Konputagailuen Arkitektura eta Teknologia Saila

Informatika Fakultatea

Medidas de distância entre línguas baseadas em corpus

Aplicação à linguística histórica do galego, português, espanhol e inglês

Memória de Tese

submetida para o título de Doutor em Informática por

José Ramom Pichel Campos

dirigida por Iñaki Alegria e Pablo Gamallo

Donostia, julho de 2020

"Falta a resposta! O galego são a mesma língua ou não?
A resposta depende de cada um de nós."
Marco Neves, O galego e o português são a mesma língua?.
Através Editora.
Junho 2019

"O galego ou é galego-português ou é galego-castelhano.
Ou somos umha forma do sistema ocidental
ou somos umha forma do sistema central.
Nom há outra alternativa"
Ricardo Carvalho Calero, "Sobre a nossa língua", em Problemas da Língua Galega, Sá da Costa Editora, 1981, pp. 19-21.

"¿Quién manda aquí, quién?
¿Quién manda aquí, quién?
Tiempo de ver como se levanta la gente.
Yo no necesito poder
Nork agintzen du, nork, nork, nork?
Nork agintzen du hemen eta hor."
Esne Beltza Ft Mala Rodríguez & Fermin Muguruza, 2014

"Mas o que salva a humanidade
É que não há quem cure a curiosidade"
Tom Zé, Vira Lata na Via Láctea, 2014

Dedicatória

No ano da distância social, lugar curioso para uma tese dedicada às distâncias, este trabalho é dedicado a todos os meus mais próximos, a quem também agradeço:

Os meus pais José Ramón Pichel e Mari Carmen Campos. Sem vós, não haveria nenhuma hipótese. Simplesmente. Também, ao meu irmão Juan Carlos e a Nei, Perla e Newton, porque na ROM do que é importante é ali onde morades.

À Sabela Fernández, pola tua extraordinária curiosidade e paixão pola educação, polas línguas e polas suas viagens. Polo teu humor surrealista. Sem a tua generosidade e entusiasmo, esta viagem com números também seria impossível.

Ao Iñaki Alegría e ao Paulo Gamalho, meus irmãos mais velhos mais do que co-directores. Convosco aprendim até a duvidar das minhas pequenas certezas. Este trabalho seria inviável sem a vossa paciência comigo, e sem o vosso conhecimento, ajuda, esforço e tempo.

Ao professor Ricardo Carvalho Calero, poeta, escritor, ensaísta. Uma figura fundamental na literatura e linguística galega do século XX e não só, injustamente marginalizado por ser reintegracionista e defender que o galego deve reencontrar-se com a sua família linguística: a portuguesa.

Finalmente, dedico todo isto a cada pessoa da minha família, especialmente aos meus sobrinhos e sobrinhas: Tiago, Mariña, Matilda, Rui e Antia. Sodes o futuro deste lugar no mundo, que sempre vai correlacionar com o futuro da nossa língua, e o futuro desta, com a viabilidade do nosso Planeta.

Agradecimentos

Os meus agradecimentos não cabem numa só secção porque na realidade a nossa história é uma engrenagem contínua de detalhes, que nos conduzirom até aqui, e só agora é que compreendemos. No entanto, arrisco a fazer uma síntese para este trabalho:

Ao Xoán Carlos Pardo, por me sugerir ainda na Universidade, que deveria ler sobre linguística computacional, porque eu ia adorar. Foi mesmo assim.

À Sabela Fernández, por abrir a possibilidade disto, deixando no congelador tantos projectos, que retomaremos.

Ao Iñaki Alegria por abrir as portas sem condições do teu humor, entusiasmo e sabedoria. Tantas viagens feitas e por fazer. Ao Paulo Gamalho por essa paixão e trabalho imensurável que pões em tudo. Essa camaradagem incondicional.

Ao José Ramon, Mari Carmen, Juan Carlos, Bea, Felisa, José Ramón, Alfonso, Bea, Ana, Martí, por estardes sempre pendentes daquilo que podia ser entregue algum dia, neste ano máscara.

A todo o reintegracionismo, porque dades o vosso melhor para que os galegos e galegas saibam que a nossa língua não é apenas nossa, mas da nossa cidade de 2,7M de pessoas numa rede de oito estados em todo o planeta.

Para imaxin|software, especialmente à Luz Castro, Diego Vázquez, Antonio Fernández, Dora Devesa, Ángel López, Antón Moura, Iago Bragado, Oscar Senra e o Paulo Malvar, por me dar durante 23 anos a oportunidade de aprender e aprender e aprender na prática, todo o pouco que sei sobre linguística computacional e seus derivados.

Para Ixa Taldea, Priberam, Eleka e Elhuyar, especialmente ao Iñaki Alegria, Kepa Sarasola, Arantza Diaz de Ilarraza, Eneko Agirre, Koldo Gojenola, Itziar González Dios, Xabier Artola, Xabier Saralegi, Gorka Labaka, Mikel Artetxe, Carlos Amaral, Josu Waliño e Josu Aztiria. Também especialmente à Olatz Arbelaitz e o Txus Pérez por tantas caminhadas e tantos

ânimos.

Ao Marcos Garcia, por todas as conversas na cafetaria perto das nossas casas, falando com a surpresa dos seus habitantes, sobre as suas línguas, os seus problemas, as suas distâncias, os seus métodos. Ao Paulo Lamas, porque sempre estiveste lá, agora em Portugal, afastando qualquer crença vestida de certeza. Quando eu mais precisava. Para o Valentim Fagim, porque desde adolescentes sonhamos com uma realidade diferente. Ao Ernesto Vázquez Souza, porque sempre colocaste essa distância que te permite ver qualquer período histórico, onde me faltou visão.

Ao Ramom Pinheiro (Chito) e à Sabela Vázquez, por me terdes acompanhado neste ano intenso com as suas estórias e histórias, as suas sanfonas, os seus humores, as suas caminhadas. Também ao Isaac González, ao Antom Meilam, à Mica Alvear e à Raquel Miragaia polas perguntas sobre os estados de saúde através do telegram. Ao Rafa Janeiro e ao Roberto Abuín, da Axóuxere Editora, pola abertura da vossa residência de escritores/as, agora também de teses. Um lugar onde as cousas mais importantes estão a ferver: amizade, humor e combate.

Ao Oscar Senra e à Noa Estevez, por aquelas perguntas sobre um trabalho que estava prestes a sair e finalmente. Ao Marcos Lorenzo, por esse entusiasmo patafísico de afinar o desafinado. Por desafinar o afinado, sempre. Ao Jorge Otero e à Laura Fernández por me abrirdes os braços em Sheffield, polo vosso entusiasmo e amizade, e por me ajudardes a ligar-me à vossa geração, uma das minhas grandes esperanças para este lugar no mundo.

À tertúlia do Burgo, durante anos, pensando e repensando a língua, com muita paixão. Ao Miguel Serrano, por teres falado da entropia naquela cafetaria. Ao Carlos Quiroga, por criar uma personagem fictícia no seu livro "Peixe babel" que molhava tortilha no café, assunto original que eu jamais teria feito. Também pola fascinação por este trabalho sempre em construção. Ao Ugio Outeiro e à Bárbara por essas fugidas muralhas. À Teresa Moure, polas conversas repousadas, alegres e caminhadas sobre a nossa língua e o seu futuro. Ao Marcos Paino, por essa banda sonora Das Kapital.

Ao Félix do Carmo, da University of Surrey, por essa sabedoria de viver com calma, acção e coerência. Por me ter chamado tantas vezes à cidade do Porto, para falar da tradução automática e inevitavelmente da língua que o Rio Minho não separa. Ao Luís Trigo, por também ter atravessado essa fronteira de conhecimento que é conhecer-nos galegos/as e portugueses/as. Com o tempo, verificamos que para além da língua, lá tudo começa.

Ao Marco Neves, por me ter permitido estar na Universidade Nova de Lisboa em 2019, com a enorme admiração que sinto polos teus extraordinários trabalhos cheios de inteligência, humildade, pedagogia e "retranca"

como dizemos na Galiza. À Vera Ferreira por sonhar e agir com humildade, constância e alegria em defesa das línguas mais desfavorecidas, como o minderico e muitas outras no mundo. Tantas portas ao exterior, Vera.

À Lucia Specia, do Imperial College London e da University of Sheffield, por me ter permitido, em 2015, fazer um pequeno estágio, iniciando todos estes trabalhos de tese, e abrindo todos os conhecimentos da sua comunidade de prestigiados investigadores em linguística computacional. Também por me tornar um membro do Rocódromo de Sheffield. A Frédéric Blain, polas conversas que me ajudaram a focar melhor as questões da tese. Ao Arkaitz Zubiaga, por me ter recebido naquela que foi a minha segunda viagem a Sheffield, lá em Birmingham. Ao José António Souto Cabo, da Universidade de Santiago, por partilhar comigo tanto conhecimento sobre a história da nossa língua, tanto na Galiza como em Portugal. Ao Xavier Varela e o Antón Santamarina, da Universidade de Santiago: por me terem facilitado o corpus de língua galega, que é o trabalho das suas vidas. Sem ele, nunca teria podido realizar a investigação relacionada com a língua na Galiza.

Conteúdo

Lista de Tabelas

Parte 1

Síntese

I. CAPÍTULO

Introdução

I.1 Motivação

Os habitantes da Galiza vivem entre duas línguas. Por um lado, o galego, língua própria da Galiza, falada e escrita desde o século XII até aos dias de hoje e variante originária do português. Por outro lado, o castelhano, também conhecido como espanhol, a língua de Castela e a única língua oficial do Estado espanhol desde a Constituição da Segunda República, em 1931 [SuanzeS-Carpegna, 2013], até aos dias de hoje.

O galego tem também uma peculiaridade que o torna ainda mais original: a sua classificação entre as línguas românicas. Assim, para linguistas como Teyssier [1982], o galego é uma língua independente, enquanto que para Freixeiro Mato [2000] é uma variedade do diassistema do português. Em qualquer caso, o que todos concordam é que o galego tem uma relação especial com as duas línguas românicas mais faladas no planeta, o português e o espanhol, o que o torna único no mundo.

Esta característica também afecta o processamento de linguagem natural, especificamente a identificação automática da língua e a tradução automática. Assim, o galego pode por vezes ser identificado como galego, por vezes como português e mesmo como espanhol, especialmente em textos curtos como os tweets. Em relação aos sistemas de tradução automática que incluem o galego, tais como *Google translate*, por vezes são geradas traduções mais compatíveis com o padrão português do que com o galego, o que leva a suspeitar que o corpus de treino para galego pode incluir parcialmente corpus de português europeu, devido a outras experiências semelhantes [Pichel et al., 2009, Malvar et al., 2010].

Uma das razões para estas singularidades pode ser encontrada na história da Galiza, que tem tido uma influência decisiva na história da língua. Primeiro, a separação do condado de Portugal do Reino de Galiza-Leão no século XII deu origem ao Reino de Portugal (actual República Portuguesa). Em segundo lugar, a satelização política da Galiza em relação a Castela, que começou em meados do século XIV e foi consumada no século XV: *"The 14th century, the frustration of attempts to separate Galicia from the Kingdom of Castile and the failure of attempts to unify Galicia with Portugal"* [Monteagudo and Santamarina, 1993].

Esta perda definitiva do poder político galego no final do século XV, foi em parte o resultado do apoio de parte da nobreza galega ao Reino de Portugal na luta pelos direitos ao trono de Castela. Assim, importantes elites galegas apoiaram o lado português liderado por Afonso V e a castelhana Joana, "A excelente Senhora" (conhecida de forma depreciativa em Castela como "La Beltraneja"), contra o lado castelhano-aragonês encarnado pelos Reis Católicos Isabel e Fernando. A derrota dos primeiros trouxe consigo: *"the final submission of the Galicians and the beginnings of their political dependence"* [Monteagudo and Santamarina, 1993], o que levou a um abandono progressivo da escrita em galego de forma maciça desde o século XVI até meados do século XIX, período conhecido nos estudos literários galegos como *"Os Séculos escuros"*.

Foi a partir do século XIX que a escrita em galego foi retomada graças a Rosalía de Castro, Curros Enriquez, Eduardo Pondal ou Johan Manuel Pintos, entre outros; adoptando, não sem controvérsia, diferentes ortografias que se afastavam das ortografias medievais e abordavam ortografias indistinguíveis do espanhol. Estas hesitações, especialmente desde o final do século XIX, foram uma consequência dos debates sobre a relação que o galego deveria ter com o português e o espanhol.

Devemos salientar aqui que estas disputas foram especialmente importantes na década de 80 do século XX, uma vez que o galego precisava de um padrão e de uma ortografia para a sua implementação no ensino obrigatório, que ocorria pela primeira vez na história. Por um lado, houve quem defendesse um padrão coerente com o medieval e o actual português, com ortografias próximas do galego medieval. Por outro lado, defendia-se um padrão baseado em opções dialectológicas com uma ortografia próxima do espanhol, língua na qual todos os galegos foram especialmente educados. Estas últimas teses, defendidas pelo catedrático asturiano Constantino García, venceram as defendidas pelo catedrático galego Carvalho Calero [Pichel Campos and Fagim, 2012]. Este último, a quem o Dia das Letras Galegas é dedicado em 2020, foi o principal promotor do chamado reintegracionismo linguístico com

o português [Collazo, 2014].

