Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 77-84 recibido 22-03-2019 revisado 05-05-2019 aceptado 15-05-2019

Cross-lingual Diachronic Distance: Application to Portuguese and Spanish

Distancia diacrónica interlingüística: aplicación al portugués y el castellano

José Ramom Pichel Campos[1], Pablo Gamallo Otero[2], Iñaki Alegria Loinaz[3]
[1]imaxin|software
[2]CITIUS-Universidade de Santiago de Compostela
[3]IXA Taldea-UPV/EHU
[1]jramompichel@imaxin.com
[2]pablo.gamallo@usc.es
[3]i.alegria@ehu.eus

Abstract: The aim of this paper is to establish a corpus-based methodology for automatically measuring the cross-lingual distance between historical periods of two languages using *perplexity*. The corpus of both has been constructed adhoc with the closest spelling to the original representing chronologically and in a balanced way fiction and non-fiction. The methodology has been applied to two related languages, Portuguese and Spanish, and measured their diachronic distances both in original orthography and in an automatically transcribed spelling.
Keywords: Corpus linguistics, Historical Linguistics, Language distance, Development of linguistic resources and tools

Resumen: El objetivo de este trabajo es establecer una metodología basada en corpus para medir automáticamente la distancia interlingüística entre períodos históricos de dos lenguas mediante *perplexity*. El corpus de los dos idiomas ha sido construído adhoc con ortografía lo más próxima a la original representando cronológicamente y de forma balanceada ficción y no ficción. Se ha aplicado la metodología a dos lenguas relacionadas, Portugués y Español, y medido sus distancias diacrónicas tanto en ortografía original como en una ortografía transcrita automáticamente.
Palabras clave: Lingüística de Corpus, Lingüística Histórica, Distancia entre Lenguas, Desarrollo de recursos lingüísticos y herramientas

1 Introduction

Languages are constantly changing throughout their history (Millar and Trask, 2015) in such a way that it is as challenging to measure the diachronic distance between periods of the same language as it is to measure the cross-lingual distance between related languages. It is also a challenge to reduce this automatic distance to a single metric to validate the hypotheses of language historians.

There have been different approaches 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), and in the field of second language acquisition (Chiswick and Miller, 2004). However, to the best of our knowledge, there is no work on how to measure cross-lingual diachronic distance of two different languages. This article proposes a corpus-driven methodology for automatically measuring a cross-lingual diachronic distance between two languages from a historical corpus.

For this general purpose, 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. The best language identification systems are based on n-gram models

ISSN 1135-5948. DOI 10.26342/2019-63-8

of characters extracted from textual corpora (Malmasi et al., 2016). 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.

The specific objective of the present article is to apply this perplexity-based measure to study and compare the cross-lingual diachronic distance among historical periods of two close-related languages: European Portuguese and European Spanish, from 12th to 20th century. To achieve this goal, we have carried out two different experiments: one applying the methodology of cross-lingual diachronic distance calculation based on perplexity to historical corpus whose texts are written with a spelling very close to the original source; and another applying the same method to the same corpus but automatically transcribed to a common orthography that approximates the two compared languages.

The results show that the two languages are not separated from the Middle Ages in a linear way, but that approximations and divergences occur along the time axis.

Finally, an additional objective of the article is to verify whether the proposed cross-lingual diachronic distance fits the opinion and analysis of philological experts.

The article is organized as follows. Some related work is introduced in Section 2. Then, the method and the corpus are described in sections 3 and 4, respectively. Section 5 introduces the experiments along with a discussion on the results. Finally, conclusions and future work are addressed in Section 6.

2 Related work

Language distance has been defined from different perspectives using different methods. We will explore two different approaches: phylogenetics and corpus based strategies.

2.1 Linguistic Phylogenetics

The objective of linguistic phylogenetics, a sub-field of historical and comparative linguistics, is to classify the languages by building a rooted tree that describes 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, Ringe, and Warnow, 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), in order to automatically measure 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 et al., (2006), present a method to build language taxonomies comparing lexical forms. The method only compares words language-internally and never cross-linguistically. Finally, Satterthwaite (2011) and Rama and Singh (2009) test four techniques to construct phylogenetic trees from corpora: cross-entropy, cognate coverage distance, phonetic distance of cognates and feature N-grams. They conclude that these measures can be very useful for languages which do not have linguistically hand-crafted lists. Finally, using perplexity-based distance, Gamallo et al., (2017), built a network that represents the current map of similarities and divergences among the main languages of Europe.

2.2 Language distance

To measure language distances, complex language models have been built from large cross-lingual and parallel corpora to obtain metrics to measure language distances. 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).

