3 shows the quantitative data of all partitions in Carvalho-EN-UK.

4.2 Portuguese Corpus

Table 4 shows the historical work, resources and samples of fictional and non-fictional documents taking part in the final Carvalho-PT-PT corpus. 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 texts 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 English version. Balanced train-test pairs might help to compute PLD measure without bias.

[9] http://www.tycho.iel.unicamp.br/corpus/index.html
[10] http://corporavm.uni-koeln.de/colonia/
[11] https://www.gutenberg.org/browse/languages/pt
[12] https://en.wikisource.org/wiki/Category:Portuguese_authors
[13] https://openlibrary.org/
[14] http://arquivopessoa.net/textos/
[15] https://www.linguateca.pt/
[16] http://alfclul.clul.ul.pt/teitok/cta/index.php?action=textos
[17] http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=16090

studies	History of Portuguese Language (**?**), Historical Phonology and Morphology of the Portuguese Language (**?**), *História da Literatura Portuguesa* (History of Portuguese Literature) (**?**), *História de Portugal em datas* (History of Portugal in a timeline) (**?**), *História de Portugal* (History of Portugal) (**?**) and *História concisa de Portugal* (Brief history of Portugal) (**?**)
sources	Tycho Brahe corpus[9] (**?**), Colonia[10] (**?**), *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) (**?**), Project Gutenberg, specially for the XIX century[11], Wiki source[12], OpenLibrary[13], Arquivo Pessoa[14], Linguateca[15], *Corpus de Textos antigos* (Corpus of old texts)[16], *Domínio Público*[17]
fiction	Cantigas de Dom Dinis, "Cancioneiro Geral de Resende", "Elegia" by Barbosa du Bocage, "A relíquia" by Eça de Queiroz, "Elegias" by Teixeira de Pascoaes, "Caim" by José Saramago
non-fiction	"Chronica de Dom João I", "Documentos Notariais", "Opúsculos" by Alexandre Herculano, "Descobrimento de Philipinas", "Páginas Archeologicas" by Felix Alves, "Este mundo da injustiça globalizada" by Saramago

Table 4. *Qualitative data on Carvalho-PT-PT corpus: historical studies, corpus resources and an ordered sample from the Middle Age to the 20th century of fictional and non-fictional documents.*

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
Train corpus (Words)	1,509,774	1,449,148	1,262,976	1,612,320	1,325,353	1,688,787
Test corpus (Words)	305,773	310,405	253,466	334,145	336,880	363,693
Proportion (Test/Train)	20.25%	21,41%	20.06%	20.72%	25,41%	21.53%

Table 5. *Size of Train and Test partitions in Carvalho-PT-PT.*

4.3 Spanish Corpus

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 corpora CORDE[18], ADMYTE[19], Corpus del español [20], 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 corpora of Spanish that have been obtained mainly from the following online repositories: Project Gutenberg, Wikisource and Open Archive.

Since medieval times, there has been a will to standardize the Castilian language, starting with Alfonso X in the 13th century (**?**). 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 (**?**). 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*

[18] `http://corpus.rae.es/cordenet.html`
[19] `http://www.admyte.com/contenido.htm`
[20] `https://www.corpusdelespanol.org/`

studies	"Historia de la lengua española" (History of Spanish Language) by Rafael Lapesa (**?**), "Los 1001 años de la Lengua española" (1001 years of Spanish Language) by Antonio Alatorre (**?**)
sources	Project Gutenberg, Wikisource, Open Archive
fiction	"Libro Buen Amor" by Arcipreste of Hita, "Don Quixote de la Mancha" by Cervantes , "La Gaviota" by Fernán Caballero, "La Regenta" by Leopoldo Alas Clarín, "Platero y Yo" by Juan Ramón Jiménez, "Pascual Duarte" by Camilo José Cela
non-fiction	"General estoria" by Alfonso X, "Naufragios" by Cabeça de Vaca, "Historia de Castilla", "Historia del Derecho español" by Eduardo Hinojosa, "Historia de la decadencia de España by Cánovas" del Castillo, "Análisis del Protágoras de Platón" by Gustavo Bueno

Table 6. *Qualitative data on Carvalho-ES-ES corpus: historical studies, corpus resources and an ordered sample from the Middle Age to the 20th century of fictional and non-fictional documents.*

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
Train corpus (Words)	1,317,635	1,302,628	1,368,232	1,315,262	1,252,998	1,231,419
Test corpus (Words)	314,428	314,596	311,032	257,119	253,039	250,198
Proportion (Test/Train)	23.86%	24.15%	22.73%	20.72%	20,19%	20.31%

Table 7. *Size of Train and Test partitions in Carvalho-ES-ES.*

vos que diere ofrenda a dios de ganados esto es de buyes o de ovejas o fazer sacrifiçios" 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.