Resumindo o que foi dito anteriormente: as dúvidas sobre a classificação filogenética histórica e actual do galego e as hesitações na identificação automática da língua e na construção e concepção de tradutores automáticos, sugerem que o cálculo automático da distância entre o galego, o português e o espanhol, a partir de textos escritos reais, pode ser um desafio interessante.

Mas ao contrário do que possa parecer, este caso não é único no mundo. Outros casos semelhantes na Europa são a convergência e divergência histórica (mesmo ortográfica) entre moldavo e romeno, flamengo e holandês, catalão e occitano, ou três variantes linguísticas dos Balcãs: sérvio, croata e bósnio [Carrera, 2014].

Além disso, a distância entre línguas não afecta apenas estas últimas, uma vez que poderíamos calcular esta distância para todas as línguas independentemente da distância ou relação de proximidade entre elas (por exemplo: distância entre inglês e basco ou entre chinês e japonês). Com este cálculo podemos confirmar as hipóteses dos linguistas ou gerar novas observações, se existirem.

Para este fim, colocámos sete questões que orientaram a nossa investigação em todos os momentos e que são detalhadas a seguir:

1. Pode a distância entre línguas ser medida automaticamente com base em corpus?

2. Que papel desempenha a ortografia na distância entre as línguas?

3. É possível traduzir esta distância numa única métrica robusta?

4. A distância calculada com essa métrica verifica as hipóteses dos linguistas? Adiciona novos dados sobre hipóteses minoritárias ou controversas?

5. Será que a distância entre períodos históricos da mesma língua muda? Como?

6. A distância entre línguas muda historicamente ou é sempre a mesma? E se mudar, esta distância entre línguas é linear?

7. Será que a distância histórica entre variantes reconhecidas da mesma língua muda?

I.2 Estrutura da tese

Explicaremos agora a estrutura do relatório de tese, tendo em conta que é o resultado de uma colecção de artigos que foram produzidos ao longo dos últimos 5 anos:

Antes da apresentação dos artigos, e depois deste capítulo introdutório, o capítulo II apresenta em maior profundidade o quadro teórico que explica o estado da arte em filogenética e dialectologia computacional, bem como em identificação automática de línguas e medições de distância entre línguas, com especial atenção à nossa proposta. Também discutimos o trabalho relacionado com a normalização fonológica, pois queremos verificar o papel da ortografia na distância entre as línguas. O corpus utilizado nas nossas experiências é também introduzido em diferentes línguas e será pormenorizado mais tarde no Anexo V e, por fim, apresentamos os pacotes de software e recursos para poder replicar as diferentes experiências.

Mais adiante, no capítulo III, encontramos o resumo das teses, assim como as hipóteses e objectivos da tese no capítulo IV.

O capítulo V inclui a compilação e discussão dos resultados mais interessantes de todas as experiências apresentadas nos artigos. A seguir, no capítulo VI, detalhamos as conclusões e contribuições deste trabalho. Finalmente, no final desse capítulo, discutiremos o trabalho futuro que planeamos fazer para aprofundar esta área de investigação.

Nos anexos, sob a Parte 3, são incluídos os diferentes artigos científicos publicados em várias revistas (ou actas de conferências). Descrevemos cada um deles a seguir:

- O Anexo I inclui a publicação fundamental relacionada com os objectivos (O1, O2, O3 e O4) da tese: estudo e avaliação das métricas e definição de uma metodologia para realizar as experiências sucessivas. Fazemos também medições com uma ortografia transcrita. Esta é a base inicial para as restantes experiências realizadas nesta tese.

- O Anexo II recolhe duas publicações que estudam e quantificam a evolução histórica do português, do espanhol e do inglês. O primeiro artigo descreve as experiências, resultados e conclusões para o português; este é alargado no segundo artigo para incluir o espanhol e inglês. Ambos os estudos são especificamente dirigidos aos objectivos (O5 e O6) da tese.

- No Anexo III, concentramo-nos numa comparação cruzada da evolução histórica das línguas relacionadas. Os estudos e resultados do primeiro

artigo são alargados no segundo, comparando a evolução do galego, português e espanhol. Estes artigos estão especificamente relacionados com os objectivos (O5 e O6) da tese.

- No Anexo IV, o último artigo analisa a convergência/divergência entre variantes geográficas da mesma língua, comparando o português de Portugal e do Brasil e o espanhol de Espanha e da Argentina. Estão também especificamente relacionados com os objectivos (O5 e O6) da tese

- No Anexo V veremos em detalhe todo o corpus histórico feito para português europeu, português do Brasil, espanhol europeu, espanhol da Argentina, inglês e galego. Estão especificamente relacionados com o objectivo (O5) da tese

Finalmente, incluímos uma bibliografia relacionada com toda a investigação realizada.

I.3 Artigos publicados

Os artigos compilados neste relatório são os seguintes:

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

- J.R. Pichel, P. Gamallo, I. Alegria. 2018. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), 145–155.

- J.R. Pichel, P. Gamallo, I. Alegria. 2019. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. Natural Language Engineering, 1-22.

- J.R. Pichel, P. Gamallo, I. Alegria. 2019. Cross-lingual Diachronic Distance: Application to Portuguese and Spanish. Procesamiento del Lenguaje Natural 63 (2019): 77-84.

- J.R. Pichel, P. Gamallo, I. Alegria, M. Neves. 2020. A Methodology to Measure the Diachronic Language Distance between Three Languages Based on Perplexity. Journal of Quantitative Linguistics (2020): 1-31. DOI: 10.1080/09296174.2020.1732177

- J.R. Pichel, P. Gamallo, M. Neves, I. Alegria. 2020. Distância diacrónica automática entre variantes diatópicas do português e do espanhol. Linguamática 12, no. 1 (2020): 117-126.

Embora estejam fora do âmbito da tese, os seguintes artigos estão incluídos na mesma linha de investigação e estão relacionados com o presente trabalho:

Identificação automática de língua

- P. Gamallo, I. Alegria, J.R. Pichel, M. Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)
- P. Gamallo, I. Alegria, J.R. Pichel. 2017. A perplexity-based method for similar languages discrimination. Proceedings of the fourth workshop on NLP for similar languages, varieties and Dialects (VarDial4)

Outros

- Pichel, José Ramom, Paulo Malvar Fernández, Oscar Senra Gómez, Pablo Gamallo Otero, and Alberto García. Carvalho: English-Galician SMT system from EuroParl English-Portuguese parallel corpus. Procesamiento del lenguaje natural 43 (2009): 379-381.
- Malvar, Paulo, José Ramom Pichel, Óscar Senra, Pablo Gamallo, and Alberto García. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. Linguamática 2, no. 2 (2010): 31-38.
- Pichel, José Ramom, Paulo Malvar Fernández, Oscar Senra Gómez, Pablo Gamallo Otero, and Alberto García. Carvalho: Un sistema de traducción estadística inglés-galego construído a partir del corpus paralelo inglés-portugués EuroParl. Procesamiento del Lenguaje Natural 43 (2009): 379-381.
- P. Gamallo, M. Garcia, J.R. Pichel. A Method to Lexical Normalisation of Tweets. 2013. Proceedings of XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural. Workshop on Tweet Normalization at SEPLN.

- A. Zubiaga, I. San Vicente, P. Gamallo, J.R. Pichel, I. Alegria, N. Aranberri, Aitzol Ezeiza, Víctor Fresno. 2016. Tweetlid: a benchmark for tweet language identification. Language Resources and Evaluation 50 (4), 729-766

- P. Gamallo, S. Sotelo, J.R. Pichel, M. Artetxe. 2019. Contextualized Translations of Phrasal Verbs with Distributional Compositional Semantics and Monolingual Corpora. Computational Linguistics, 1-27.

I.4 Recursos gerados

O corpus utilizado nas nossas experiências e descrito no Anexo V está livremente disponível e consiste em:

- *Corpus de 44 línguas da Europa*[1], sendo um corpus sincrónico comparável de 44 línguas europeias que contém textos extraídos de Bíblias e de web-crawling.

- Corpus *Carvalho*[2], sendo um corpus diacrónico contendo cinco variedades linguísticas: Carvalho-PT-PT (português Europeu) e Carvalho-PT-BR (português do Brasil) para o português; Carvalho-ES-ES (espanhol Europeu) e Carvalho-ES-AR (espanhol da Argentina) para o espanhol e finalmente Carvalho-EN-UK (inglês britânico) para o inglês. Embora tenhamos criado também o corpus Carvalho-GL para o galego, este não pode ser descarregado por razões de direitos de autor.

  Quanto à dimensão do corpus, foram seguidos os critérios de dois autores do Corpus Helsinki de Historical English [Rissanen et al., 1993b], os quais indicam que: *"The size of the basic corpus is c. 1.5 million words"*. Portanto, o corpus de cada um dos períodos históricos de todas as línguas (galego, português, espanhol, inglês), desde a Idade Média até ao final do século XX, e variantes (português do Brasil, espanhol da Argentina) desde o final do século XX até aos dias de hoje, tem pelo menos esta dimensão.

  Finalmente, para que o corpus *Carvalho* seja representativo, tendo em conta a representatividade definida por Biber [1993], incluí-

[1] `https://gramatica.usc.es/pln/projects/distance.html`
[2] `https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho`

mos sistematicamente para cada período, de forma equilibrada, 50% textos de ficção e 50% de não-ficção.

Em relação ao software, libertámos os scripts[3] para o cálculo da distância linguística com base em *perplexity*, que foram desenvolvidos em *Perl* e *Bash*. Para as utilizar é necessário seguir as instruções incluídas no ficheiro README.

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

II. CAPÍTULO

Quadro Teórico

Neste capítulo analisaremos diferentes abordagens que procuram medir as mudanças que ocorrem interna ou externamente nas línguas.

Com este fim, iremos descrever metodologias baseadas em recursos linguísticos, especificamente no campo da filogenética e da dialectologia computacional, e depois aprofundar nas metodologias baseadas em corpus, um campo especialmente relacionado com a nossa investigação.

Dentro destas metodologias, daremos especial atenção ao trabalho relacionado com a identificação automática das línguas, o cálculo da distância automática entre as línguas em geral e as que utilizam a métrica de *perplexity* em particular.

A seguir, abordaremos o estudo de *perplexity*, métrica utilizada na nossa metodologia de cálculo da distância entre línguas. Também examinaremos a normalização fonológica que utilizámos para medir o impacto da ortografia como um factor relevante na distância entre línguas.

Finalmente, descreveremos os corpora utilizados nas diferentes línguas de estudo e o software criado necessário para medir esta distância.

II.1 Medidas de distância entre línguas

Muitas pessoas sentem intuitivamente que a distância entre línguas está relacionada com a capacidade de compreender línguas. A nível

oral, quanto menor for a compreensão, maior deve ser a distância, e vice-versa. A nível escrito, quanto pior compreendermos o alfabeto, maior a distância, e vice-versa.

Contudo, a ciência que estuda as línguas, a linguística, indica que, para conhecer a distância entre as línguas, devemos ter em conta não só o momento presente mas também as contínuas mudanças internas que nelas ocorrem ao longo da sua história. Temos também de ver se comparamos padrões, registos populares ou variedades dialectais, entre outros. Por outro lado, devemos decidir que aspectos da língua queremos medir: ortográficos, lexicais, morfológicos, sintácticos, semânticos ou pragmáticos.

A sociolinguística, o ramo da linguística que estuda as relações de convergência ou divergência entre línguas, também oferece contribuições que devemos ter em conta para compreender os resultados da nossa investigação. Assim, de acordo com Kloss, Heinz [1967], as línguas dividem-se em duas categorias em termos da sua relação com outras: "línguas por distância" (chamadas *Abstand*), separadas umas das outras por uma distância linguística significativa (por exemplo, basco e português), e "línguas por elaboração" (*Ausbau*) (por exemplo, sérvio, croata e bósnio), em que a distância pode convergir ou divergir de forma diferente ao longo da sua história. Também iremos considerar os sistemas *policêntricos* de línguas, línguas que têm diferentes centros de poder político e económico [Da Silva, 2018], que geram padrões linguísticos diferentes [Muhr, 2013].

Por estas razões, quantificar todos estes aspectos das línguas e reduzi-los automaticamente a uma medida robusta de distância que esteja alinhada com os pressupostos aceites da linguística é um grande desafio. Além disso, esta métrica, para ser consistente, deve não só quantificar a distância actual entre línguas, mas também a distância interna entre períodos históricos da mesma língua, a distância histórica entre línguas e mesmo a distância histórica entre variantes diatópicas das línguas.

Historicamente, tem havido abordagens diferentes em diferentes campos, como a filogenética, a dialectologia computacional, a aprendizagem de segunda língua e a identificação automática da língua dentro do processamento de linguagem natural.

O nosso trabalho enquadra-se numa área complementar à identificação das línguas. Assim, queremos investigar através de diferentes experiên-

cias uma metodologia que, utilizando modelos linguísticos baseados em corpus, consiga obter uma métrica robusta capaz de calcular diferentes distâncias entre línguas: sincrónica e diacrónica, intralinguística e interlinguística. Todos estes cálculos serão feitos a partir de um corpus equilibrado de ficção e não-ficção de diferentes línguas.

