Measuring language distance among historical varieties using perplexity. Application to European Portuguese.

Jose Ramom Pichel
imaxin|software,
Santiago de Compostela,
Galiza
jramompichel@imaxin.com

Pablo Gamallo
CiTIUS
Univ. of Santiago
de Compostela. Galiza
pablo.gamallo@usc.es

Iñaki Alegria
IXA group
Univ. of the Basque Country
UPV/EHU
i.alegria@ehu.eus

Abstract

The objective of this work is to quantify, with a simple and robust measure, the distance between historical varieties of a language. The measure will be inferred from text corpora corresponding to historical periods. Different approaches have been proposed for similar aims: Language Identification, Phylogenetics, Historical Linguistics or Dialectology. In our approach, we used a perplexity-based measure to calculate language distance between all the historical periods of that language: European Portuguese. Perplexity has already proven to be a robust metric to calculate distance between languages. However, this measure has not been tested yet to identify diachronic periods within the historical evolution of a specific language. For this purpose, a historical Portuguese corpus has been constructed from different open sources containing texts with spelling close to the original one. The results of our experiments show that Portuguese keeps an important degree of homogeneity over time. We anticipate this metric to be a starting point to be applied to other languages.

1 Introduction

In this article, we deal with the concept of diachronic language distance, which refers to how different one historical period of a language is from another. The prevailing view is that language distance between two languages cannot be measured appropriately by using a well-established score because they may differ in many complex linguistic aspects such as phonetics and phonology, lexicography, morphology, syntax, semantics, pragmatics, and so on. In addition, languages change internally as well as in relation to other languages throughout their history (Millar and Trask, 2015).

Quantifying all these aspects by reducing them to a single distance score between languages or between historical periods of a 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. Also, there is not any standard methodology to define a metric for language distance, even though there have been different attempts to obtain language distance measures, namely in phylogenetic studies within historical linguistics (Petroni and Serva, 2010), in dialectology (Nerbonne and Heeringa, 1997), in language identification (Malmasi et al., 2016), or in studies about learning additional languages within the field of second language acquisition (Chiswick and Miller, 2004).

In the present work, we consider that the concept of language distance is closely related to the process of language identification. Actually, the more difficult the identification of differences between two languages or language varieties is, the shorter the distance between them. 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 (Malmasi et al., 2016) . Thus, character n-grams not only encode lexical and

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 145–155
Santa Fe, New Mexico, USA, August 20, 2018.

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, introducing prepositional phrases.

In our previous work, perplexity-based measures were used for language identification (Gamallo et al., 2016) and for measuring the distance between languages (Gamallo et al., 2017a). Now, the main objective of our current work is to extend this approach in order to measure distance between periods of the same language (diachronic language distance), also based on perplexity. This method has been applied to a case of study on European Portuguese from 12th to 20th century. Two experiments are reported: the first one uses our "perplexity-based" method in a historical corpus of Portuguese with an orthography closely related to that of the original texts, and the second experiment was applied using a transliterated corpus trying to use the same orthography for the whole corpus. The article is organized as follows: First, we will introduce some studies on language distance (Sec. 2). Then, our language distance measure is described in Section 3. In Section 4, we introduce the experimental method and finally, in Section 5, we describe the two above mentioned experiments and discuss the results. Conclusions are addressed in Section 6.

2 Related Work

Linguistic 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 (Wieling and Nerbonne, 2015). According to Borin (2013), genetic linguistics (also known as "phylogenetics" or "comparative-historical linguistics") and dialectology are the most popular fields dealing with language distance. This author stated: (Borin, 2013, p. 7) "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 methods to be useful in both cases" (Borin, 2013, p. 7).

Degaetano-Ortlieb et al. (2016) present an information-theoretic approach, based on entropy, to investigate diachronic change in scientific English.

In the following sections, we introduce some relevant work on phylogenetics and dialectology, but also on corpus-based approaches.