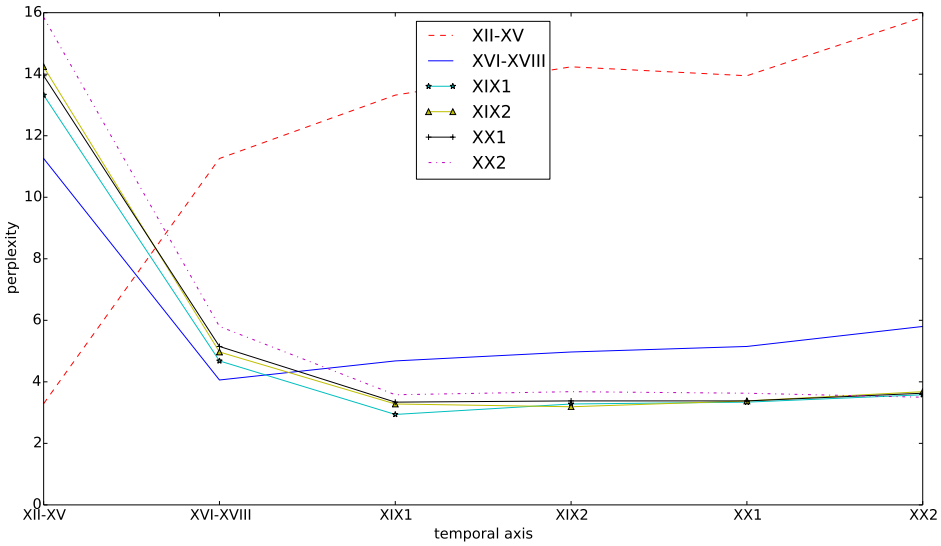
5 Experiments

Since our aim is to test our methodology in different languages (English, Portuguese and Spanish), linguistic models were generated using a collection of documents in various periods of each language, as explained above. These documents are not translations of each other and are made up of a balanced combination of genres (both fiction and non-fiction) from period to period. As a result, we created a set of comparable and balanced corpora of fiction and non-fiction in six different periods of the three languages containing relevant text in original orthography. The experiments consist in calculating the PLD distance between pairs of periods within each language in two steps: first, using texts written with original spelling, and then using the same texts automatically transcribed into a common ortography.

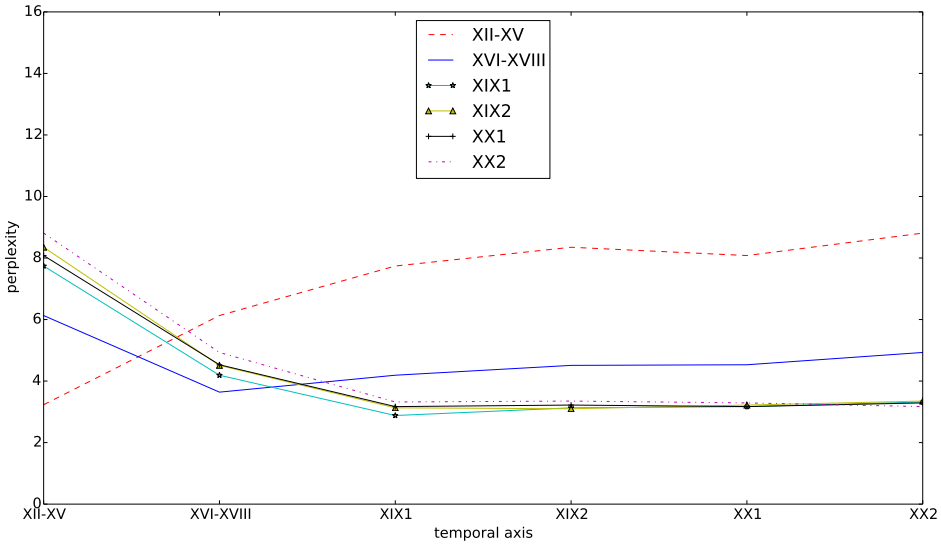
To perform these experiments, a set of scripts has been developed (https://github.com/gamallo/Perplexity) to create a train 7-gram diachronic language model, period by period. As a result, six 7-gram diachronic language models are obtained. Then, we have generated 7-gram models from all test corpora.

Once all models have been created, PLD is computed for each possible train-test pair of models in original spelling. Next, the experiments are performed again to obtain the PLD between transcribed models (as it was described in the Methodology section).

Next, we will show the PLD computation for each language with original and transcribed spelling, and discuss the results.



(a) Original spelling



(b) Transcribed spelling

Fig. 1. In (a) we compare the English PLD distances between XII-XV and XX-2 across all periods in original spelling. In (b) the same comparison using a transcribed spelling.

5.1 English

First, we will see in Table 8 the results of calculating the PLD in original orthography between all periods of English within the Carvalho-EN-UK corpus. Second, Table 9 shows the results of performing the same experiment but with the characteristic of transcribing all periods to the same spelling. Finally, in order to see more clearly the data, Figure 1(a) compares the distance evolution across all periods in original spelling while Figure 1(b) compares the same but with transcribed spelling.

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	3.29	11.26	13.32	14.24	13.95	15.85
XVI-XVIII	11.26	4.06	4.68	4.97	5.15	5.80
XIX-1	13.32	4.68	2.94	3.28	3.34	3.58
XIX-2	14.24	4.97	3.28	3.19	3.38	3.68
XX-1	13.95	5.15	3.34	3.38	3.38	3.63
XX-2	15.85	5.80	3.58	3.68	3.63	3.50

Table 8. *PLD diachronic measure in original spelling (Carvalho-EN-UK corpus)*

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	3.23	6.13	7.74	8.35	8.08	8.81
XVI-XVIII	6.13	3.64	4.19	4.51	4.53	4.93
XIX-1	7.74	4.19	2.88	3.13	3.17	3.32
XIX-2	8.35	4.51	3.13	3.10	3.22	3.35
XX-1	8.08	4.53	3.17	3.22	3.17	3.29
XX-2	8.81	4.93	3.32	3.35	3.29	3.17

Table 9. *PLD diachronic measure in a common transcribed spelling (Carvalho-EN-UK_norm corpus.)*

5.1.1 PLD with original spelling

Different phenomena can be observed in Table 8 different phenomena: first, the medieval period is steadily and considerably distanced from all other periods: the PLD distance from XVI-XVIII is 11.26 while its distance from the second half of the XX century is 15.85; second, from XVI-XVIII to XX-2 figures are quite homogeneous: the highest PLD value between different periods is 5.80 while the lowest point is 3.28 (between the two halves of the 19th century).