Para o cálculo desta distância, avaliámos diferentes medidas procurando sempre um compromisso entre a robustez das medidas e a obtenção de resultados alinhados com as hipóteses dos especialistas. Especificamente na nossa investigação, tentámos observar se as línguas evoluem historicamente de acordo com a opinião dos linguistas históricos, se a distância actual entre línguas (sincrónica) está em conformidade com a distância aceite pelos estudiosos das famílias linguísticas e das suas relações internas e externas, e se existe ou não linearidade na distância histórica entre línguas. Além disso, em alguns casos, obtivemos dados novos. Estes dados lançam luz sobre hipóteses controversas, que se revelam complexas de provar por outros meios.

Finalmente, em todas estas distâncias, observámos até que ponto a ortografia é um factor que contribui para distanciar ou aproximar línguas ou variantes linguísticas, e tentámos, como cálculo adicional, isolar este factor utilizando uma projecção fonológica dos textos no corpus.

II.1.1 Introdução e campos de aplicação

A medida da distância entre línguas foi abordada a partir de diferentes campos, começando com a comentada filogenética [Petroni and Serva, 2010], a dialectologia computacional [Nerbonne and Heeringa, 1997a] ou a aprendizagem de línguas [Chiswick and Miller, 2004]. Também tem sido realizada investigação a partir do estudo diacrónico sobre a evolução das línguas [Lai et al., 2018], no campo da dialectologia [Lui and Cook, 2013], ou no domínio da identificação automática de línguas e variantes [Jauhiainen et al., 2019, Zampieri et al., 2015, Molina et al., 2019].

O tema também despertou interesse em campos como os estudos de economia [Isphording and Otten, 2013], a distância cultural [West and Graham, 2004], estudos sobre a dinâmica da sobrevivência linguística [Mira and Paredes, 2005], o estudo da inteligibilidade mútua entre línguas [Gooskens et al., 2007] ou a aquisição de uma segunda língua [Chiswick and Miller, 2004].

Entre as diferentes propostas de métodos para calcular a distância entre línguas, algumas baseiam-se em comparações baseadas em recursos linguísticos (listas de palavras, dicionários, bases de dados com informações sobre diferentes aspectos linguísticos, etc.) e outras em comparações baseadas em corpus. O primeiro grupo de técnicas é utilizado principalmente em filogenética e dialectologia computacional e o segundo é comum no processamento de linguagem natural (NLP), especificamente na identificação automática da língua [Malmasi et al., 2016].

Em seguida, iremos rever os dois conjuntos de técnicas, baseados em recursos linguísticos e baseados em corpus, associando-os às áreas em que foram preferencialmente aplicados: filogenética e dialectologia, em que predominam as abordagens lexical e fonética/fonológica, e as áreas da NLP, em que predominam as abordagens baseadas em corpus. Centrar-nos-emos especialmente no segundo grupo, pois é o que mais se aproxima dos objectivos da nossa investigação.

II.1.2 Metodologias baseadas em recursos linguísticos

Filogenética

De acordo com Borin [2013], a filogenética e a dialectologia são os campos que tradicionalmente têm lidado com a distância entre línguas. Assim, este autor afirma que: *"traditionally, dialectological investigations have focused mainly on vocabulary and pronunciation, whereas comparative-historical linguists put much stock in grammatical features"* e *"we would expect the same kind of [language distance] methods to be useful in both cases"* [Borin, 2013, p. 7].

No caso da filogenética, um subcampo da linguística histórica e comparativa, o objectivo é classificar por meio de uma árvore a evolução histórica de um grupo de línguas, variantes linguísticas e línguas independentes. A partir desta classificação, podemos observar a distância de proximidade ou distância entre línguas.

A fim de construir automaticamente as árvores que representam a forma como um conjunto de línguas evolui ao longo do tempo [Barbançon et al., 2013], muitos investigadores têm utilizado a técnica conhecida como *lexicoestatística*. Esta abordagem da linguística comparativa faz uma comparação quantitativa de cognatos lexicais, que são palavras com uma origem histórica comum [Nakhleh et al., 2005,

Brown et al., 2008, Holman et al., 2008, Bakker et al., 2009, Petroni and Serva, 2010, Barbançon et al., 2013].

Mais especificamente, a *lexicoestatística* é baseada em listas de palavras interlinguísticas (cross-lingual word lists) (por exemplo, a Swadesh list [Swadesh, 1952] ou a base de dados AJSP [Brown et al., 2008]), medindo automaticamente a distância entre as línguas a partir da percentagem de cognatos partilhados. Entre as investigações relacionadas, podemos destacar Kolipakam et al. [2018], List et al. [2018] e Satterthwaite-Phillips [2011].

Outros métodos utilizam a distância de edição ou Levenshtein, nas suas diferentes variantes, para construir automaticamente estas árvores. Assim, a distância de Levenshtein é utilizada em Petroni and Serva [2011], também com uma distância padronizada de Levenshtein em Yujian and Bo [2007], ou uma relação entre línguas baseada na distância renormalizada de Levenshtein em Serva and Petroni [2008]. Em Petroni and Serva [2010], outra pesquisa baseada na distância de Levenshtein, o objectivo não era distinguir cognatos e não-cognatos, comparando exclusivamente a distância de Levenshtein entre palavras de uma lista multilingue aberta, mas encontrar uma média de todas as distâncias entre pares da lista.

Podemos também destacar os trabalhos de Müller et al. [2010], que utilizava técnicas baseadas no algoritmo de distância de Levenshtein e de neighbour-joining, método anteriormente utilizado para representar relações filogenéticas em biologia [Saitou and Nei, 1987]. Este algoritmo aplica-se a uma matriz de similaridade lexical de todos os pares possíveis de 4350 línguas uma comparação com palavras para 40 referências, gerando uma rede de distâncias entre as diferentes línguas. Destacamos também Ellison and Kirby [2006], que apresentaram um método chamado *PHILOLOGICON*, que constrói taxonomias linguísticas através da comparação de matrizes de semelhanças linguísticas internas de cada língua. Destacamos também, em relação às obras filogenéticas, a aplicação da distância de Levenshtein às línguas de interesse no nosso trabalho, como é o caso do galego em relação a outras línguas românicas em Alecha and González [2016].

É importante sublinhar que, embora tenhamos destacado na *Filogenética* as técnicas lexicais não baseadas em corpus, também vale a pena mencionar neste campo os trabalhos de Satterthwaite-Phillips [2011] e Rama and Singh [2009], que testaram várias técnicas para construir ár-

vores filogenéticas a partir de corpus: *cross-entropy, cognate coverage distance, distância fonética de cognados e n-gramas*. Estes investigadores concluíram que estas medidas podem ser muito úteis para as línguas que não têm listas manuais e mesmo para as que já as têm. Além disso, salientamos Singh and Surana [2007], que aplicou a *cross-entropy* para identificar famílias de línguas diferentes no subcontinente indiano.

Finalmente, uma estratégia diferente é a baseada em técnicas tradicionais de machine learning. O conjunto de dados anotados contém diferentes tipos de características linguísticas que representam informação tipológica [Michael, 2015, Nichols and Warnow, 2008]. As características não são apenas lexicais, mas também podem ser fonológicas ou mesmo sintácticas. Um interessante conjunto de dados para a aprendizagem destes modelos é descrito em Carling et al. [2018].

Dialectologia computacional

Tal como na filogenética, a maioria dos trabalhos em dialectologia baseia-se em listas de palavras. A abordagem computacional, também conhecida como *dialectometria*, funciona geralmente a partir de listas de parâmetros linguísticos correspondentes aos diferentes dialectos.

Se a dialectologia estuda as características das variedades linguísticas de uma língua, especialmente as relacionadas com o léxico e a fonologia, a *dialectometria* estuda a distância entre dialectos com base na comparação de grandes quantidades de dados que codificam as características dialectais de um determinado espaço geográfico.

Isto nem sempre é fácil, devido, por um lado, às diferentes características que cada língua possui (vocabulário, rasgos fonéticos, prosódicos, sintácticos o doutra índole) e, por outro lado, à riqueza e heterogeneidade que pode existir em cada um dos dialectos; razões suficientes para aceder a técnicas computacionais para processar estas distâncias.

Entre os trabalhos mais notáveis estão os de Séguy [1971] que iniciaram a investigação no campo da *dialectometria*, ou os de Goebl [1982a,b, 2006], que marcou o início e reforçou a chamada Escola de Dialectometria de Salzburgo. Por outro lado, as obras de Nerbonne and Kretzschmar Jr [2013], Nerbonne and Kretzschmar [2006], Nerbonne et al. [1996] e Heeringa and Nerbonne [2001, 2013] prestigiaram a "Escola de Dialectometria de Groningen". É importante destacar

aqui o trabalho de Dubert and Sousa [2016], que desenvolveram uma metodologia específica para uma das variedades românicas em estudo: o galego.

Além disso, vale a pena mencionar o trabalho de Wieling and Nerbonne [2015], que expõe o estado da arte dos avanços na *dialectometria*, e os de Donoso and Sánchez [2017], que aplicam a análise dialectométrica a textos curtos e ruidosos, como os encontrados no Twitter.

Finalmente, embora as metodologias baseadas em corpus sejam discutidas na seguinte secção, devemos também citar neste campo vários trabalhos de Szmrecsanyi [2008, 2011], que faz investigação em *dialectometria* com técnicas baseadas em corpus.

II.1.3 Metodologias baseadas em corpus

No processamento de linguagem natural, este tipo de metodologia tem sido aplicada com sucesso em várias tarefas. Destaca-se a identificação automática das línguas ou o cálculo da distância entre línguas ou variedades de línguas pelos seus bons resultados. O que une ambas as tarefas é que são produzidas a partir de corpus, tão grande quanto possível, com textos paralelos multilingues ou multidialectos, de preferência. Para ambas as tarefas, foram utilizadas diferentes técnicas, que serão detalhadas a seguir. Destacam-se as que utilizam modelos linguísticos construídos a partir de n-gramas.

Identificação automática de línguas

A identificação linguística é um subcampo da linguística computacional que tem sido amplamente estudado ao longo dos últimos cinquenta anos [Jauhiainen et al., 2019]. É considerado um problema resolvido quando as línguas são distantes e os textos não são excessivamente curtos. No entanto, quando as línguas estão próximas ou muito próximas, ainda é considerado um desafio a ser resolvido [Baldwin and Lui, 2010]. Portanto, as áreas de maior actividade neste campo são a identificação automática da língua em textos curtos e/ou ruidosos, tais como tweets [Gamallo et al., 2014, Zubiaga et al., 2015, 2016] e a classificação de línguas ou variedades diatópicas estreitamente relacionadas [Malmasi et al., 2016, Zampieri et al., 2018, Kroon et al., 2018].

O desafio nesta área é identificar correctamente os textos escritos em línguas muito próximas umas das outras (por exemplo, bósnio e croata), consideradas por Kloss, Heinz [1967] como línguas *Ausbau*, ou "línguas por elaboração" e variedades diatópicas de uma língua (por exemplo, distinguir o espanhol argentino e europeu, ou o português brasileiro e o português europeu)

Além disso, quando as línguas podem ser consideradas línguas historicamente independentes ou variantes da mesma língua, e são escritas em ortografias diferentes [Suzuki et al., 2002], como é o caso do bósnio (ortografia latina) e do sérvio (ortografia cirílica), a identificação automática da língua é uma tarefa mais complicada [Tiedemann and Ljubešić, 2012].

Tradicionalmente, técnicas baseadas em recursos linguísticos têm sido utilizadas na identificação automática da língua: bolsas de palavras, dicionários baseados em listas de palavras ou diferentes tipos de heurísticas (ortográficas, morfológicas e sintácticas), bem como abordagens estatísticas.

Entre as numerosas investigações nestas abordagens, destacamos o artigo "N-gram-based text categorization" [Cavnar et al., 1994] que é um dos primeiros artigos a utilizar n-gramas para identificação automática de línguas e "Statistical Identification of Language" [Dunning, 1994].

Os modelos basados en n-gramas de palavras e n-gramas de caracteres extraídos a partir de corpus costumam ser os melhores na identificação de línguas, especialmente os n-gramas de caracteres [Cavnar et al., 1994, Jauhiainen et al., 2019]. A razão provável para que estes sistemas se destaquem dos outros é que os n-gramas de caracteres não só codificam informação lexical e morfológica, mas também características fonológicas, uma vez que os sistemas fonográficos escritos estão relacionados com a forma como as línguas eram pronunciadas no passado. Se os n-gramas são suficientemente amplos, também codificam as relações sintácticas, pois podem representar o fim de uma palavra e o início da seguinte numa sequência [Pichel et al., 2019b].

Na investigação com n-gramas, destacamos, pela sua qualidade e pela sua relação com as nossas línguas de estudo, o trabalho de Zampieri and Gebre [2012], que aplica um método de estimação (log-likelihood) sobre n-gramas de caracteres, alcançando excelentes resultados na distinção entre português europeu e português brasileiro, com uma pre-

cisão de 99.5%. Também em Zampieri et al. [2013], onde diferentes técnicas baseadas em n-gramas de caracteres, unigramas de palavras, bigramas de palavras e informação morfológica e POS foram aplicadas para distinguir as variantes do espanhol argentino e do espanhol mexicano. Como resultado, obtém-se também uma precisão extremamente elevada de 99.9% na distinção entre o espanhol mexicano e o argentino.