2.1 Phylogenetics

The objective of linguistic phylogenetics, a sub-field of historical and comparative linguistics, is to build 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 (Nakhleh et al., 2005; Holman et al., 2008; Bakker et al., 2009; Petroni and Serva, 2010; Barbançon et al., 2013). More precisely, lexicostatistics is based on cross-lingual word lists (e.g. Swadesh list (Swadesh, 1952) or ASJP database (Brown et al., 2008)) to automatically compute distances using the percentage of shared cognates. Levenshtein distance among words (Yujian and Bo, 2007) in a cross-lingual list is one the most common metrics used in this field (Petroni and Serva, 2010). Ellison and Kirby (2006) present a method, called PHILOLOGICON, for building language taxonomies comparing lexical forms. The method only compares words language-internally and never cross-linguistically.

Rama and Singh (2009) test four techniques for constructing 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.

2.2 Dialectology

As in phylogenetics, Levenshtein distance among list of words is employed very often in dialectology (Nerbonne and Hinrichs, 2006; Nerbonne et al., 1999).

In addition to raw Levenshtein distance, (Nerbonne and Hinrichs, 2006) proceed to measuring pronunciation differences, focusing on differences in the pronunciation of the same words in different varieties. Results are validated using measurements based on the degree to which they correlate with dialect speakers' judgments about those differences. Also, Heeringa et al. (2006) evaluated several string distance algorithms for dialectology, but always based on pairs of words.

2.3 Corpus-Based Approaches

To measure language distances, very recent approaches construct complex language models not from word lists, but from large cross-lingual and parallel corpora. 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 (Liu and Cong, 2013; Gao et al., 2014; Asgari and Mofrad, 2016).

It is worth noting that most techniques in language identification also use corpus-based approaches, mainly based on n-gram language models. Language identification is considered as being a pretty solved task (McNamee, 2005), specially for languages by distance, also called *Ausbau* languages (Kloss, 1967). However, there are already big challenges to classify some closely related varieties of the same language (e.g. Nicaraguan Spanish and Salvadoran Spanish) or *Abstand* languages (Kloss, 1967) (e.g. Czech and Slovak). Two specific tasks of language identification have attracted a lot of research attention in recent years, namely discriminating among closely related languages (Malmasi et al., 2016) and language detection on noisy short texts such as tweets (Gamallo et al., 2014; Zubiaga et al., 2015). Reasonable results have been achieved even for very closely related varieties using corpus-based strategies. For instance, Zampieri et al. (2013) 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 a extremely high accuracy of 0.998 for distinguishing between European Portuguese and Brazilian Portuguese, and 0.990 for Mexican and Argentinian Spanish.

2.4 Historical Portuguese

Historical periods of the Portuguese language are reported in several language monographies: *História da Literatura Portuguesa* (History of Portuguese Literature) (Saraiva, 2001) and *História da Língua Portuguesa* (Portuguese Language History) (Teyssier, 1982), Historical Phonology and Morphology of the Portuguese Language (Williams, 1962), as well as in different books of History of Portugal: *História de Portugal em datas* (History of Portugal in a timeline) (Capelo et al., 1994), *História de Portugal* (History of Portugal) (Mattoso and Ramos, 1994) and *História concisa de Portugal* (Brief history of Portugal) (Saraiva, 1978).

3 Perplexity

Perplexity 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 (González, 2015), classification of related languages (Gamallo et al., 2016) and measuring distances among languages (Gamallo et al., 2017a).

3.1 Perplexity of a language model

Perplexity is frequently used as a quality measure for language models built with n-grams extracted from text corpora (Chen and Goodman, 1996; Sennrich, 2012). This is a metric about how well a language model is able to fit a text sample. A low perplexity indicates the language model is good at predicting the sample. On the contrary, a high perplexity shows the language model is not good to predict the given sample. It turns out that we could use perplexity to compare the quality of language models in relation to specific textual tests.

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

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

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

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

Equation 2 estimates the n-gram probability by dividing the observed frequency (C) of a particular sequence of characters by the observed frequency of the prefix, where the prefix stands for the same sequence without the last character. To take into account unseen n-grams, we use a smoothing technique based on linear interpolation.