In the case of Figure 1(a) it can be seen more clearly how the medieval period (XII-XV) is progressively separated from the other periods of English, the distance being very large with respect to all periods. In the case of the XV-XVIII period, the distance with regard to the medieval period is much larger than with regard to the rest of the periods (XIX and XX). Finally, it is perceived that there is very little difference between the four subperiods of the nineteenth and twentieth centuries: 0.40 between the maximum value and the minimum one.

5.1.2 PLD with transcribed spelling

In a second experiment, we have converted the Carvalho-EN-UK corpus into a new common spelling: Carvalho-EN-UK_norm. After this transformation, the same experiment as for Carvalho-EN-UK has been performed. The same will be done for Portuguese (Carvalho-PT-PT and Carvalho-PT-PT_norm) and Spanish (Carvalho-ES-ES and Carvalho-ES-ES_norm).

By unifying the same orthography between all the English periods, we see that the medieval period is much less distant, though still far away, from the rest of the English periods. Thus, in Table 9, we can observe how the PLD drops from 11.26 to 6.13 with regard to the XVI-XVIII period , an important decrease in distance, only caused by orthographic normalization.

In the case of Figure 1(b) we can see again a significant drop in the distance between the medieval period and the rest of the English periods, making the distance even smoother with the second half of the twentieth century. At the same time, we can also observe that the distance between the XV-XVIII period and the rest of the periods (XIX and XX) is no significantly smaller. Finally, the four periods of the nineteenth and twentieth centuries (submatrix 4x4), once normalized, are practically identical in terms of PLD distance:

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	*2.91*	5.47	6.80	7.21	7.69	7.73
XVI-XVIII	5.47	*2.79*	6.60	6.84	7.11	7.40
XIX-1	6.80	6.60	*3.97*	4.40	4.38	5.08
XIX-2	7.21	6.84	4.40	*3.09*	4.13	4.79
XX-1	7.69	7.11	4.38	4.13	*3.77*	4.69
XX-2	7.73	7.35	5.08	4.79	4.69	*3.08*

Table 10. *PLD diachronic measure in original spelling (Carvalho-PT-PT corpus)*

from 3.13 (the lowest value in Table 9) to 3.35 (the highest one). In fact, these values are on the same scale as the ones we get when we compare the periods with themselves on the diagonal.

5.1.3 Discussion for English results

These results allowed us to find that the distance between the medieval period and the second half of the twentieth century taking into account the original spelling is very substantial (PLD: 15.85 with the same size for train and test). In addition, it can be observed how the distance starts from the Renaissance period with a PLD of 11.29 and progressively reaches the PLD mentioned above.

But after converting all periods to a comparable spelling, the PLD falls to 8.81, a distance slightly greater than that indicated by perplexity, in the same article (**?**), between current Spanish and current Portuguese, with a PLD of 7.77. That is to say, it could be claimed that medieval English (XII-XV) and the English of the last half of the XX-2 century are different but very close-related languages after sharing a common spelling.

Finally, we can see how since the Renaissance period (XV-XVIII) the English language does not undergo important changes, with only a small distance between this period and the rest of historical periods (XIX and XX). The diachronic distance is practically irrelevant between these last two periods.

5.2 Portuguese

The same two experiments will be performed for Portuguese: the first one consists in applying PLD measure on a Portuguese historical corpus (Carvalho-PT-PT) keeping the original spelling to all and between all historical periods. In the second experiment, we apply the same PLD measure to the same historical documents, but transcribed automatically by means of a normalization process.

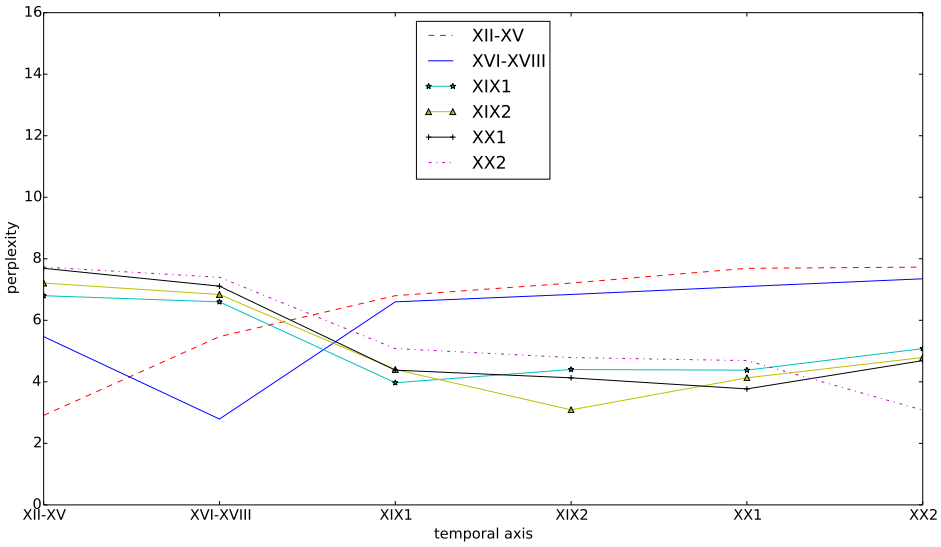
5.2.1 PLD with original spelling

We can observe in Table 10 the following phenomena:

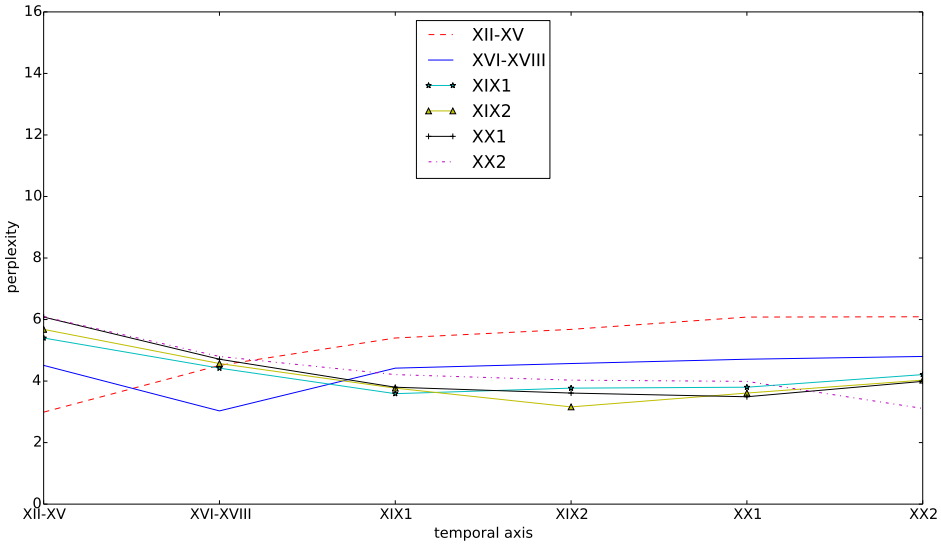
First, the medieval period is progressively but gently distant from the rest of the language periods: 5.47 from XVI-XVIII, 6.80 from the first half of the XIXth century, and 7.73 from the second half of the XXth century.

Second, the differences in PLD between the recent periods (XIX-1, XIX-2, XX-1 and XX-2) are small but distinguishable, namely almost 1 point between the extreme values: the distance in the 4x4 sub-matrix from the 19th-1st century to the 20th-2nd century has a maximum PLD value of 5.08 between the first half of the 19th century and the second half of the 20th century, while the minimum PLD value is 4.13 between the second half of the 19th century and the first half of the 20th century.

Finally, Figure 2(a) helps us see that the distance between the XIX and XXth centuries with regard to the two oldest periods (XII-XV and XVI-XVIII) is quite wide but quite similar. Hence, since the XIXth century, the two previous periods are seen as distant but almost indistinguishable.



(a) Original spelling



(b) Transcribed spelling

Fig. 2. In (a) we compare the Portuguese PLD distances between XII-XV and XX-2 across all periods in original spelling. In (b) the same comparison using a transcribed spelling.

5.2.2 PLD with transcribed spelling

In this new experiment on Carvalho-PT-PT_norm, the PLD distances shown in Table 11 are very similar to those of the previous experiment (Tab 10). However, if we look carefully at Table 11, it can be observed that the orthographic transformation approximates some periods that were separated in the original orthography.

In Figure 2(b), we can see how the transformation of orthography turns a relevant leap between the Renaissance period (XV-XVIII) with respect to the 19th-1st and successive centuries, into a much shorter distance. Tables 10 and 11 show how the difference drops: with original orthography, the PLD distance between XVI-XVIII and XIX-1 is 6.60, by contrast, for the same periods, the distance drops to 4.42 with normalized spelling. This trend continues until the last half of the XX-2 century, where the PLD falls from 7.40 in original orthography to 4.80 in normalized one.

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	*2.99*	4.51	5.40	5.68	6.08	6.09
XVI-XVIII	4.51	*3.03*	4.42	4.57	4.71	4.80
XIX-1	5.40	4.42	*3.59*	3.77	3.80	4.21
XIX-2	5.68	4.57	3.77	*3.16*	3.61	4.03
XX-1	6.08	4.71	3.80	3.61	*3.49*	3.99
XX-2	6.09	4.80	4.21	4.03	3.99	*3.11*

Table 11. *PLD diachronic measure in a common transcribed spelling (Carvalho-PT-PT_norm corpus.)*

Finally, it can be observed that the differences in PLD between the periods XIX-1, XIX-2, XX-1 and XX-2 when orthography is normalized remain small but still distinguishable. The distance in the 4x4 sub-matrix from the 19th-1st period to 20th-2nd period has its maximum PLD value at 4.21 between the first half of the 19th century and the second half of the 20th century, while its minimum PLD score is 3.61 between the second half of the 19th century and the first half of the 20th century.

5.2.3 Discussion for Portuguese results

The XII-XV and XVI-XVIII periods have a PLD distance of 5.47 with the original orthography and 4.51 with the normalized spelling. From this, we can infer that the Portuguese of the Middle Ages (*galego-português*) and the Portuguese of the Renaissance, even if they keep some distance, have small orthographic differences.