Nas investigações desta área, queremos destacar o workshop *Natural Language Processing for Similar Languages, Varieties and Dialects* (VarDial). Durante os últimos anos, as suas campanhas de avaliação são uma referência essencial nesta sub-área, sendo os relatórios de revisão dos sistemas participantes, o estado da arte [Zampieri et al., 2018, 2019].

Neste workshop, podemos ver como, por exemplo, no Shared Task GDI (German Dialect Identification) de 2016, os melhores sistemas de identificação linguística baseavam-se em n-gramas de caracteres [Malmasi et al., 2016]. Em 2018 os dois melhores sistemas baseavam-se também em modelos de n-gramas. O melhor resultado foi alcançado por um sistema baseado em 4-gramas de caracteres. Finalmente no ano 2019 com cinco Shared Tasks (GDI, CMA, DMT, MRC e CLI) Zampieri et al. [2019] existem diferentes algorimos de classificação como *SVM*, *Bayes*, *Random forest* e *neuronais* utilizando a maioria dos mais bem classificados n-gramas de caracteres.

Por fim, gostaríamos de comentar que se esperava que as abordagens baseadas na aprendizagem profunda *deep learning* [Lopez-Moreno et al., 2014, Gonzalez-Dominguez et al., 2014, Criscuolo and Aluisio, 2017] pudessem trazer melhorias na identificação automática de idiomas. No entanto, na *Evaluation Campaign* mais recente organizada no Workshop on VarDial-2019 foi confirmado que as abordagens mais sofisticadas baseadas neste tipo de sistemas avançados não superam as estratégias mais tradicionais baseadas em n-gramas e classificadores com *Naive Bayes* ou *Support Vector Machine* [Zampieri et al., 2019].

Também em Jauhiainen et al. [2019] em relação a esta questão comenta-se que: *"Barman et al. (2014b) extracted features from the hidden layer of a Recurrent Neural Network ("RNN") that had been trained to predict the next character in a string. They evaluated several features with a SVM classifier, and found the RNN-extracted features alone were far inferior to character n-grams. However, the RNN features slightly im-*

proved the results when they were added".

Por fim, em [Zampieri et al., 2019] conclui-se que: *"From the obtained results we can see that sophisticated approaches involving Deep Learning models do not necessarily outperform the traditional methods like Naive Bayes or SVM ".*

Distância automática entre línguas

Nas metodologias baseadas em corpus para o cálculo automático da distância entre línguas, existem diferentes técnicas, tais como: distâncias lexicais, fonológicas, baseadas em n-gramas ou neuronais.

No que diz respeito ao cálculo da distância linguística a partir de *distâncias lexicais*, destacamos aqueles que utilizam modelos linguísticos complexos construídos a partir de informação distributiva das palavras obtidas a partir dos corpora. Assim, em Liu and Cong [2013], Gao et al. [2014], a partir de corpus paralelo, as redes de co-ocorrência de palavras são construídas para classificar as línguas em detalhe. Também são utilizadas medidas de distância baseadas em word embeddings (WELD) a partir de corpora paralelos como é no caso de Asgari and Mofrad [2016].

Outros métodos têm-se baseado em *distâncias fonológicas e fonéticas* entre línguas. Eden [2018] mede distâncias fonológicas enquanto Nerbonne and Heeringa [1997b] realiza uma comparação interlinguística de formas fonéticas, embora alguns investigadores, tais como Singh and Surana [2007] *"have argued against the possibility of obtaining meaningful results from crosslingual comparison of phonetic forms"* .

Em qualquer caso, as técnicas mais utilizadas para calcular a distância entre línguas baseiam-se principalmente em n-gramas e não em palavras ou fonemas. Estas técnicas, que foram anteriormente utilizadas para medir distâncias entre *strings* como em Kondrak [2005], foram muito bem sucedidas, e para estes cálculos são utilizadas diferentes métricas, tais como a *entropia*, a *cross-entropy* ou a *perplexity.*

Deve-se notar que os n-gramas de caracteres extraídos do corpus [Malmasi et al., 2016], apesar de serem unidades estatísticas e não linguísticas, são capazes de codificar informação lexical e morfológica e até características fonológicas e até outras relações mais complexas. Isto deve-se em parte ao facto de os sistemas de escrita estarem relacionados com a forma como as línguas eram pronunciadas no passado.

Por outro lado, quando os caracteres n-gramas têm um certo tamanho (mais de cinco é geralmente um tamanho adequado), são capazes de codificar relações sintácticas e sintagmáticas porque podem representar o fim de uma palavra e o início da seguinte na mesma sequência.

Por exemplo, o 7-grama *ion#de#* (onde '#' representa um espaço em branco) é uma sequência frequente de letras partilhada por várias línguas romance (p.e. espanhol, francês ou galego). Poderíamos considerar este 7-grama como uma instância de um padrão genérico *"nome-preposição-nome"*, porque *ion* é um sufixo do substantivo e *de* uma preposição muito frequente, que normalmente introduz frases preposicionais.

Para além de línguas diferentes, foram utilizadas técnicas de distância baseadas em n-gramas para medir a distância entre variantes linguísticas estreitamente relacionadas (p.e. espanhol de Nicarágua e de El Salvador) ou em textos muito curtos, onde é necessária maior precisão, como em Purver [2014], Porta and Sancho [2014] e Goutte et al. [2016].

Foram também desenvolvidas diferentes técnicas para distinguir línguas de dialectos [Wichmann, 2016], medir as diferenças entre as variedades linguísticas da mesma língua [Nerbonne and Hinrichs, 2006, Heeringa, 2004, Nerbonne and Heeringa, 1997b, Kessler, 1995] ou classificar sistemas policêntricos de línguas [Zampieri and Gebre, 2012]. Além disso, foram utilizados para medir a evolução histórica de uma língua através do cálculo da distância diacrónica entre textos na mesma língua [Zampieri et al., 2016].

Além disso, em Boldsen et al. [2019] foi utilizada uma metodologia que combina *perplexity*, *Redes Neuronais Recorrentes (RNN)* e *clustering* mediante *algoritmo K-Means* para identificar tendências temporais num corpus de documentos medievais.

Finalmente, estas técnicas de distância automática entre línguas foram aplicadas a outros campos e problemas diferentes dos já mencionados. Assim, Degaetano-Ortlieb et al. [2016] apresentam uma abordagem informativa-teórica baseada na *entropia* para investigar a mudança diacrónica em inglês científico, em Buckley and Vogel [2019] para investigar as mudanças diacrónicas no inglês medieval e Degaetano-Ortlieb and Teich [2018] utilizam a *entropia relativa* para a detecção e análise de períodos de mudança linguística diacrónica. Também foi aplicada a outras tarefas linguísticas computacionais do ponto de vista histórico, tais como *stance evolution* (evolução de posições) [Lai et al., 2018], ou

para medir como o inglês científico evolui diacronicamente através do uso da entropia em Degaetano-Ortlieb et al. [2016].

Distância automática entre línguas baseada em *perplexity*

O nosso trabalho é enquadrado dentro das metodologias de distância entre línguas baseadas em n-gramas a partir de corpora multilingues. A intenção da nossa abordagem tem sido verificar se a métrica de *perplexity* era capaz, não só de verificar hipóteses assumidas pela linguística, mas também de lançar novos dados sobre hipóteses minoritárias e mesmo controversas. Também quisemos observar o papel desempenhado pela ortografia na relação de distância entre línguas.

Para tal, a nossa metodologia tem sido aplicada a diferentes corpora multilingues construídos ad hoc em ortografia original, e com uma base equilibrada de ficção e não-ficção. É importante notar que o método é totalmente automático e pode ser aplicado a qualquer língua ou variante linguística.

A seguir detalhamos as diferentes experiências que realizámos aplicando *perplexity* para o cálculo das seguintes distâncias:

- Cálculo da distância sincrónica entre línguas [Gamallo et al., 2017a]. Neste trabalho, medimos a distância entre 44 línguas europeias utilizando duas medidas, uma das quais é *perplexity*. Como resultado, podem ser observadas as relações no seio das diferentes famílias linguísticas europeias e as línguas isoladas da Europa.

- Cálculo da distância histórica aplicada a uma língua (português) [Pichel et al., 2018]. Neste trabalho, medimos a história da língua portuguesa comparando mediante *perplexity* a distância entre os seus períodos históricos desde a Idade Média até ao final do século XX.

- Cálculo da distância histórica aplicada a mais do que uma língua (inglês, português, espanhol) [Pichel et al., 2019b]. Neste trabalho, melhorámos a metodologia anterior para medir utilizando *perplexity* a história de cada uma delas, desde a Idade Média até ao final do século XX.

– Cálculo da distância histórica entre duas línguas próximas (português e espanhol) [Pichel et al., 2019a]. Aqui medimos a relação histórica de convergência e divergência entre estas duas línguas românicas usando *perplexity*.

– Cálculo da distância histórica entre duas línguas próximas e outra intimamente relacionada com ambas [Pichel et al., 2020a]. Neste trabalho, melhorámos a metodologia para medir a distância histórica entre três línguas *Ausbau* (português, galego e espanhol), em que é controverso se uma delas (galego) pode ser uma língua independente das outras duas (português e espanhol) ou se é uma variante do diassistema linguístico do português [Carvalho, 1979].

– Cálculo da distância histórica entre as variedades diatópicas de duas línguas (português europeu/português do Brasil) e (espanhol europeu/espanhol da Argentina) [Pichel et al., 2020b]. Neste caso, quisemos comparar a distância em dois períodos históricos entre pares de variedades diatópicas de português e espanhol: português europeu - português do Brasil e espanhol europeu - espanhol da Argentina.

II.2 Perplexity

Um dos objectivos fundamentais desta tese é a escolha e avaliação de uma medida de distância entre línguas que seja efectiva e robusta.

Em trabalhos anteriores e graças à participação em sucessivas tarefas partilhadas VarDial [Gamallo et al., 2016, 2017b], a *perplexity* foi escolhida e avaliada como a medida mais adequada. A razão é que, por um lado, é simples e robusta porque não precisa de ser ajustada para cada tarefa e, por outro lado, aproxima-se dos melhores resultados das tarefas partilhadas. Além disso, não se baseia num classificador, mas num valor que pode ser interpretado como distância, por oposição a valores mais opacos de classificadores.

Uma vez escolhida como candidata para os objectivos da tese, esta medida foi utilizada e avaliada nas diferentes tarefas que fazem parte deste trabalho, confirmando a hipótese inicial de que poderia ser uma medida adequada aos nossos objectivos.

Em seguida, iremos definir formalmente a medida e a sua configuração para as tarefas de medição da distância entre línguas.

II.2.1 Definição de *perplexity*

Perplexity é uma medida amplamente utilizada para avaliar a qualidade dos modelos linguísticos construídos a partir de n-gramas extraídos de corpus de texto [Chen and Goodman, 1996, Sennrich, 2012a, Dieguez-Tirado et al., 2005].

Perplexity mede o quão bem um modelo linguístico prevê uma amostra de texto nunca antes vista. Portanto, se a *perplexity* é baixa, o modelo linguístico é eficiente na previsão da amostra de texto. Pelo contrário, com uma *perplexity* alta, o modelo linguístico não é aceitável para prever a amostra de texto em questão. Ou, de outro ponto de vista, se a medida for baixa, a amostra pertence à língua representada pelo modelo e se for alta, não.

Esta métrica tem sido utilizada em diferentes tarefas. Para além das mencionadas (identificação automática de línguas relacionadas [Gamallo et al., 2016] e distância entre línguas [Gamallo et al., 2017a]), foi utilizada na classificação de tweets formais e coloquiais [González, 2015], na estimação da dificuldade das tarefas de reconhecimento da fala [Jelinek et al., 1977], em várias tarefas dentro da tradução automática [Sennrich, 2012b] ou mais próximo da nossa investigação, na avaliação dos sistemas de reconhecimento de voz para galego e espanhol [Dieguez-Tirado et al., 2005].

Formalmente, a *perplexity* (PP para abreviar) de um modelo de língua que prevê um teste, é a probabilidade inversa do teste normalizado pelo número de caracteres:

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

onde as probabilidades $P(\cdot)$ dos n-gramas são definidos da seguinte forma:

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

A equação II.2 estima a probabilidade de n-gramas, dividindo a frequência observada (C) de uma sequência particular de caracteres e onde o prefixo representa a mesma sequência sem o último caractere. Para ter em conta os n-gramas não vistos, utilizámos uma técnica de suavização baseada na interpolação linear.

A partir desta definição de *perplexity*, definimos a nossa distância entre línguas chamada PLD, utilizado pela primeira vez no nosso trabalho Gamallo et al. [2017b]. Assim, para calcular a PLD entre duas línguas e para podermos ter resultados comparáveis nas diferentes experiências realizadas, construímos modelos de língua baseados em 7-gramas de corpus de línguas, períodos históricos de línguas ou variedades diatópicas de línguas, de pelo menos 1.25 milhões de palavras para o corpus do modelo de lingua (MDL) e de 250 mil palavras para o corpus de teste.