3.2 Perplexity-Based Language Distance (PLD)

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

4 Methodology

Our methodology is based on applying PLD measure to a historical corpus of a language (also called "diachronic corpus"), in order to obtain a diachronic language distance between periods. A representative and balanced historical corpus is required. This corpus is divided into two parts: train and test corpora. Also, train and test must be divided into different language periods, which should be previously defined according to philological criteria. Finally, 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.

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

1. First, we need to define historical periods of a language. For this purpose, it will be necessary to take into account philological studies on the specific language at stake. For Portuguese,

the periods were defined according to the ideas reported in two pieces of work about, on the one hand, the History of Portuguese Language (Teyssier, 1982) and, on the other, about Historical Phonology and Morphology of the Portuguese Language (Williams, 1962). As a result of this philological research, Portuguese language may be divided into a medieval period (XII-XVth centuries), a renaissance period (XVI-XVIIth), XVIIIth, first half XIXth, second half XIXth, first half XXth, and second half XXth century. Yet, considering the lack of documents for some of these periods, we had to merge renaissance and XVIIIth into one single period. Thus, we have selected the following 6 periods: XII-XV, XVI-XVIII, XIX-1, XIX-2, XX-1, and XX-2.

2. In the second step, we select a representative and balanced historical corpus. For this purpose, texts from several genres must be retrieved. For our corpus, we collected texts from both nonfiction and literature. In addition, we consider that it is important to get documents with a spelling as close as possible to the original one. It is quite relevant to bear in mind that the oldest period (medieval) is where there are more differences between texts, since language was not standarized at that time. Unlike other historical Portuguese corpora (Galves and Faria, 2010), in the construction of the corpus we have paid special attention to maintain the original spelling for every text. Bearing this aim in mind, adapted or edited versions have been ruled out (for example, in the 19th century, the spelling "ph" was used for the phoneme */f/*, and in many available digital versions the texts are adapted to modern spelling by replacing "ph" with "f", but we discarded these versions).

3. Then, text corpus is divided into both train and test partitions. As soon as we get documents in their original spelling and they are classified in the pre-defined historical periods, we must decide if these documents must belong to either the train or the test corpus, each one also divided in the same 6 periods. The size of each period of the test corpus is about 20% of the size of the corresponding period in the train corpus.

4. Finally, PLD is applied to the previously organized train/test dataset and results are evaluated. The results obtained by using PLD between periods are compared with those obtained between well-established languages and reported in Gamallo et al. (2017a), where the distance among more than 40 languages was analyzed. Considering that two historical periods belong to the same language, for Portuguese the PLD score between two periods should not be greater than the perplexity between two recognized languages. Therefore, given that the perplexity-based distance between Catalan and Spanish is about 8, the distance between two Portuguese periods should be lower than that value; otherwise we consider that there might be some problems with, at least, one aspect of our methodology: either the corpus or the measure.

5 Experiments

5.1 Corpus

As we aim to test our methodology on Portuguese, the language models were generated by making use of a collection of documents in several periods of Portuguese language. These documents are not translations of each other and are constituted by a balanced combination of genres (both literature and nonfiction) period by period. As a result, we collected comparable and balanced corpus from literature and nonfiction in six different periods of languages from different sources. Our method to compile the historical corpus was the following.

First, in order to know which were the most relevant nonfiction and literature documents in Portuguese for each historical period, we took into account information reported in historical work cited above in Sec. 2.4. As a result, we selected a set of relevant candidate documents to be part of our experiments.

Second, we searched for these candidate texts in open repositories such as *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) (Xavier et al., 1994), Project

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
Train corpus (Words)	1,509,774	1,426,636	1,327,045	1,612,320	1,325,353	1,688,787
Test corpus (Words)	305,773	310,405	296,712	334,145	293,952	363,693
Proportion (Test/Train)	20.25%	21.75%	22.35%	20.72%	22.17%	21.53%

Table 1: Number of words using in Train and Test corpus

Gutenberg, specially for the XIX century[1], Wiki source[2], OpenLibrary [3], Tycho Brahe corpus[4] (Galves and Faria, 2010), Domínio Público[5], Arquivo Pessoa[6], Linguateca[7], *Corpus de Textos antigos* (Corpus of old texts)[8] and Colonia corpus[9] (Zampieri, 2017).