Degaetano-Ortlieb et al., (2016) present an information-theoretic approach based on entropy to investigate diachronic change in scientific English. Rama et al., (2015) use cross-entropy to measure distances, while Singh (2007) uses phonetic distances. These studies can be seen as the most related to our work, which is corpus-driven and has been previously applied to the diachronic varieties of the same language (Pichel, Gamallo, and

Alegria, 2018).

3 Methodology

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 (Chen and Goodman, 1996; Sennrich, 2012). It has also been used in very specific tasks, such as to classify formal and colloquial tweets (González, 2015), and to identify close-related languages (Gamallo et al., 2016). In Gamallo et al., (2017), a specific perplexity-based distance, called PLD, has been defined and applied to compute the distance of different European languages. In a previous work (Pichel, Gamallo, and Alegria, 2018), we applied PLD to measure the diachronic distance between different historical periods of the same language. In the current work, our aim is to apply PLD to measure cross-lingual diachronic distance between two different languages in the same historical periods. In order to be able to compare the perplexity distances we have obtained with those reported in Gamallo et al., (2017), we use the same PLD configuration: namely, 7-gram language models, smoothing technique based on linear interpolation, and train/test corpora with 1,25M/250K words, respectively.

3.2 Task Description

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

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

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

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

4. Once the textual sources of our corpus have been selected and the periods have been established, two subcorpora are created for each period: train and test. In the train partition, we include for each period texts in original spelling in fiction and non-fiction. In order to facilitate a better representation of the language for each period, the fiction and non-fiction texts in both the train and the test were balanced at approximately 50% (the test and train texts are distinct sets). It is worth mentioning that the train and test partitions are not manually annotated as our method is fully unsupervised.

5. A spelling normalization is applied to all the texts and a transcribed version is obtained for each corpus. The common alphabet consists of 34 symbols, representing 10 vowels (including accents) and 24 consonants, designed to cover most of the commonly occurring sounds, including several consonant palatalizations and a variety of vowel articulation. The encoding is thus close to a phonological one and, then, makes it possible to simplify and homogenize cases in which similar sounds (generally palatalizations) are transcribed differently in different languages. For instance, the palatalized nasal sound is transcribed by our normalizer as "ny", thus unifying the Portuguese spelling "nh" and the Spanish "ñ". Similarly, the palatalized lateral is transcribed as "ly", simplifying the

OS	TS	Edited
Com seu meneio hipócrita, calando. Na alma lodosa da blasfémia o grito. Então exultarão os bons, e o ímpio, (...)	com seu meneio hipocrita calando na alma lodosa da blasfemia o grito então exultarão os bons e o impio (...)	Com seu meneio hipócrita, calando. Na alma lodosa da blasfémia o grito. Então exultarão os bons, e o ímpio, (...)

Table 1: Portuguese excerpt in three versions: original spelling (OS), transcribed (TS), and edited text.

two different spellings "lh" in Portuguese and "ll" in Spanish.

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

In order to allow researches to apply the methodology to any language, we have developed a pipeline architecture in Perl, which is freely available[1]. With this implementation, we have built train partitions giving rise to six different 7-gram diachronic language models per language. Then, we have analyzed all test documents so as to generate six 7-gram files per language.

4 Corpus

The Corpus that we have built and used in our experiments, called *Carvalho*, is freely available and contains the diachronic corpus for the two languages: Carvalho-PT-PT (European Portuguese) and Carvalho-ES-ES (European Castillan, also known as Spanish of Spain).

Our initial aim was to classify the corpus for both languages into historical periods with three fundamental stages: medieval period (XII-XV), modern age (XVI-XVIII), and contemporary age (XIX-XX), following the classification provided by Corpus Helsinki (Rissanen and others, 1993).

However, as Portuguese and Spanish have a large volume of texts and different orthographic standards in the 19th and 20th centuries, we have decided to divide these two centuries into two subperiods (XIX-1, XIX-2, XX-1 and XX-2).

Regarding the different orthographic standards in Portuguese, there was a first orthographic standard in 1779 promoted by the *Academia das Ciências de Lisboa*, which was later reformed in the years: 1885, 1911, 1945, 1973 and 1990. In the case of Spanish, the orthographic standard of 1741 promoted by the *Real Academia Española* was consolidated in the two successive centuries.