Especificamente, o cálculo da PLD é o resultado do cálculo da média aritmética da PP do modelo de língua da *Língua1* (LM_L1) e o texto de test da *Língua2* (CH_L2) e a PP do modelo de língua da *Língua2* (LM_L2) e o texto de test da *Língua1* (LM_L1). Esta comparação deve ser feita em duas direcções, uma vez que a PP é uma divergência com valores assimétricos.

Assim, a nossa distância entre línguas baseada em *perplexity* PLD é definida da seguinte forma:

$$PLD(L1, L2) = \frac{1}{2}(PP(CH_{L2}, LM_{L1}) + PP(CH_{L1}, LM_{L2})) \quad \text{(II.3)}$$

Quanto mais baixa a *perplexity* tanto de CH_{L2} dado LM_{L1} como de CH_{L1} dado LM_{L2}, menor será a distância entre os idiomas $L1$ e $L2$. É importante notar que a PLD é a média das duas divergências assimétricas: $PP(CH_{L2}, LM_{L1})$ e $PP(CH_{L1}, LM_{L2})$.

Dado que realizámos diferentes tarefas para medir a distância entre línguas, fizemos pequenas adaptações que não comprometem esta fórmula geral. Foi assim que o aplicámos ao cálculo da distância entre períodos históricos em três línguas diferentes [Pichel et al., 2019b], distâncias interlinguísticas entre períodos históricos de duas línguas próximas [Pichel et al., 2019a], distâncias interlinguísticas diacrónicas entre duas línguas próximas e uma muito próxima entre as duas [Pichel et al., 2020a], ou distâncias diacrónicas entre variedades diatópicas de duas línguas diferentes [Pichel et al., 2020b].

II.2.2 Abordagem através dum exemplo

Vamos explicar de uma forma mais intuitiva, através de um exemplo, o cálculo da medida. É um exemplo *"de brincadeira"*, para espanhol e português, que utiliza textos muito curtos, mas que serve como uma boa aproximação. Para isso precisaremos de dois textos modelo e dois textos de teste para ambas as línguas. Os textos modelo que mostramos a seguir, foram extraídas das duas primeiras frases das entradas *Portugal* e *España* das respectivas wikipédias.

Texto para o modelo em português:

Portugal, oficialmente República Portuguesa é um país soberano unitário localizado no sudoeste da Europa, cujo território se situa na zona ocidental da Península Ibérica e em arquipélagos no Atlântico Norte. O território português tem uma área total de 92 090 km^2, sendo delimitado a norte e leste por Espanha e a sul e oeste pelo oceano Atlântico, compreendendo uma parte continental e duas regiões autónomas: os arquipélagos dos Açores e da Madeira.

Texto para o modelo em espanhol:

España, también denominado Reino de España, es un país transcontinental, miembro de la Unión Europea, constituido en Estado social y democrático de derecho y cuya forma de gobierno es la monarquía parlamentaria. Su territorio, con capital en Madrid está organizado en diecisiete comunidades autónomas, formadas a su vez por cincuenta provincias; y dos ciudades autónomas.

Para os dois textos de teste, teríamos uma frase em cada língua.

Em português: *Nós falamos português.*

Em espanhol: *Nosotros hablamos español.*

Como foi indicado que a distância é assimétrica, temos de calcular o valor de PP (equação II.1) em cada sentido (LM espanhol, CH português) e (LM português, CH espanhol), obtendo posteriormente a média aritmética como o valor final de PLD na equação II.3.

Os resultados da distância que são obtidos para estes textos modelo e textos teste, estão reflectidos na Tabela II.1.

É importante lembrar novamente que quanto mais baixo for o valor da medida mais próximas as línguas são consideradas. É, portanto, surpreendente que o valor entre *pt* e *es* é inferior à que existe entre *es*

Modelo	Test	Mod. a Test	Test a Mod.	Média
es	pt	2492.71	214.12	1353.42
pt	pt	121.42	121.42	121.42
es	es	428.04	428.04	428.04

Tabela II.1: Valores de *perplexity* para o exemplo

e *es*, mas isto deve-se ao pequeno tamanho do corpus de teste (uma única frase), e ao do modelo (um pequeno parágrafo de duas frases).

Para o nosso exemplo, escolhemos o cálculo de PP no sentido de (LM espanhol, CH português), sendo o nosso objectivo explicar os passos que dão origem ao valor 242.12 na Tabela II.1, desde o modelo de língua espanhol e o teste de português (LM espanhol, CH português).

Para o efeito, veremos todas as fases do algoritmo de cálculo da PP(LM espanhol, CH português) e que explicamos em pormenor a seguir:

1. No nosso exemplo, como escolhemos o sentido (LM espanhol, CH português), temos primeiro de calcular o modelo de língua espanhol. Para tal, calculamos todas as probabilidades dos n-gramas de caracteres (2<=n<=7) a partir do início do texto. A probabilidade destes n-gramas será entre [0,1].

2. A seguir, faremos cálculos no texto do teste português. Para este fim, construímos grupos de n-gramas dos textos do teste. Para tal, começaremos na posição 0 do início do texto "Nós falamos português", realizando progressivamente desde o seu 7-grama até o seu bigrama e assim por diante até que o texto esteja completo. Por exemplo, se iniciarmos a construção do grupo de n-gramas na posição 10 desde o começo do texto de teste, estes serão todos os n-gramas do grupo: 7-grama[o,s,#,p,o,r,t], 6-grama[o,s,#,p,o,r], 5-grama [o,s,#,p,o], 4-grama [o,s,#,p], 3-grama [o,s,#], 2-grama[o,s]. Depois continuaríamos na posição 11 e assim sucessivamente.

3. Para cada um dos grupos de n-gramas em cada posição procuramos a probabilidade de cada n-grama no modelo de língua do espanhol. Por exemplo: O trigrama [o,s,#] tem uma probabilidade no texto do modelo de língua do espanhol de: 1, ou o trigrama [l,a,m] tem uma probabilidade no texto do modelo de língua de

n	n-grama	probabilidade
3	o s #	1
3	s # f	0.1111
3	l a m	0.3333
3	o s #	1
4	# p o r	1
3	p o r	1

Tabela II.2: N-gramas comuns a ambos textos (com n maior a 2)

espanhol de: 0.3333. Por outras palavras, os valores diferentes de zero correspondem com os n-gramas ($n >= 2$ y $n <= 7$) da frase em português que aparecem no corpus de treino do modelo em espanhol. Estes valores são especificados na Tabela II.2 (removendo os bigramas a fim de não alongar) juntamente com a probabilidade dependendo do modelo de treinamento:

4. Depois realizamos a suavização por interpolação linear para cada grupo de n-gramas. Para isso multiplicaremos as probabilidades de cada n-grama em cada grupo por um peso atribuído aos tipos de n-grama (2<=n<=7). O peso p_i é maior quanto maior é o n-grama e a soma dos pesos deve ser 1. Mostramos um exemplo deste cálculo na equação II.4 cujo resultado é -3.2712.

$$log(0 * p_7 + 0 * p_6 + 0 * p_5 + 0 * p_4 + 1 * p_3 + 0.0357 * p_2) \quad \text{(II.4)}$$

5. Este é o resultado do grupo de n-gramas na posição 10 do texto do teste português. A soma de todos os grupos de n-gramas no texto corresponde ao denominador da equação II.1. Após completar a equação PP obteremos como resultado o valor (214.12), a partir do modelo em espanhol e o teste em português.

6. Uma vez calculado PP no sentido (LM espanhol, CH português), vamos voltar à primeira fase deste algoritmo para calcular a direcção oposta PP(LM português, CH espanhol).

7. Finalmente, calcularemos a PLD da equação II.3, o que resulta na distância 1353.42, como indicado na última coluna da Tabela II.1

II.3 Normalização fonológica

Ao calcular a distância entre línguas, a influência da ortografia pode ser um factor muito importante, que merece ser estudado. Por este motivo, foram tomadas duas decisões metodologicamente relevantes no nosso trabalho: por um lado reunir corpora com a ortografia o mais próxima possível do original e, por outro lado, nas nossas experiências, medir a distância com a ortografia original e com uma ortografia transcrita fonologicamente de forma automática por meio de um transcritor desenvolvido *ad hoc*. O objectivo é observar o papel que a ortografia desempenha nessa distância. A partir de agora, vamos nomear a ortografia original com o acrónimo (OS) de *original spelling* e a ortografia transcrita com o acrónimo (TS) de *transcribed spelling*.

Como trabalhos relevantes em transcrição ou transliteração ortográfica (*machine transliteration*), podemos salientar Knight and Graehl [1998], que desenvolveu uma série de métodos para a transliteração automática (back-transliteration) entre japonês e inglês, e a Haizhou et al. [2004] que propôs uma transliteração baseada em n-gramas para o par chinês-inglês. Podemos também destacar o trabalho de Al-Onaizan and Knight [2002] que desenvolveram um sistema de transliteração de nomes entre o árabe e o inglês utilizando transdutores.

Na subárea da normalização de textos históricos, Bollmann [2019] compilou os diferentes métodos (baseados em regras, medidas de distância, métodos estatísticos ou métodos neuronais) que têm sido utilizados com o objectivo de projectar variações ortográficas de palavras num lema actual. Schneider [2002] desenvolveu um sistema de padronização ortográfica chamado Zenspell, que projecta palavras de textos do século XVIII para a actual ortografia inglesa e Reynaert et al. [2012] propõe dois métodos estatísticos para a padronização ortográfica de textos históricos em português.

Em relação à transcrição fonológica, destacamos o trabalho de Porta et al. [2013], que criaram transdutores que incluem um transcritor fonológico, entre outros dispositivos, para padronizar antigas formas de espanhol, e o de Satapathy et al. [2017] que *"proposes a phonetic-based framework for normalizing microtext to plain English and, hence, improve the classification accuracy of sentiment analysis"*.

Destacamos também as abordagens neuronais, como nos trabalhos de Tang et al. [2018], que aplicam modelos de Neural Machine Translation

(NMT) ao problema da padronização da ortografia histórica de diferentes línguas: inglês, alemão, húngaro, islandês e sueco; conseguindo melhores resultados do que o Statistical Machine Translation (SMT).

No nosso trabalho fizemos uma simples transcrição *ad hoc* que normaliza a grafia dos textos independentemente da língua, período histórico ou variedade diatópica, para uma grafia artificial próxima da fonológica. Para o efeito, o nosso transcritor utiliza como alfabeto final um conjunto de 34 símbolos, representando 10 vogais (que incluem variação de pronúncia dentro da mesma língua) e 24 consoantes, concebidos para cobrir a maior parte dos sons mais comuns, incluindo várias palatalizações de consoantes.

Esta normalização quase fonológica permite simplificar e homogeneizar os casos em que sons semelhantes (geralmente palatalizações) são transcritos de forma diferente em línguas diferentes. Assim, o som nasal palatizado é transcrito pelo transcritor como "ny", unificando, por exemplo, a ortografia portuguesa "nh" e a galega e a espanhola "ñ". Da mesma forma, a palatal lateral é transcrita como "ly", unificando as duas ortografias diferentes: "lh" em português e "ll" em galego e espanhol. O som africado palatal em galego e espanhol, bem como em português, representado pela ortografia "ch", é transcrito como "ĉ".

Vejamos na Tabela II.3, a diferença entre um texto português do século XIX na grafia original (OS) e a grafia transcrita automaticamente (TS) usando o nosso transcritor. Pode ver-se a negrito os n-gramas onde o transcritor age construindo uma ortografia artificial e a diferença que existe na transcrição com uma ortografia modernizada editada manualmente (*Editado*).

OS	**TS**	**Editado**
Deus, a vida, os grandes problemas, n**ã**o s**ã**o os **ph**ilos**ph**os que os resolvem, s**ã**o os pobres vivendo (...)	**d**eus, a vida, os grandes problemas, n**ä**o s**ä**o os **f**ilos**f**os que os resolvem, s**ä**o os pobres vivendo (...)	Deus, a vida, os grandes problemas, não são os **f**ilós**f**os que os resolvem, são os pobres vivendo (...)

Tabela II.3: Extracto de texto em português na ortografia original (OS), ortografia transcrita (TS), e texto editado.

O programa que faz esta transcrição ortográfica (próxima da fonoló-

gica) realiza a normalização de acentos, eliminação de letras maiúsculas, normalização de africadas, palatais, aspiradas, nasais e laterais palatais, geminadas e vogais. Além disso, também normaliza caracteres especiais de textos históricos em português, galego, espanhol e inglês. O seu objectivo é converter para a mesma ortografia qualquer que seja a ortografia original da língua estudada.

II.4 Corpora

Os corpora necessários para as nossas experiências para cada língua ou variante linguística diatópica foram concebidos tendo em conta que devem ser representativos, de tamanho suficiente, divididos em diferentes períodos históricos relevantes (no caso de cálculos de distância diacrónica) e escritos com uma ortografia tão próxima quanto possível dos textos originais (OS), a fim de medir a importância da ortografia como parâmetro de distância entre línguas ou variantes de línguas. Estes corpora são descritos em pormenor no Anexo V.

II.5 Pacotes de software e outros recursos

Para a preparação e programação de experiências e para permitir aos investigadores medir através de PLD as distâncias entre línguas, períodos históricos de qualquer língua, distâncias entre períodos históricos de duas ou mais línguas ou distâncias entre períodos históricos de variedades diatópicas, desenvolvemos uma arquitectura pipeline em Perl, que está disponível em código aberto sob licença GPL-v3 [1].