It is worth noting that the further back we go in historical texts (e.g.: renaissance, medieval), the more spelling differences between texts are found due to a lack of a stable spelling standard. Also, there were high rates of illiteracy since there was not any kind of public schools to learn how to read or write the language. Actually, the first relevant language standard for Portuguese is defined and applied at the end of XVIIIth century, as it also happened in other Romance languages such as French or Spanish. *Academia das Ciências de Lisboa* (Lisbon Academy of Sciences), one of the bodies that regulate the standardization of European Portuguese language, was created in 1779 in Lisbon.

Then, we checked whether the documents selected in the previous step were in the original spelling. If so, they were indexed and their OCR errors were cleaned; otherwise they were not considered.

All texts with original spelling were digitized and cleaned. It resulted in a new diachronic corpus, we call Diachronic Portuguese Corpus (DiaPT). To compute PLD measure between all periods, each period of DiaPT (i.e. XII-XV, XVI-XVIII, XIX-1, XIX-2, XX-1, XX-2) was divided into two partitions: train and test. As a result, each training partition is constituted by about 1,3/1.5M word tokens. Balanced train-test pairs allows us to compute PLD measure without bias.

5.2 Results

The objective of the current experiments is to compare six language periods of European Portuguese language using PLD. The specific implementation of PLD consists of 7-gram models and a smoothing technique based on linear interpolation. Two experiments have been performed. The first one consists of applying PLD measure on a Portuguese historical corpus keeping the original spelling. In the second experiment, we apply the same PLD measure to the same historical documents, but previously transcribed by means of a normalization process.

5.2.1 PLD with original spelling

In this experiment, we have developed a set of scripts (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. Table 2 shows the diachronic language distance between all historical Portuguese periods with original spelling using PLD. Some representative samples of these distances are depicted in Figure 1. More precisely, Figure 1(a) compares the distance evolution across all periods of the two

[1] https://www.gutenberg.org/browse/languages/pt

[2] https://en.wikisource.org/wiki/Category:Portuguese_authors

[3] https://openlibrary.org/

[4] http://www.tycho.iel.unicamp.br/corpus/index.html

[5] http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=16090

[6] http://arquivopessoa.net/textos/

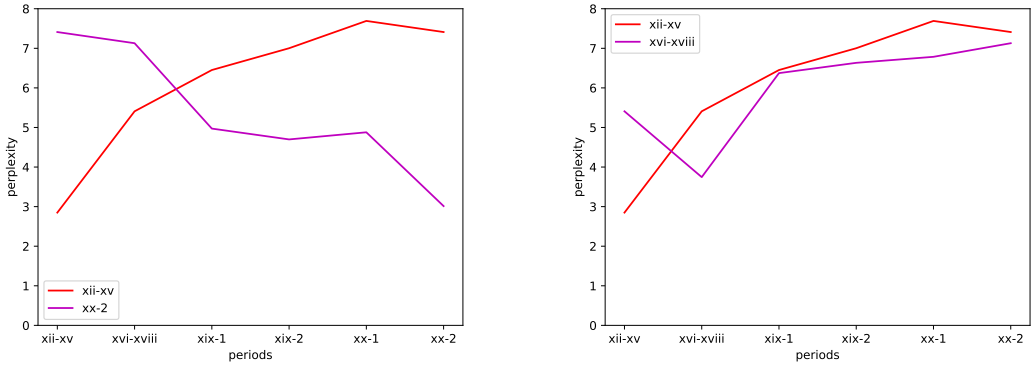
[7] https://www.linguateca.pt/

[8] http://alfclul.clul.ul.pt/teitok/cta/index.php?action=textos

[9] http://corporavm.uni-koeln.de/colonia/

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	2.849	5.408	6.451	7.002	7.692	7.411
XVI-XVIII	5.408	3.745	6.373	6.633	6.785	7.128
XIX-1	6.451	6.373	2.990	4.081	3.965	4.972
XIX-2	7.002	6.633	4.081	3.037	3.937	4.698
XX-1	7.692	6.785	3.965	3.937	2.872	4.878
XX-2	7.411	7.129	4.972	4.698	4.878	3.013

Table 2: PLD diachronic measure in original spelling (DiaPT corpus)



(a) XII-XV / XX-2

(b) XII-XV / XVI-XVIII

Figure 1: Original spelling. In (a) we compare the PLD distances of XII-XV and XX-2 across all periods. In (b) the same comparison is made between XII-XV and XVI-XVIII.

further away periods, namely medieval (XII-XV) and second half XXth period (XX-2), whereas Figure 1(b) compares two close historical periods: XII-XV and XV-XVIII.