We have chosen to use documents with a spelling as close as possible to the original text. This decision makes it possible to compute the cross-lingual diachronic distance between texts in both original and transcribed spelling. Table 1 shows three excerpts of the same text, belonging to the book *A Harpa do crente* by Alexandre Herculano (1810-1877). On the left, we show the original spelling (OS) of the document we have selected to be part of our corpus. In the middle, the same text has been transcribed to a common spelling (TS), including lower-case transformation. On the right, we show an edited version adapted to the current Portuguese. Only OS and TS versions have been selected. No edited version has been introduced in our corpus.

To create the Portuguese Carvalho-PT-PT corpus, we identified and selected documents from the following repositories: Tycho Brahe corpus[2] (Galves and Faria, 2010), Colonia[3] (Zampieri, 2017), *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) (Xavier, Brocardo, and Vincente, 1994), Project Gutenberg, specially for the XIX century[4], Wiki source[5], OpenLibrary[6], Arquivo Pessoa[7],

[1] https://github.com/gamallo/Perplexity

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

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

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

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

[6] https://openlibrary.org/

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

Carvalho PT/ES	Train-pt	Test-pt	Train-es	Test-es
XII-XV	1.509M	305K	1.317M	314k
XVI-XVIII	1.449M	289K	1.302M	314K
XIX-1	1.262M	253K	1.368M	311K
XIX-2	1.464M	312K	1.315M	257K
XX-1	1.325M	336K	1.252M	253K
XX-2	1.688M	363K	1.231M	250K

Table 2: Size of Train and Test corpora in six historical periods of Portuguese and Spanish

Linguateca[8], *Corpus de Textos antigos* (Corpus of old texts)[9], *Domínio Público*[10]

Concerning Spanish, Carvalho-ES-ES was built from the following repositories: Project Gutenberg, specially for the XIX century[11], OpenLibrary[12], Wiki source[13].

Finally, the two corpora were partitioned into train and test parts so as to compute the perplexity-based measure (PLD). Table 2 shows the size of both Train and Test corpora across the 6 periods of each language.

5 *Experiments*

The experiments we have carried out consist of measuring the cross-lingual diachronic distance between the different historical periods of Portuguese and Spanish. First, we applied the PLD distance to Carvalho-PT-PT / Carvalho ES-ES in original spelling (OS). Then, PLD was applied to the same corpus but transcribed into a common spelling (TS).

5.1 Results

Table 3 shows the results of applying PLD to OS and TS versions of the Portuguese and Spanish corpora period by period. More precisely, we compared each period cross-lingually: for instance, the PLD distance between the Spanish and Portuguese Medieval periods (XII-XV) in OS is 11,49, but in TS is, as expected, lower: 8,9. And we did the same with the rest of the periods. Figure 1 depicts the same information in a plot so as to better observe how the two languages behave in relation to each other throughout history.

Periods	PLD (OS)	PLD (TS)
XII-XV	11.48	8.9
XVI-XVIII	12.12	8.59
XIX-1	11.54	8.72
XIX-2	9.78	7.49
XX-1	13.20	9.34
XX-2	11.99	9.04

Table 3: Cross-lingual diachronic distance (PLD) between Spanish and Portuguese across six historical periods in original spelling (OS) and transcribed (OS).

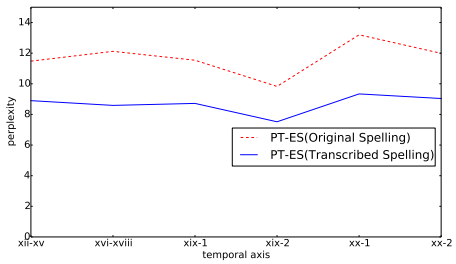

Figure 1: Cross-lingual diachronic distance between Spanish and Portuguese through time axis in OS and TS.

5.2 Discussion

The maximum PLD distance in OS is 13.2, which was reached in the first half of the 20th century (XX-1), while the minimum PLD distance is 9.83, obtained in the second half of the 19th century (XIX-2). In TS, the maximum distance is 9.34 in XX-1, while the smallest one is 7.52 in XIX-2. According to the results reported in Gamallo et al., (2017), the PLD scores of close-related languages of the same family range from 7 (e.g., Croatian

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

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

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

[11] https://www.gutenberg.org/browse/languages/es

[12] https://openlibrary.org/

[13] https://en.wikisource.org/wiki/Category:Spanish_authors

and Bosnian) to 9 (e.g., Czech and Slovak). Those values were obtained from transcribed spelling (TS). Therefore, the distance between all the historical periods of Portuguese and Spanish is always framed in a typical distance of very close languages if they were using a common transcribed spelling.