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

III. CAPÍTULO

Resumo da tese

As línguas têm sofrido alterações ao longo da sua história, tanto interna como externamente, em relação a outras línguas. A fim de medir esta evolução, foram propostas abordagens diferentes a partir de estudos filogenéticos, na dialectologia ou na área da aquisição de segunda língua. No domínio do processamento de línguas naturais, este papel tem cabido à identificação automática das línguas e à distância entre línguas.

O principal objectivo desta tese é propor e verificar uma metodologia baseada em corpus que quantifique automaticamente a distância sincrónica e diacrónica entre línguas e/ou variantes linguísticas. Para este fim, utilizámos técnicas já verificadas para identificar línguas, procurando as mais robustas que possam quantificar o quão próximo está um texto de um modelo de língua. Como objectivo secundário, investigámos o papel que a ortografia desempenha como factor de divergência e convergência entre as línguas.

A medição da distância que cumpriu os requisitos e que identificámos como a mais precisa e robusta baseia-se em *perplexity*. Para avaliar esta métrica, foi feito inicialmente uma experiência descrita no Anexo I, comparando textos contemporâneos em quarenta e quatro línguas europeias, com um corpus feito *ad hoc*. As distâncias foram calculadas para todos os pares de línguas possíveis apresentados numa rede. Finalmente, verificámos que estas distâncias estão correlacionadas com as publicadas pelos especialistas.

No passo seguinte descrito no Anexo II, usámos a mesma medida para quantificar a evolução da língua durante vários períodos históricos. Inicialmente foi aplicada ao português europeu, e mais tarde também ao espanhol e inglês europeus. Para isso, construímos o corpus de download gratuito *Carvalho* que contém textos históricos para cada período e para cada língua. Os textos que constituem *Carvalho* foram compilados a partir de diferentes fontes de corpus aberto contendo textos tão próximos quanto possível da sua ortografia original. Os resultados das experiências mostraram que a distância linguística diacrónica calculada, baseada em *perplexity*, detecta a evolução linguística explicada pelos historiadores das três línguas. A robustez do método deve ser salientada, uma vez que não foi efectuada qualquer supervisão ou revisão humana no que diz respeito aos cálculos da distância linguística.

Posteriormente, no Anexo III, fizemos uma comparação histórica por pares entre três línguas relacionadas: galego, português e espanhol. O objectivo foi tentar detectar se a evolução histórica destas línguas foi convergente ou divergente durante os vários períodos históricos. Para este fim, aplicámos a metodologia inicialmente à relação histórica entre português e espanhol desde a Idade Média até ao final do século XX. Os resultados mostram que, durante o período histórico em estudo, estas duas línguas aproximaram-se em certos sub-períodos e distanciaram-se noutros, e que a ortografia desempenha um papel importante no distanciamento destas línguas relacionadas. Seguidamente, aplicámos a metodologia à relação histórica que o galego teve com o português e o espanhol ao longo da sua história, tanto em OS como em TS. Os resultados quantitativos são contrastados com hipóteses extraídas de especialistas em linguística histórica, mostrando que o galego e o português são variedades da mesma língua na Idade Média e que o galego diverge e converge com o português e o espanhol desde o último período do século XIX. Neste processo, a ortografia também desempenha um papel relevante.

No Anexo IV, tentámos medir a distância entre as variedades diatópicas do português (português europeu e português do Brasil) e espanhol (espanhol europeu e espanhol da Argentina). Os resultados mostram distâncias muito próximas entre as variedades diatópicas do português e do espanhol, com ligeiras convergências/divergências desde meados do século XX até aos nossos dias.

No Anexo V, podemos ver o corpus histórico *Carvalho* que criámos *ad hoc* para as nossas experiências

Por fim, gostaríamos de salientar novamente que o método é totalmente automático e pode ser aplicado a qualquer língua ou grupo de línguas.

IV. CAPÍTULO

Hipóteses e Objectivos

IV.1 Hipóteses

O objectivo específico do nosso trabalho foi tentar confirmar empiricamente as hipóteses descritas mais abaixo, que estão intimamente relacionadas com as questões que colocamos como motivação no capítulo I.1:

- H1: A identificação automática de línguas e a distância entre línguas estão intimamente relacionadas.
- H2: A distância linguística pode ser medida automaticamente através de uma medida quantitativa robusta baseada em corpus.
- H3: A experimentação para o estudo da robustez do método de quantificação deve levar ao estabelecimento de uma metodologia para o cálculo empírico das distâncias entre corpora textuais de diferentes línguas.
- H4: A ortografia é um factor relevante na distância entre línguas, sendo capaz de aproximar ou afastar línguas, períodos históricos da mesma língua e mesmo variantes da mesma língua.
- H5: Para que os cálculos da distância entre línguas sejam comparáveis, o corpus das línguas deve ter tamanho suficiente e ser equilibrado em relação aos géneros textuais.

– H6: A medida deve ser capaz de calcular a distância entre textos actuais em diferentes línguas ou variantes (síncrona) e deve estar correlacionada com as avaliações qualitativas dos especialistas, podendo também gerar novas observações sobre hipóteses minoritárias ou controversas.

– H7: A medida deve ser capaz de calcular a distância entre diferentes períodos da mesma língua (distância diacrónica), a fim de determinar e medir a evolução histórica de uma língua. Os resultados devem estar correlacionados com as avaliações qualitativas dos especialistas.

– H8: A medida deve ser capaz de calcular a distância histórica entre línguas relacionadas, observando possíveis convergências e divergências históricas entre línguas ou variantes. Os resultados devem estar correlacionados com as avaliações qualitativas dos especialistas.

– H9: A medida deve ser capaz de medir a distância histórica entre variantes diatópicas da mesma língua, a fim de determinar possíveis convergências/divergências históricas. Como nos casos anteriores, a distância deve estar correlacionada com as avaliações qualitativas dos especialistas.

– H10: As línguas não convergem e divergem (evoluem) num sentido linear, mas podem historicamente convergir ou divergir de múltiplas formas.

IV.2 Objectivos

A fim de verificar as hipóteses anteriormente explicadas relacionadas com o cálculo automático da distância entre línguas, realizámos previamente um estudo bibliográfico e definimos uma série de objectivos a fim de orientar as experiências a realizar, que se detalham a seguir:

– O1: Encontrar uma métrica utilizada na identificação automática da língua que seja um candidato para uso na distância linguística.

– O2: Contrastar a métrica na tarefa de identificação linguística, permitindo assim que o cálculo das distâncias entre línguas, a partir do corpus, seja robusto e eficiente.

– O3: Definir uma metodologia flexível que utilize a métrica baseada em corpus, e que possa medir a distância entre línguas, entre períodos históricos da mesma língua, entre períodos históricos de duas ou mais línguas e entre períodos históricos de variantes diatópicas da mesma língua.

– O4: Incluir na metodologia o cálculo da distância entre as línguas na ortografia original e numa ortografia transcrita automaticamente através de uma transcrição ortográfica que converte o corpus multilingue numa ortografia comum próxima de uma ortografia fonológica.

– O5: Reunir corpora de tamanho suficiente, na ortografia original, comparáveis e equilibrados em diferentes línguas, tanto de textos actuais como históricos, a fim de realizar experiências relevantes. As línguas escolhidas, por interesse e conhecimento, foram o galego, o português, o espanhol e o inglês, bem como períodos históricos da variante brasileira do português e da variante argentina do espanhol.

– O6: Para quantificar e analisar a distância entre diferentes línguas, períodos históricos, variantes diatópicas, verificar, por um lado, se os resultados estão de acordo com as hipóteses dos linguistas e, por outro lado, verificar se existem novas observações que possam conduzir a novas hipóteses, ou verificar hipóteses minoritárias ou controversas.

V. CAPÍTULO

Discussão

V.1 Introdução

Este trabalho partiu da ideia de verificar se as técnicas utilizadas na identificação automática das línguas eram úteis para calcular as distâncias entre línguas.

A oportunidade surgiu no *shared tasks Discriminating between Similar Languages* (DSL) e *German Dialect Identification* (GDI), para o caso da identificação automática das variedades do alemão) no âmbito do workshop de referência na Identificação Automática de Línguas (Vardial 2017). Estas tarefas partilhadas servem: *"to evaluate how state-of-the-art systems perform in identifying similar languages and varieties, we decided to organize the Discriminating between Similar Languages"* [Zampieri et al., 2014].

Os sistemas apresentados competiram utilizando diferentes técnicas para identificar as variantes diatópicas do espanhol de Argentina e espanhol de Perú, português europeu e português do Brasil ou línguas *Ausbau* como bósnio, croata e sérvio, entre outras.

Embora o sistema baseado em *perplexity* não estivesse entre os primeiros classificados, estava muito próximo dos classificadores mais sofisticados que utilizam *SVM, Randon Forest, Naive Bayes*, etc. Este não foi o caso do único sistema baseado em redes neuronais, que não teve um bom desempenho: *"all teams except deepCybErNet obtained similar scores"*. O nosso sistema baseado em *perplexity*, ao contrário

do resto, era um sistema robusto que oferecia uma boa aproximação para o cálculo automático de distâncias.

A métrica *perplexity*, em particular, conseguiu na DSL task o nono lugar entre onze participantes (precisão: 0.903, enquanto o melhor obteve 0.927) e oitavo lugar entre os dez participantes na tarefa GTD (precisão: 0.630 frente à melhor com 0.680). Finalmente, é de notar que apenas dois sistemas na DSL 2016 task excederam a precisão de *perplexity*.

Como resultado destas descobertas, que foram discutidas na secção "Trabalhos futuros" do artigo Gamallo et al. [2017b], quisemos investigar se *perplexity* era também capaz de medir a distância entre línguas ou variantes diatópicas de forma diacrónica.

Assim, na nossa investigação de tese, criámos uma metodologia para o cálculo automático das distâncias entre línguas com base na métrica *perplexity* a partir de corpora de diferentes línguas. Verificámos também que papel desempenha a ortografia nessa distância.

Neste capítulo explicaremos e discutiremos, em pormenor, os principais resultados das diferentes experiências e as conclusões parciais. Para uma visão mais profunda dos resultados das diferentes experiências, recomendamos a leitura dos diferentes anexos descritos a seguir:

- No Anexo I descrevemos a criação de corpora síncronos em 44 línguas europeias em ortografia original e as experiências realizadas para identificar que a métrica *perplexity* é uma métrica robusta para calcular a distância entre as línguas.

- No Anexo II, é descrito a criação de um corpus diacrónico para português, espanhol e inglês em ortografias tão próximas quanto possível do original, e uma metodologia baseada em *perplexity* para o cálculo da história da distância entre períodos históricos destas línguas.

- No Anexo III, adaptámos a metodologia com base em *perplexity* a fim de medir a distância histórica entre as línguas *Ausbau* próximas ou muito próximas, como são o galego, o português e o espanhol. Neste caso, queríamos verificar se *perplexity* era capaz de verificar as hipóteses dos linguistas sobre as distâncias entre línguas ou variantes linguísticas tão próximas e efectuar outras observações.

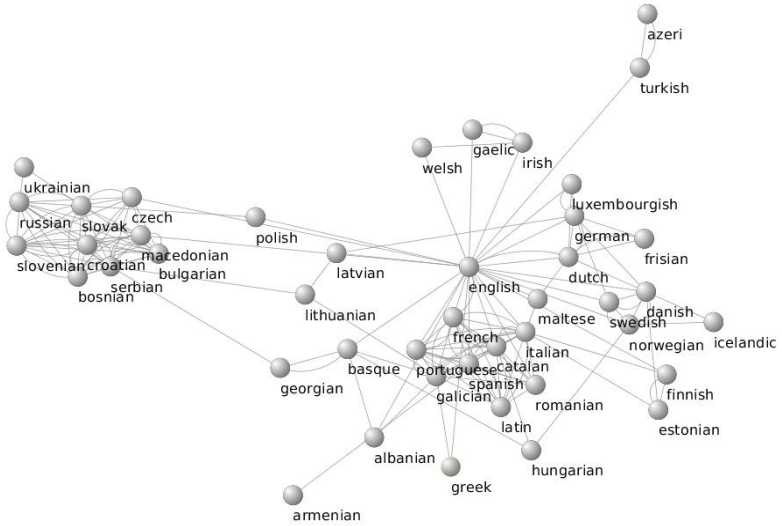


Figura V.1: Mapa de distâncias entre línguas na Europa, construído a partir de uma distância baseada em *perplexity* e o corpus web (*perp-web* strategy).

– Finalmente, no Anexo IV, aproximámos ainda mais o foco para medir a distância entre os pares de variantes diatópicas do português e do espanhol.

V.2 Medidas de distância síncronas entre línguas

No Anexo I, verificamos que *perplexity* é capaz de medir a distância síncrona entre quarenta e quatro línguas europeias. Assim, observámos que os modelos de língua básicos de n-gramas de caracteres extraídos a partir de corpus textuais podem ser utilizados, não só para classificar línguas ou variedades como na tarefa tradicional de identificação de línguas, mas também para medir a distância entre pares de línguas de uma forma geral. *Perplexity* provou ser uma métrica eficaz para comparar modelos, mas certamente não a única. Outras estratégias, tais como as diferentes técnicas utilizadas nos classificadores, também podem ser aplicadas à tarefa de definir uma medida da distância, trabalhando com n-gramas.