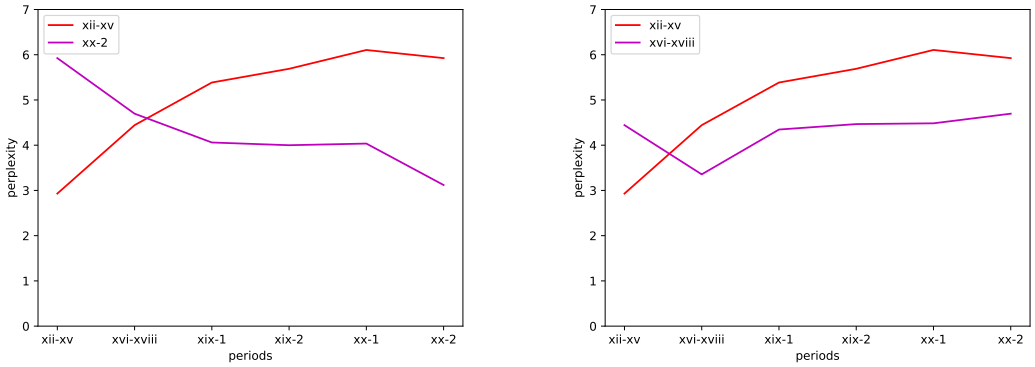
Figure 1(a) plots how XII-XVth diverges from all the periods in a regular basis: there is an almost linear growth from 4.48 for XVI-XVIII (the closest PLD distance), up to 7.69 for XX-1 (the furthest one), even though the distance grows smoothly from XIX-1 and decreases slightly in XX-2. The same pattern can be observed for XX-2, but in the reverse direction: distance grows slightly until XIX-1, but there is a more pronounced divergence with regard to the furthest periods.

On the other hand, Figure 1(b) compares XII-XVth and XVI-XVIIIth periods. The most relevant information in this plot is the following: XVI-XVIII is more distant from the modern periods (6.37 with regard to XIX-1) than from the medieval period, (5.4 with regard to to XII-XV). In addition, as it was expected, the distance grows very slowly from XIX, in the same way as XII-XV with regard to the modern periods.

In general, distance between periods is correlated with chronology.

5.2.2 PLD with transcribed spelling

In a second experiment, we have converted DiaPT corpus into a new one in which documents of all periods share a common spelling: DiaPT_norm. To do so, all Portuguese historical periods were both transliterated into Latin script and normalized using a generic orthography closer to phonological issues. The encoding of the final spelling normalization consists of 34 symbols, representing 10 vowels and 24 consonants, designed to cover most of the commonly occurring sounds, including several consonant palatalizations and a variety of vowel articulation. As the encoding is close to a phonological one, the new spelling might be seen as a pointer to phonology. After this transformation we have carried out the same experiment as for DiaPT (described in the previous subsection).



(a) XII-XV / XX-2 (b) XII-XV / XVI-XVIII

Figure 2: Transcribed spelling. In (a) we compare the PLD distances of XII-XV and XX-2 across all periods. In (b) the same comparison is made between XII-XV and XVI-XVIII.

	XII-XV	XVI-XVIII	XIX-1	XIX-2	XX-1	XX-2
XII-XV	2.937	4.443	5.386	5.689	6.106	5.925
XVI-XVIII	4.443	3.355	4.346	4.467	4.484	4.697
XIX-1	5.386	4.346	3.118	3.676	3.620	4.060
XIX-2	5.689	4.467	3.676	3.137	3.569	4.000
XX-1	6.106	4.484	3.620	3.569	2.997	4.036
XX-2	5.925	4.697	4.060	4.000	4.036	3.120

Table 3: PLD diachronic measure in a common transcribed spelling (DiaPT_norm corpus.)