Furthermore, it can be concluded that the distance between the medieval period and the second half of the 20th century is not very high, taking into account both original and transcribed orthography. After spelling normalization this distance goes from 7.73 to 6.09. By considering the results reported in **?**, this last score is in the same range as the distance between diatopic varieties or *Ausbau* languages (e.g. Bosnian-Croatian, $PLD = 5.90$). We could affirm that medieval Portuguese and Portuguese from the second half of the 20th century are historical variants of the same language.

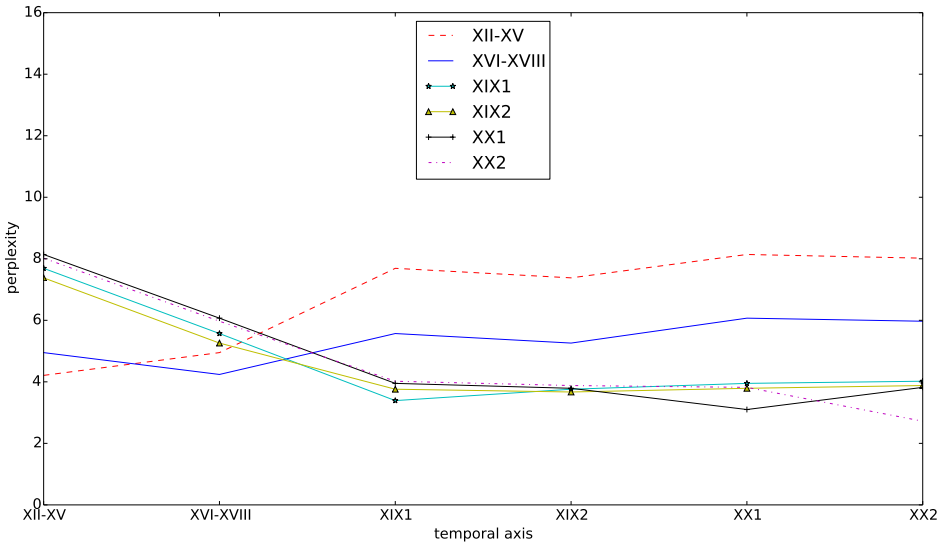
For the rest of the periods, we can infer that orthography is relevant in the first half of the 19th century to mark differences with the medieval and Renaissance periods. PLD distance between periods XVI-XVIII and XIX-1 goes from 6.60 (with original orthography) to 4.42 (with transcribed orthography). It is worth noting that, in the last quarter of XVIIIth century, Portuguese language started to deploy an etymological orthography very related to Latin and Greek (e.g. *philosofia* instead of *filosofia*).

We also see in Figures 2(a) and 2(b) that the distance, although small, between the different subperiods of the nineteenth and twentieth centuries with original orthography does not disappear if the orthography is normalized. From this, we can deduce that orthography is not totally relevant to keep significant distances in the 19th and 20th centuries in European Portuguese.

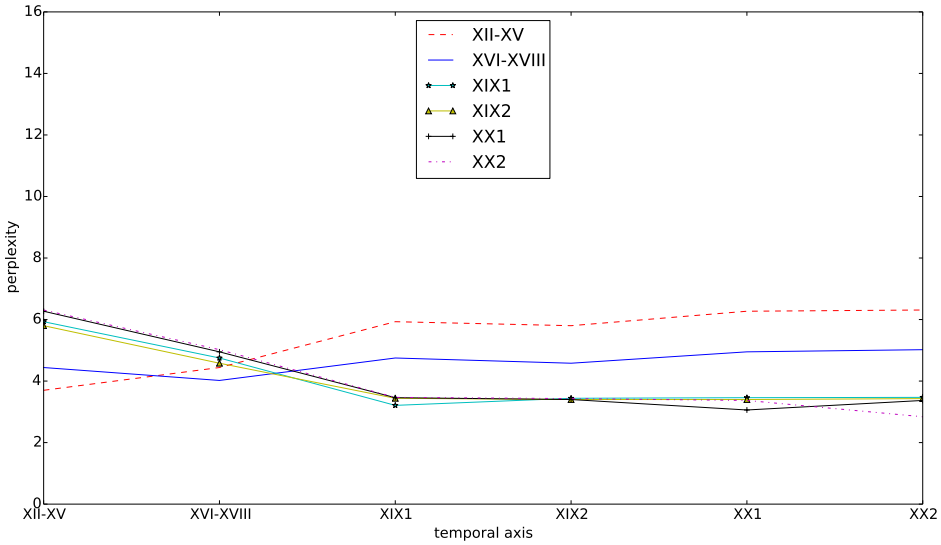
To sum up, we may claim, on the one hand, that historically Portuguese, with a distance ($PLD = 7.73$) between the medieval period and the second half of the twentieth century in the original spelling, has a relevant distance. And on the other hand, if we take into account the PLD distances between the European languages reported in **?** built with a common transcribed orthography, and compare it with the distance in a common orthography for all periods of Portuguese, we can say that Portuguese, regarding its most distant periods (XII-XV vs XX-2: 6.09), is in the same range as the distance between diatopic varieties or languages *Ausbau*. (e.g. Bosnian-Croatian, PLD =5.90).

5.3 Spanish

Here, too, both experiments have been carried out. The first one consists in applying PLD measure on a Spanish historical corpus (Carvalho-ES-ES) and a second one applying the same PLD measure to the normalized documents.



(a) Original spelling



(b) Transcribed spelling

Fig. 3. In (a) we compare the Spanish PLD distances between XII-XV and XX-2 across all periods in original spelling. In (b) the same comparison using a transcribed spelling.