Com este objectivo, criámos dois corpora comparáveis diferentes a partir de ferramentas de web-crawling e de textos de Bíblias na ortografia

original de cada língua para ver o papel que a ortografia desempenha na distância entre as línguas. Embora estes corpora não sejam representativos em tamanho, como o indicado pelo Corpus de Helsinki [Rissanen et al., 1993a], servem para apontar tendências nas distâncias linguísticas. Além disso, cada um deles foi dividido em corpora do modelo de língua (MDL), com 120K palavras, e teste, com 40K palavras. Como resultado construímos um gráfico que representa o mapa actual de semelhanças e divergências entre as principais línguas da Europa e que pode ser visto na Figura V.1.

A partir destes corpora fizemos o cálculo da distância numa grafia transcrita automaticamente construída a partir do normalizador fonológico previamente comentado e que mostramos para alguns pares de línguas na Tabela V.1.

Estes resultados mostram as distâncias entre diferentes línguas (bósnio, croata, sérvio, checo, eslovaco, etc.), para além das principais línguas estudadas na nossa investigação: inglês, galego, português e espanhol. A partir destes resultados, iremos discutir alguns dos aspectos mais relevantes:

- Na distância com base em *perplexity* existe uma assimetria na distância entre pares de línguas, já explicada no capítulo II (*perplexity*).
- Se fizermos uma média aritmética entre as distâncias entre os pares de línguas observamos que o bósnio e o croata têm a mesma distância 5.90 que o espanhol e o galego 5.87.
- O português e o espanhol têm uma relação de proximidade 7.72 equivalente à relação entre o espanhol e o catalão 7.64.

Além disso, observámos no *site* que mostra a relação de distância para todos os pares das quarenta e quatro línguas europeias[1] o seguinte:

- As línguas românicas partilham a maioria das características com as outras línguas, o que as torna centrais.
- O inglês é uma língua central na relação entre as famílias românica, germânica, celta e eslava.

[1] `https://gramatica.usc.es/~gamallo/php/distance/index.php`

target language	closest languages	distance
bosnian	croatian	5
bosnian	slovene	8
bulgarian	macedonian	15
bulgarian	serbian	20
catalan	spanish	8
catalan	galician	10
croatian	bosnian	7
croatian	serbian	11
czech	slovak	9
czech	slovene	21
english	french	16
english	dutch	31
french	catalan	14
french	spanish	15
georgian	basque	37
georgian	serbian	47
irish	gaelic	9
irish	english	33
maltese	italian	24
maltese	english	25
portuguese	galician	6
portuguese	spanish	8
serbian	croatian	13
serbian	bosnian	13
spanish	galician	6
spanish	portuguese	8
swedish	danish	12
swedish	norwegian	13
turkish	azeri	20
turkish	english	46

Tabela V.1: Distância entre pares de línguas extraídos de *perp-web*. Na segunda coluna encontram-se as línguas mais próximas da primeira. Na terceira coluna, a distância entre ambas.

– Dentro das "línguas por distância" *Abstand*, o maltês, língua com origem no árabe, está ligado ao italiano e ao inglês, e o basco e o georgiano estão ligados, fornecendo uma observação relacionada com uma das hipóteses de parentesco do basco apontada por [Trask, 1995].

Como conclusões finais destes resultados, podemos concluir que:

– *Perplexity* é uma medida capaz de identificar línguas e medir a distância entre línguas, mostrando assim uma relação íntima entre os dois campos e confirmando a hipótese H1.

– *Perplexity* é uma medida robusta no que toca à aplicação do coeficiente de Spearman: *"We observe that there is a strong correlation (75.481) between the two methods based on perplexity,* perp-web *and* perp-bible *even though they are applied on two very different corpora."* [Gamallo et al., 2017a]. Isto confirma, portanto, a hipótese H2.

– Podemos criar uma metodologia flexível baseada em *perplexity* que nos permite calcular automaticamente as distâncias entre línguas. Esta metodologia irá evoluir para poder calcular outros aspectos relevantes na distância entre línguas ou variantes diatópicas. Isto confirma a hipótese H3.

– Finalmente, identificámos através de *perplexity* a existência de fortes ligações e interacções entre as línguas, as suas distâncias, de um ponto de vista sincrónico, confirmando a hipótese H6.

V.3 Medidas de distância diacrónicas entre línguas

V.3.1 Distância diacrónica intralinguística de galego, português, espanhol e inglês

No Anexo II e a parte do Anexo III relacionada com o galego, tivemos como principais objectivos melhorar a metodologia baseada em *perplexity* a fim de poder medir a distância entre períodos históricos da mesma língua e também para criar um corpus histórico chamado *Carvalho*, a fim de realizar experiências para calcular distâncias intralinguísticas entre os períodos históricos das quatro línguas em estudo.

Relativamente à metodologia, criámos a medida da distância entre as línguas PLD baseada em *perplexity*. Esta métrica identifica a evolução entre períodos históricos de qualquer língua, no nosso caso o galego, o português e o espanhol, que são línguas próximas, e o inglês, que pertence a outra família linguística. Esta metodologia pode ser aplicada a qualquer língua.

Para o efeito, foi concebido e criado *Carvalho*, que é um corpus histórico contendo apenas textos em ortografia original, a fim de ajudar a investigar o papel que a ortografia desempenha na distância entre períodos históricos de uma língua. As características gerais de *Carvalho* são explicadas no Anexo V.

Com esta metodologia flexível, que foi adaptada com PLD para medir distâncias entre períodos linguísticos históricos e o corpus histórico *Carvalho*, foram realizadas duas experiências em quatro línguas (galego, português, espanhol e inglês), um com os textos na ortografia original e o outro com textos transcritos por meio de um normalizador fonológico. Como resultado dos resultados das experiências, verificamos as seguintes hipóteses:

- Hipótese H3, pois a experimentação para estudar a robustez do método de quantificação levou a melhorias na metodologia para o cálculo empírico das distâncias entre períodos históricos em diferentes línguas.
- Hipótese H4, ao confirmar-se que a ortografia é um factor relevante na distância histórica entre línguas, aproximando ou distanciando, no nosso caso, períodos históricos da mesma língua.
- Hipótese H5, ao confirmar-se que, para ter resultados de distância entre línguas através de PLD deve haver um corpus histórico de tamanho suficiente e equilibrado como *Carvalho*.
- Hipótese H7, ao confirmar-se que PLD mede a distância entre diferentes períodos da mesma língua, em correlação com as avaliações qualitativas dos especialistas, e também produzindo alguns dados novos.

Embora no Anexo II possamos ver com mais profundidade os resultados e conclusões das experiências em relação à história destas línguas, resumimos a seguir as conclusões mais relevantes a que chegámos a partir dos resultados alcançados com a utilização de PLD:

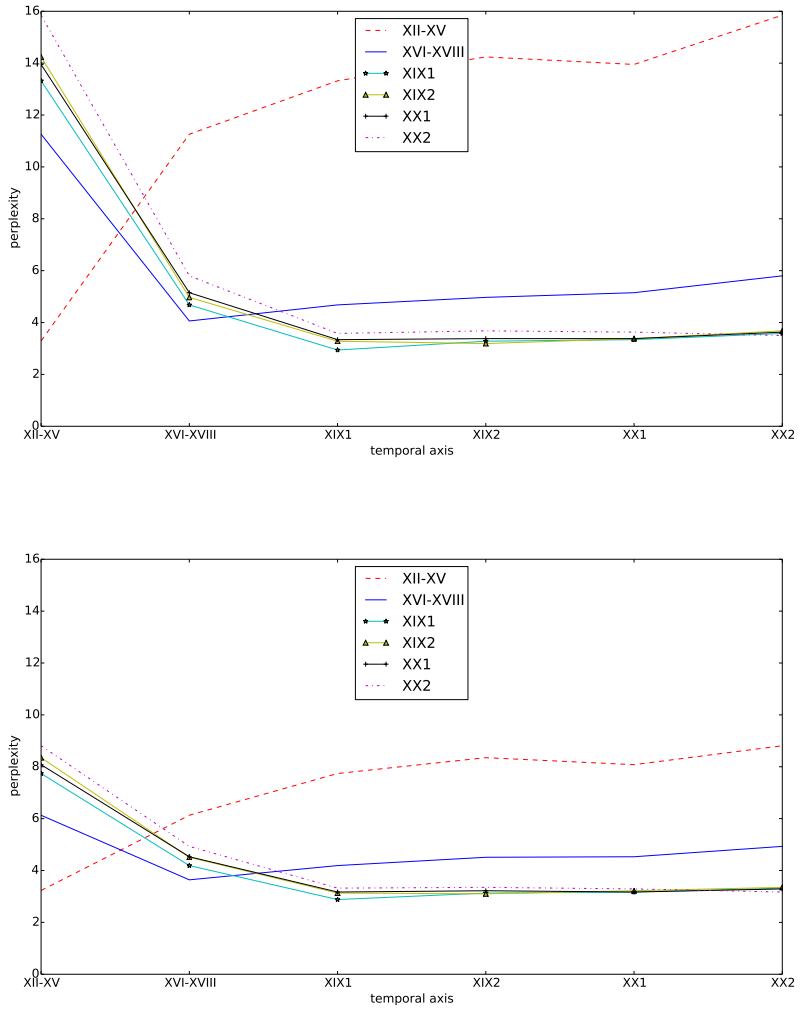


Figura V.2: Em (a) distância histórica do inglês na ortografia original. Em (b) a mesma comparação na ortografia transcrita.

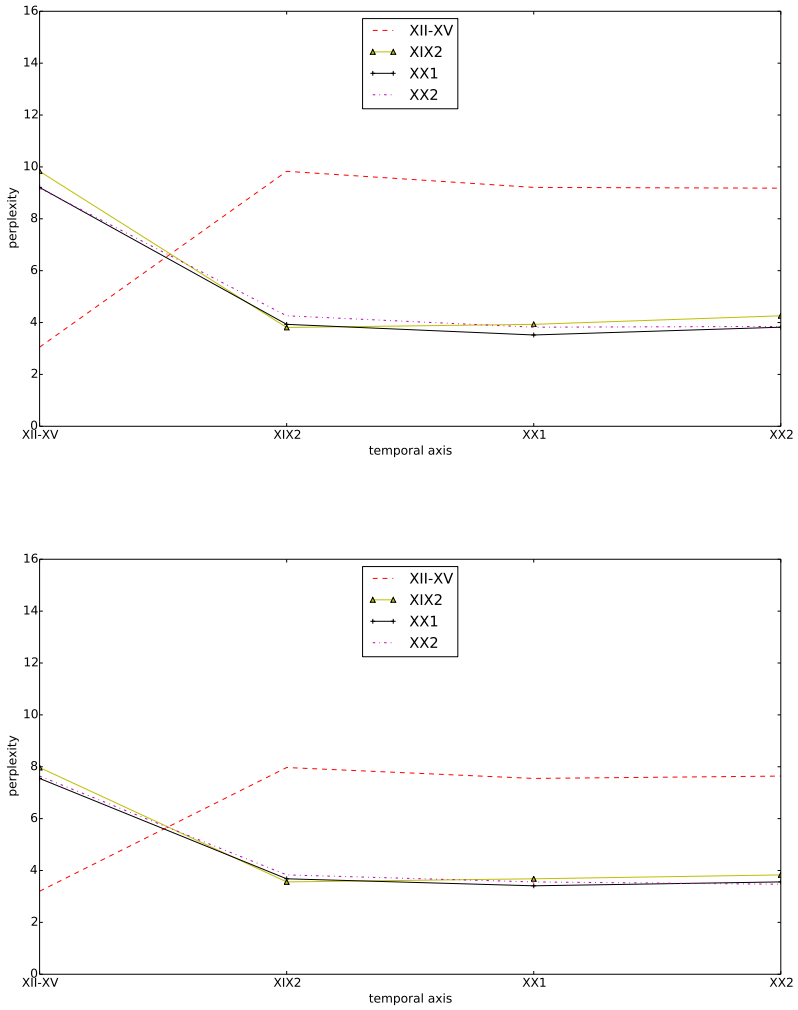


Figura V.3: Em (a) a distância histórica do galego na ortografia original. Em (b) a mesma comparação na ortografia transcrita.