In this new experiment on DiaPT_norm, the PLD distances shown in Table 3 are very similar to those of the previous experiment (Tab 2). The pattern of distances is the same in both experiments, even though in DiaPT_norm there is a closer approximation between periods since there is lower divergence in general as a result of using normalized orthography.

5.3 Discussion

The results obtained in our experiments allow us to conclude that there are only three clearly separated historical periods of Portuguese: XII-XV, XVI-XVIII and XIX-XX. If we look in depth our results, we can observe that the distance between the modern periods (from XIX to XX) could be too low to justify the existence of different periods in terms of language variation.

The results also lead us to observe that European Portuguese language is historically a compact language. There is not a large divergence within the different historical periods of European Portuguese language. The longest difference between XII-XV and XX-2 is over 6.19, which drops to 5.92 with a normalized orthography for all periods. By considering the results reported in (Gamallo et al., 2017b), this score is in the same range as the distance between diatopic varieties or *Ausbau* languages (e.g. Bosnian-Croatian, perplexity = 5.90), and is not larger than the distance between languages considered undoubtedly different but closely related (e.g. Spanish-Portuguese, perplexity=7.74).

6 Conclusions and Future Work

6.1 Conclusions

We have defined a new diachronic language distance measure, PLD, to identify the main evolution phases of a language and measure how much these phases differ from one another. Even though a similar measure was used to compute language distance in our previous work (Gamallo et al.,

2017b), as far as we know, this is the first attempt to use it for measuring distance between periods in a diachronic perspective. Its application to Portuguese language allows us to quantify its historical evolution as well as its main standarization changes over time.

Three main periods of Portuguese have been identified, and the distance between ancient periods and the modern ones is not bigger than the distance between language varieties from a diatopic perspective. So, Portuguese keeps an important degree of homogeneity over time.

Another contribution of our work is that a new diachronic Portuguese corpus in original spelling has been created: DiaPT. This corpus has been collected from different open historical corpora and texts repositories, priorizing those who have original spelling [10].

PLD is a robust measure since the transcription of the corpus with a shared ortography has not had any impact in changing the distance of Portuguese periods. On the contrary, this change has compacted the internal distance between language periods, but has not generated different relations between them.

6.2 Further work

Based on these results, we are planning to test diachronic distance on another languages and linguistic varieties. Also, we aim at using PLD with different language models: e.g. n-grams calculated from relevant linguistic words, phonological rules modifying the spelling, etc. Additionally we would like to test this technique for labeling undated texts. Finally, we will use PLD to enhance precision on other NLP tools, such as language identification, specially for *Ausbau* languages and closely related varieties.

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 is due José António Souto Cabo of the University of Santiago de Compostela for his expertise in medieval historical linguistics of Galician-Portuguese. This work has been partially supported by a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, by TelePares (MINECO, ref:FFI2014-51978-C2-1-R) and TADeep (MINECO, ref: TIN2015-70214-P) projects. 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

Ehsaneddin Asgari 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.

Dik Bakker, 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.

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

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

Cecil H. Brown, 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).

[10] https://github.com/gamallo/Perplexity-tree/master/resources/DiaPT

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

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

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

Stefania Degaetano-Ortlieb, 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)*.

T Mark Ellison 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.

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

Pablo Gamallo, 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.

Pablo Gamallo, 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.

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

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

Yuyang Gao, 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.

Meritxell González. 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.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances*, pages 51–62. Association for Computational Linguistics.

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

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

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

Shervin Malmasi, 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.

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

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 3:94–101.

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

Luay Nakhleh, 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.

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

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

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison*, 15.

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

Taraka Rama 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.

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

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

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

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

Paul Teyssier. 1982. *História da língua portuguesa*. Livraria Sá da Costa Editora.

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

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

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

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

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

Marcos Zampieri. 2017. Compiling and processing historical and contemporary Portuguese corpora. *arXiv preprint arXiv:1710.00803*.

Arkaitz Zubiaga, 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.