5.3.1 PLD with original spelling

Table 12 shows that the Renaissance and Medieval periods are quite similar: only 4.95 between the two periods. However, between the medieval period and the first half of the 19th century (XIX-1), the PLD reaches 7.69, increasing progressively in later periods and reaching the maximum in the second half of the 20th century (XX-2), with a PLD of 8.02.

Besides, the differences in PLD between the periods XIX-1, XIX-2, XX-1 and XX-2 are small but distinguishable: almost 1 point, as in Portuguese. More precisely, the distance has a maximum PLD point of 5.08 between the first half of the 19th century and the second half of the 20th century, and a minimum PLD point of 4.13 between the second half of the 19th century and the first half of the 20th century.

In Figure 3(a), it can be clearly seen that the first half of the nineteenth century (XIX-1) is almost equally

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	*4.21*	4.95	7.69	7.38	8.14	8.02
XVI-XVIII	4.95	*4.24*	5.57	5.26	6.07	5.97
XIX-1	7.69	5.57	*3.39*	3.76	3.95	4.02
XIX-2	7.38	5.26	3.76	*3.67*	3.79	3.88
XX-1	8.14	6.07	3.95	3.79	*3.10*	3.82
XX-2	8.02	5.97	4.02	3.88	3.82	*2.72*

Table 12. *PLD diachronic measure in original spelling (Carvalho-ES-ES corpus)*

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	*3.70*	4.44	5.93	5.80	6.27	6.31
XVI-XVIII	4.44	*4.02*	4.75	4.58	4.95	5.02
XIX-1	5.93	4.75	*3.21*	3.44	3.46	3.47
XIX-2	5.80	4.58	3.44	*3.40*	3.40	3.43
XX-1	6.27	4.95	3.46	3.40	*3.06*	3.37
XX-2	6.31	5.02	3.47	3.43	3.37	*2.84*

Table 13. *PLD diachronic measure in a common transcribed spelling (Carvalho-ES-ES_norm corpus.)*

distant from the medieval period as from the Renaissance period. Finally, the XIX-1 period is almost linearly distanced from the rest of recent periods: XIX-2, XX-1 and XX-2.

5.3.2 PLD with transcribed spelling

If we look at Table 13 we see that the orthographic transformation approximates in a significant way the medieval period and the Renaissance periods from the rest, as in Portuguese (recall that in English only the medieval period approached the rest with the transcribed orthography).

Table 13 shows that the Renaissance and Medieval periods continue to be very similar with a PLD of 4.44 between the two periods. Furthermore, between the medieval period and the first half of the 19th century (XIX-1)the PLD decreases to 5.69, an important leap that increases progressively in later periods and reaches the maximum in the second half of the 20th century (XX-2), with a PLD of 6.31, well below the PLD of 8.02 in original orthography.

Figure 3(b) shows more clearly how orthography is relevant for approaching the medieval (XII-XV) and Renaissance (XV-XVIII) periods with respect to the XIX-1 and successive centuries.

Finally, it is also observed that orthography unifies the distances of the four Spanish subperiods in the 19th and 20th centuries, so the greatest PLD distance between these subperiods is just 0.1: it goes from 3.37 (lowest value) to 3.47 (highest value).

5.3.3 Discussion for Spanish results

The XII-XV and XVI-XVIII periods have a PLD distance of 4.95 in the original spelling and 4.44 in the normalized spelling. From this, it is deduced that the medieval Spanish and the Golden Age periods are not very different and, in addition, there are no important spelling changes, as the transcription to a generic spelling does not influence the PLD distance. It can be stated that medieval Spanish and the Spanish of the Golden Age have not diverged too much.

On the contrary, it has been discovered that the distance between the medieval period and the second half of the twentieth century taking into account the original orthography is relevant ($PLD = 8.02$). If

we normalize this orthography between all periods, this distance falls to 6.31. By considering the results reported in **?**, the score in the second case, it is in the same range as the distance between diatopic varieties or *Ausbau* languages (e.g. Bosnian-Croatian, $PLD = 5.90$).

Looking at the PLD distance it can be stated that with original orthography, medieval Spanish and Spanish of the second half of the 20th century might be considered as different but very close languages, and with transcribed orthography, these two Spanish periods become historical varieties of the same language.

It is also worth noting that there is an important distance (7.69, and 5.57) between the first half of the 19th century (XIX-1) with regard to both the medieval period (XII-XV) and the so-called Golden Age (XVI-XVIII). If we normalize orthography in all periods, there is no such distance, since it falls to 5.93 and 4.75, respectively. Therefore, it seems that, at the end of the XVIII century, the orthographic changes of the *Real Academia Española*, already commented in Section 4, had an impact on the distance between the oldest and the more recent periods concerning the original spelling.

On the other hand, as Figure 3 shows, orthographic normalization makes the distances between the four subperiods of the 19th century and the 20th century minimal.

6 Final Discussion

The medieval and Renaissance periods are not very distant in Portuguese and Spanish in both original and transcribed orthography (over 4.5 when the texts are normalized). On the contrary, in English there is a great difference between these two periods, even though the distance decreases considerably (PLD: $11.26 \rightarrow 6.13$) when we normalize orthographies. In the case of Portuguese, the difference between the two periods decreases a little when spelling is normalized (PLD: $5.47 \rightarrow 4.51$), but in the case of Spanish there are almost no differences (PLD: $4.95 \rightarrow 4.44$). Therefore, spelling is an important distance mark in English, while it is not very important in the case of Portuguese and Spanish for these two ancient periods.