Another important finding is the following. In all historical periods, the rate of decrease in the distance between the OS and TS varies between 3.86 in XX-1 and 2.31 in XIX-1. This significant drop in PLD seems to suggest that spelling is an important factor in making the difference between the two languages. With a common orthography, Portuguese and Spanish have a very small distance, similar to that of two variants of the same language. By contrast, with two well-differentiated orthographies (as they currently have), the distance widens to more than 13 PLD and resembles that of two clearly different (even if closely related) languages, such as Spanish and Catalan, which have a PLD distance of 14 according to Gamallo et al., (2017).

Yet, The most important observation that can be extracted from the results is the following. The two languages do not separate linearly along the time axis, as might be expected from two languages that start from the same root tongue and standardize independently. On the contrary, their evolution takes place with convergences and divergences not necessarily related to the chronological order. In the first half of the 19th century (XIX-1), both languages diverge with a similar distance to the medieval distance (XII-XV), whereas in the second half of the 19th century (XIX-2) is when their distance converge the most. Later, in the following period (XX-1), their distance increases again reaching the maximum distance but immediately decreases until it reaches values in XX-2 close to those of the Middle Ages.

There may be socio-political motives explaining the consecutive approaches/separations between the two languages. The rapprochement in the second post-Renaissance period (XVI-XVIII) could be explained for the political and cultural hegemony that Castile had in that period that influenced the Portuguese elites, in addition to Portugal's political dependence during the seventeenth century which also influenced cultural and supposedly linguistic issues. Because of this, Spanish words were taken in with ease, as if they were not truly foreign words, but family words (Venâncio, 2014). Also, the promoters of vernacular Portuguese in the Modern Age accentuated and made symbolic use of the difference against the competing language (Spanish). And orthography, above all, served for such a delimiting process (Corredoira, 1998).

The following period of rapprochement between the two languages, in the second half of the 19th century (XX-2), could be due, in part, to the global effects of French and its influence on Roman languages after the Enlightenment period (Curell, 2006). The subsequent distancing between Portuguese and Spanish at the beginning of the 20th century (XX-1) would be partially explained, in addition to the new orthographic rules for Portuguese approved in those years, by the influence of Romanticism, the concept of nation-state and the linguistic *casticism* that derives from this national sentiment.

6 Conclusion and Further work

The present work consists of the automatic calculation of the cross-lingual diachronic distance from two historical corpus of different languages in original orthography. This perplexity-based measure, PLD, was previously used to calculate language distance (Gamallo, Pichel, and Alegria, 2017) and diachronic language distance between different historical periods of the same language (Pichel, Gamallo, and Alegria, 2018).

The experiments we carried out led us to conclude that orthography is an important factor in the distance between Portuguese and Spanish. We also observed that the their distance does not increase chronologically but that historical periods of divergence are followed by periods of convergence and the other way around.

In addition to all these observations, one of the main contributions of this work is the compilation of a freely available diachronic corpus for two languages in closer original spelling: Carvalho-PT-PT and Carvalho-ES-ES[14]. This corpus has been collected from different open historical corpora and texts repositories.

Based on these results, we are planning to use PLD to measure the distance between di-

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

atopic varieties such as European and Brazilian Portuguese or Latin American Spanish and European Spanish.

Acknowledgments

The authors thanks the referees for thoughtful comments and helpful suggestions. We are very grateful to Fernando Venâncio from the University of Amsterdam, José António Souto Cabo and Carlos Quiroga from the University of Santiago de Compostela for his expertise in Portuguese and Spanish Language history. This work has received financial support from the DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), and the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Asgari, E. and M. 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, D., A. Muller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, and E. W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181.

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

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

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

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

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

Corredoira, F. V. 1998. *A construção da língua portuguesa frente ao castelhano: o galego como exemplo a contrario*.

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

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

Ellison, T. M. and S. 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, C. and P. Faria. 2010. Tycho Brahe parsed corpus of historical Portuguese. *URL: http://www. tycho. iel. unicamp. br/˜ tycho/corpus/en/index. html*.

Gamallo, P., I. Alegria, J. R. Pichel, and M. 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, P., J. R. Pichel, and I. Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Gao, Y., W. Liang, Y. Shi, and Q. 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, M. 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.

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

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

Malmasi, S., M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, and J. 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.

Millar, R. M. and L. Trask. 2015. *Trask's historical linguistics*. Routledge.

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

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

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

Pichel, J. R., P. Gamallo, and I. 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, T., L. Borin, G. Mikros, and J. Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification.

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

Rissanen, M. et al. 1993. The helsinki corpus of english texts. *Kyttö et. al*, pages 73–81.

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

Sennrich, R. 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.

Singh, A. K. and H. 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.

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

Venâncio, F. 2014. O castelhano como vernáculo do português.

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

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

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