Concerning the most distant periods (XII-XV and XX-2), the distance in English is very large, giving resulting in separate languages, particularly if we consider the original orthography. However, this distance is shortened by more than half with normalization (PLD: $15.85 \rightarrow 8.81$), being equivalent to the distance between the medieval period and XX-2 in original orthography in Spanish and Portuguese. Also, it can be observed that the orthographic normalization in Portuguese and Spanish gives rise now to significant changes bringing these language periods much closer. More precisely, Portuguese goes from 7.73 to 6.09 and Spanish from 8.02 to 6.31. The same trend is observed when comparing the medieval period with the other periods of the nineteenth and twentieth century.

The importance of the orthographic changes in Portuguese and Spanish is probably due to the official reforms of mid and late eighteenth century. In the case of Portuguese, the language of "Os Lusíadas" (XVI-XVIII) is much closer to the language of the first half of the nineteenth century with the transcribed orthography than with the original one. In the case of Spanish, the same situation is found: the recent periods have similar values with both original and transcribed spelling, but the distances are smoother with the transcribed text than with the original one.

Finally, in the case of English, it is observed that from the Renaissance to the present day this language does not undergo great changes, regardless of the original spelling being considered. This long period represents one block separated from the medieval period. On the contrary, in the case of Portuguese and Spanish, although languages are more compact in their history than English, there are two distinct historical blocks. A first block that encompasses the medieval (XII-XVI) and Renaissance (XVI-XVIII) periods, and a second block that encompasses the nineteenth and twentieth centuries, both marked by the emergence of Academies of Languages of prescriptive character. In the case of Portuguese with more orthographic variations than the second one.

7 Conclusions and Future Work

7.1 Conclusions

A new diachronic language distance measure, PLD, has been defined to measure the distance between historical language periods. This measure was previously used to calculate the distance between different languages at present (**?**) and diachronic language distance applied to a language (**?**), and as far as we

know, this is the first attempt to use it to measure the distance between historical periods from a diachronic perspective for several languages: two related languages belong to the same linguistic family (European Portuguese and European Spanish) and one is more distant as it belongs to another family (European English). Thus, its application to both of them allows to quantify and compare its historical evolution as well as its main standardization changes over time.

The experiments performed let us conclude that medieval English is far distant from the rest of the historical periods of English, if the original orthography is considered. However, using a common transcribed orthography we see that the distance from the rest of the English periods decreases considerably, although not sufficiently to keep a significant distance from them. Therefore, the orthography in English is an important factor of separation between medieval and modern periods, but it is no longer a factor for change within the modern ones. Thus, it is noted that English has a soft and linear historical evolution since the Renaissance period (XVI-XVIII), similar to the one maintained in later centuries (XIX and XX) by Portuguese and Spanish.

By contrast, Spanish and Portuguese maintain a smoother and more linear evolution along all the historical periods, being the orthography an important factor of separation, especially between the periods of the 19th and 20th centuries with respect to Middle Age and Renaissance (specially in Portuguese).

Therefore, taking into account the experiments, it can be stated that historical language distance is not only related to grammatical or lexical matters since orthography also helps to distance or approximate the different periods.

In addition to all these observations, one of the main contributions of this work is the compilation of freely available diachronic corpora for three languages in closer original spelling: *Carvalho*. These corpora have been collected from different open historical corpora and texts repositories, [21].

7.2 Further work

Based on these results, we are planning to test PLD to measure inter-linguistic language distance to quantify the diachronic convergence/divergence among languages. For example, between languages that have had historical periods of convergence/divergence with other ones they are intimately related with: Spanish, Galician and Portuguese; Serbian, Bosnian and Croatian; Flamish and Dutch and Moldavian and Romanian. In order to do this, we will take into account works already done in Slavic languages that analyse the relationship between orthography and distance between languages such as **?**, **?**, and **?**.

In addition our aim is to apply PLD to measure the synchronic and diachronic distance between diatopic varieties of languages such as Portuguese and Spanish (e.g., testing if the distance between Mexican Spanish and European Spanish is increasing or decreasing?)

Besides we would also like to investigate the relationship between the language distance using PLD and Quality estimation (**?**).

Moreover, we aim at using PLD with different language models: e.g. n-grams calculated from relevant linguistic words, more complex phonological rules modifying the spelling, word embeddings, etc.

Finally we will test our diachronic corpora *Carvalho* [22] with other divergence measures, namely Kullback–Leibler divergence (KLD). For this we will take into account the work of **?** which studies how to validate corpora for analysis of cultural and linguistic evolution, the research performed by **?** where KLD is applied to Google Books Corpus to compare historically the change in the frequency distribution of words within one language and across languages. Also, **?** measure the diachronic change at the lexical and grammatical level in scientific writing and **?** apply KLD to investigate how ideas evolve in Parliamentary transcripts of the French Revolution Corpus. Finally, KLD has been also used to measure the divergence between different social groups (old and young people, people with and without university studies, etc) in relation to the language used (**?**).

[21] https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho
[22] https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho

Acknowledgments

The authors thanks the referees for thoughtful comments and helpful suggestions. We are very grateful to Marcos Garcia of the University of A Coruña for his contributions to the development of the experiments. Special acknowledgment are due to José António Souto Cabo and Carlos Quiroga of the University of 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 in linguistics, and Alfonso Barata Villapol for his bibliographical contributions on the history of the English language and proofreading support. This work has been partially supported by the DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE). It also has received financial support from the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

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

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.

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.

Barron, Alexander TJ, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences* 115(18):4607–4612.

Baugh, Albert C and Thomas Cable. 1993. *A history of the English language*. Routledge.

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

Bochkarev, Vladimir, Valery Solovyev, and Sören Wichmann. 2014. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface* 11(101):20140841.

Borin, Lars. 2013. The why and how of measuring linguistic differences. *Approaches to measuring linguistic differences, Berlin, Mouton de Gruyter* pages 3–25.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupilla. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals* 61(4).

Capelo, Rui Grilo, A Monteiro, J Nunes, A Rodrigues, L Torgal, and F Vitorino. 1994. *História de Portugal em datas*. Círculo de Leitores, Lisboa.

Cavnar, William B, John M Trenkle, et al. 1994. N-gram-based text categorization. *Ann Arbor MI* 48113(2):161–175.

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

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.

Del Valle, José. 2013. *A political history of Spanish: The making of a language*. Cambridge University Press.

Dunning, Ted. 1994. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University.

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 44th annual meeting of the association for computational linguistics*, pages 273–280.

Galves, Charlotte and Pablo Faria. 2010. Tycho Brahe parsed corpus of historical Portuguese. *URL: http://www. tycho. iel. unicamp. br/˜ tycho/corpus/en/index. html* .

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. 2017a. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications* 484:152–162.

Gamallo, Pablo, Jose Ramom Pichel, Santiago de Compostela, and Inaki Alegria. 2017b. A perplexity-based method for similar languages discrimination. *VarDial 2017* page 109.

Gamallo, Pablo, Susana Sotelo, and José Ramom Pichel. 2014. Comparing ranking-based and naive bayes approaches to language detection on tweets. In *Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014*. Girona, Spain.

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.

Gooden, Philip. 2009. *The story of English: How the English language conquered the world*. Quercus Books.

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.

Jágrová, Klára, Tania Avgustinova, Irina Stenger, and Andrea Fischer. 2019. Language models, surprisal and fantasy in slavic intercomprehension. *Computer Speech & Language* 53:242–275.

Jágrová, Klára, Irina Stenger, Roland Marti, and Tania Avgustinova. 2016. Lexical and orthographic distances between bulgarian, czech, polish, and russian: A comparative analysis of the most frequent nouns. In *Language Use and Linguistic Structure: Proceedings of the Olomouc Linguistics Colloquium*, pages 401–416.

Jurić, Dragana. 2013. *The historical development of the English spelling system*. Ph.D. thesis, Josip Juraj Strossmayer University of Osijek. Faculty of Humanities and Social Sciences.

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.

Kroon, Martin, Masha Medvedeva, and Barbara Plank. 2018. When simple n-gram models outperform syntactic approaches: Discriminating between dutch and flemish. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 244–253.

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 .

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.

Malmasi, Shervin, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL Shared Task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, pages 1–14. Osaka, Japan.

Mastin, Luke. 2011. The history of english.

Mattoso, José and Rui Ramos. 1994. *História de portugal*. Editorial Estampa.

Millar, Robert McColl and Larry Trask. 2015. *Trask's historical linguistics*. Routledge.

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. 1997a. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pages 11–18.

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

Pechenick, Eitan Adam, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one* 10(10):e0137041.

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

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.

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

Rama, Taraka and Anil Kumar Singh. 2009. From bag of languages to family trees from noisy corpus. In *Proceedings of the International Conference RANLP-2009*, pages 355–359.

Rissanen, Matti, Merja Kytö, and Minna Palander-Collin. 1993. *Early English in the computer age: Explorations through the Helsinki Corpus*. No. 11. Walter de Gruyter.

Sanders, Andrew. 1994. *The short oxford history of english literature.*. Oxford: Clarendon Press.

Saraiva, António José. 2001. *História da literatura portuguesa*. Porto: Porto Editora, 2001.

Saraiva, José Hermano. 1978. *História concisa de Portugal*. Publ. Europa-América.

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.

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.

Smith, Jeremy. 2003. *An historical study of English: Function, form and change*. Routledge.

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

Stenger, Irina, Klára Jágrová, Andrea Fischer, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2017. Modeling the impact of orthographic coding on czech–polish and bulgarian–russian reading intercomprehension. *Nordic Journal of Linguistics* 40(2):175–199.

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 .

Th. Gries, Stefan and Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora* 3(1):59–81.

Wieling, Martijn and John Nerbonne. 2015. Advances in dialectometry .

Williams, Edwin Bucher. 1962. *From Latin to Portuguese: Historical Phonology and Morphology of the Portugese Language*. Univ. Pennsylvania Press.

Xavier, Maria Francisca, Maria Teresa Brocardo, and MG Vincente. 1994. Cipm–um corpus informatizado do português medieval. *Actas do X Encontro da Associação Portuguesa de Linguística* 2:599–612.

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. 2017. Compiling and processing historical and contemporary portuguese corpora. *arXiv preprint arXiv:1710.00803* .

Zampieri, Marcos, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN*, vol. 2, pages 580–587.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging: The second vardial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17.

Zubiaga, Arkaitz, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2015. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation* pages 1–38.

Álvaro Iriarte, Pablo Gamallo, and Alberto Simões. 2018. Estratégias lexicométricas para detetar especificidades textuais. *Linguamática* 10(1):19–26.