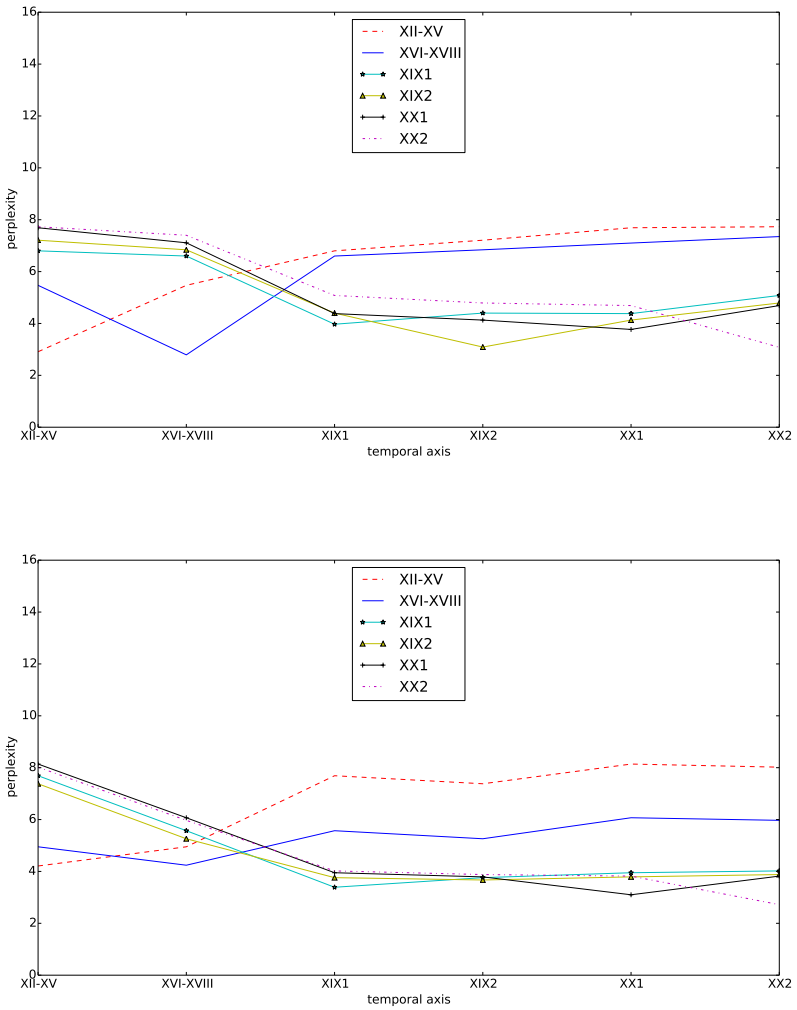


Figura V.4: Em (a) a distância histórica do português na ortografia original com PLD. Em (b) a distância histórica do espanhol na ortografia original com PLD.

– **O inglês e o galego têm dois períodos historicamente distantes: o período medieval e o resto**. Na Figura V.2, que representa a distância histórica do inglês, observamos que o inglês medieval (XII-XV) tem uma grande distância na ortografia original do resto dos períodos históricos, que mantêm uma importante homogeneidade. No que diz respeito ao galego, na Figura V.3 vemos que entre o período medieval (o período conhecido como galego-português) e o período em que a escrita em galego foi significativamente reavivada (XIX-2), existe uma distância significativa com uma PLD: 9.83. Nos outros períodos históricos, os textos estão muito próximos uns dos outros.

– **O português e o espanhol são historicamente mais homogéneos do que o inglês e o galego**. No caso do português e do espanhol, observamos na Figura V.4 um afastamento gradual entre os períodos medieval (XII-XVI) e renascentista (XVI-XVIII), que não estão muito distantes um do outro, e os outros períodos (XIX e XX), que também estão próximos. Além disso, notamos que o português é ligeiramente mais homogéneo do que o espanhol, uma vez que a PLD entre os períodos históricos mais longínquos (XII-XV e XX-2) é 7.73, frente a ($PLD = 8.02$) no caso do espanhol, diminuindo para 6.09 e 6.31 com uma ortografia normalizada. Ao considerar os resultados reportados em Gamallo et al. [2017b], esta pontuação está no mesmo intervalo que a distância entre as variedades diatópicas ou idiomas *Ausbau*, como o caso do bósnio-croata com uma *perplexity* = 5.90, e não é maior do que a distância entre línguas consideradas indubitavelmente diferentes mas estreitamente relacionadas (por exemplo, espanhol-português, *perplexity* = 7.74.

– **O galego é mais homogéneo em períodos históricos recentes do que o português, inglês e espanhol**. Em relação ao galego, observamos que é muito homogéneo desde o século XIX, praticamente sem diferenças relevantes na distância entre períodos históricos até ao final do século XX.

– **O galego tem uma distância não linear entre períodos históricos, ao contrário do inglês, português e espanhol**. Assim, embora o período de distanciamento entre a Idade Média e a segunda metade do século XIX atinja a sua distância máxima e, tendo em conta o que acontece com outras línguas, devesse

aumentar essa distância em períodos históricos subsequentes, tal não acontece. De facto, a PLD entre o período medieval XII-XV e XX está a aproximar-se progressivamente (9.21 e 9.18 em XX-1 e XX-2, respectivamente). Isto verifica a hipótese H10, uma vez que observamos que nem todas as línguas se distanciam em períodos históricos de uma forma linear.

– **A ortografia em inglês desempenha um papel muito importante na distância entre a Idade Média e os outros períodos**. Na Figura V.2(b) observamos que a distância entre o período medieval e o resto dos períodos ingleses diminui consideravelmente quando se utiliza a mesma ortografia entre todos os períodos históricos. Assim, se compararmos os períodos mais distantes (XII-XV e XX-2), a distância em inglês na ortografia original é mais de metade como resultado da normalização (PLD : 15.85 $\rightarrow$ 8.81), sendo equivalente à distância entre o período medieval e XX-2 na ortografia original em espanhol e português.

– **A ortografia no galego é um factor relevante de distância entre o período medieval e o resto dos períodos.** Na Figura V.2(b) vemos que a distância entre o período medieval (XII-XV) e os outros períodos se reduz: PLD: 7.97 em XIX-2, PLD: 7.55 em XX-1 e PLD: 7.64 em XX-2. No entanto, entre os diferentes períodos do medieval do galego, a ortografia não desempenha qualquer papel. Isto pode dever-se ao facto de o espanhol ter servido de base ao modelo ortográfico para os galegos desde o século XIX: *"Of course, Spanish was a model they could not ignore as it was the language they had learned to write in."* [Ramallo and Rei-Doval, 2015].

– **A ortografia em português e espanhol desempenha um papel importante na distância entre períodos históricos**. No caso do português e do espanhol, a ortografia aproxima os períodos medieval (XII-XV) e renascentista (XVI-XVIII) em oposição ao resto dos períodos (XIX e XX). Assim, a ortografia separa especialmente os períodos XVI-XVIII e XIX-1 ao descer a PLD de 6.60 (com a ortografia original) até 4.42 (com a ortografia transcrita). Note-se que, no último quartel do século XVIII, a língua portuguesa começou a desenvolver uma ortografia estreitamente relacionada com o latim e o grego (por exemplo, *philo-*

sofia em vez de *filosofia*). No caso do espanhol, a ortografia diferencia especialmente os períodos medieval, período renascentista e resto de períodos. Contudo, se utilizarmos a mesma ortografia em todos os períodos, os períodos medievais, renascentistas e os outros estão mais próximos. As alterações ortográficas da *Real Academia Española* no final do século XVIII influenciaram este distanciamento em ortografia original.

V.3.2 Distância diacrónica interlinguística entre línguas próximas: galego, português e espanhol

Para o cálculo da distância diacrónica interlinguística, realizámos experiências entre línguas próximas, com base no melhoramento da metodologia e na extensão do corpus *Carvalho* para incluir também o galego. As experiências realizadas confirmaram a hipótese H3, que indica que a experimentação para o estudo da robustez do método de quantificação deve levar ao estabelecimento de uma metodologia para o cálculo empírico das distâncias entre corpora textuais de diferentes línguas.

A inclusão do galego em *Carvalho* tem duas características únicas. Por um lado, como o galego, desde o século XVI até à segunda metade do século XIX, sofreu um abandono progressivo da sua produção escrita, não há um corpus suficiente para as nossas experiências em todos os períodos históricos. Por outro lado, os textos em galego em ortografia original que pertencem a dois corpus informatizados, TMILG (Tesouro Medieval Informatizado da Lingua Galega) [2] e TILG (Tesouro Informatizado da Lingua Galega) [3], não podem ser descarregados a partir do corpus *Carvalho* por causa do copyright.

Nas experiências que realizámos, aplicámos esta metodologia ao corpus *Carvalho*, para medir primeiro a distância diacrónica, em ortografia original (OS) e transcrita (TS), entre duas línguas próximas (português e espanhol). Em segundo lugar, medimos a distância nos períodos históricos disponíveis de galego com português e galego com espanhol. Este caso é muito interessante, como já discutimos no capítulo I.1, porque o galego é considerado por alguns linguistas como uma língua independente e por outros como uma *"variante dialectal*

[2] `https://ilg.usc.es/tmilg/`
[3] `https://ilg.usc.es/TILG/`

dentro do diasistema lingüístico do que hoxe en día se coñece como portugués" [Collazo, 2014].

As hipóteses mais relevantes que verificámos, para além das já indicadas nas nossas experiências, foram as seguintes:

- H8: A medida calcula a distância histórica entre línguas relacionadas, observando convergências/divergências históricas entre línguas ou variantes. Os resultados correlacionam-se com as avaliações qualitativas dos especialistas, e geram algumas novas observações sobre a relação entre as línguas.
- H10: As línguas não convergem e divergem (evoluem) num sentido linear, mas podem historicamente convergir ou divergir de múltiplas formas.

Apesar de, no Anexo III, podermos ver mais em detalhe as características das experiências e a discussão dos resultados e conclusões, abaixo detalhamos para cada par de línguas as conclusões mais relevantes. Estes pares são os seguintes: português-espanhol, galego-português e galego-espanhol.

Distância diacrónica interlinguística entre português e espanhol

Na Figura V.5 observamos a relação histórica de distância interlinguística entre português e espanhol em todos os períodos históricos, desde a Idade Média até ao presente. A partir daí, tiramos as seguintes conclusões principais:

- **Não há distanciamento linear entre as duas línguas**: A conclusão mais importante que pode ser tirada dos resultados é que a Hipótese H10: português e espanhol não se separam linearmente ao longo do eixo temporal. Pelo contrário, a sua evolução sofreu convergências e divergências não relacionadas com a ordem cronológica. Assim, vemos que português e espanhol convergem progressivamente da Idade Média até à segunda metade do século XIX com uma distância mínima de (PLD: 9.78) semelhante à que tinham na Idade Média. Ao contrário do que se poderia esperar, é a partir da primeira metade do século XX que divergem mais, atingindo uma distância máxima de (PLD : 13.2).

Mais tarde convergem novamente na segunda metade do século XX, tendo uma distância de PLD perto do que tinham na Idade Média.

– **A ortografia desempenha um papel importante na distância entre português e espanhol**. Os resultados das experiências realizados levaram-nos a concluir que a ortografia é um factor importante na distância entre o português e o espanhol, confirmando a Hipótese H4. Assim, podemos ver que entre os períodos de distância máxima e distância mínima, a ortografia desempenha um papel relevante uma vez que, se o português e o espanhol utilizarem a mesma ortografia, a distância entre as línguas diminui significativamente entre a ortografia original e a transcrita (PLD de 13.2 a 9.34 e PLD de 9.78 a PLD: 7.49) nos mesmos períodos.

  Observámos também que, com uma ortografia comum, a distância mínima entre o português e o espanhol, PLD: 7.49, é inferior à distância entre línguas próximas como o checo e o eslovaco com PLD: 8.46, ou espanhol e catalão com PLD: 8.63 também em ortografia transcrita segundo Gamallo et al. [2017a]. Em contraste, com ortografias diferenciadas (cálculos em ortografia original), a distância máxima entre português e espanhol (PLD: 13.2), é equivalente à distância entre francês e catalão de utilizarem uma mesma ortografia (PLD: 13.94) também segundo Gamallo et al. [2017a].

Distância diacrónica entre galego/português e galego/espanhol

Na Figura V.6 podemos observar a relação histórica da distância em PLD entre galego e português e a relação entre galego e espanhol na Figura V.7. As conclusões específicas tiradas desta relação de distância diacrónica são as seguintes:

– **A convergência do galego com o português não é linear desde o período de menor distância na Idade Média.** Os valores de PLD tanto em OS como em TS mostram que na Idade Média (período XII-XV) o galego e o português estavam muito próximos (PLD: 5.49 em OS e PLD: 4.84 en TS), equivalente à distância entre dois períodos históricos do espanhol

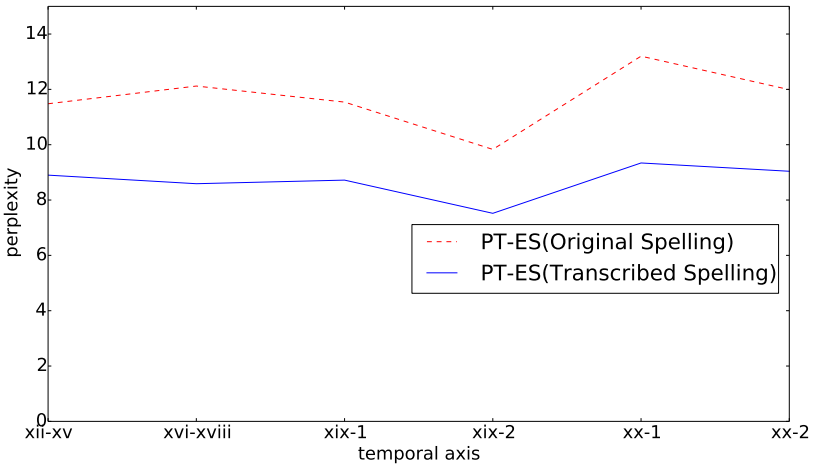


Figura V.5: Distância diacrónica interlinguística entre português e espanhol em OS e TS.

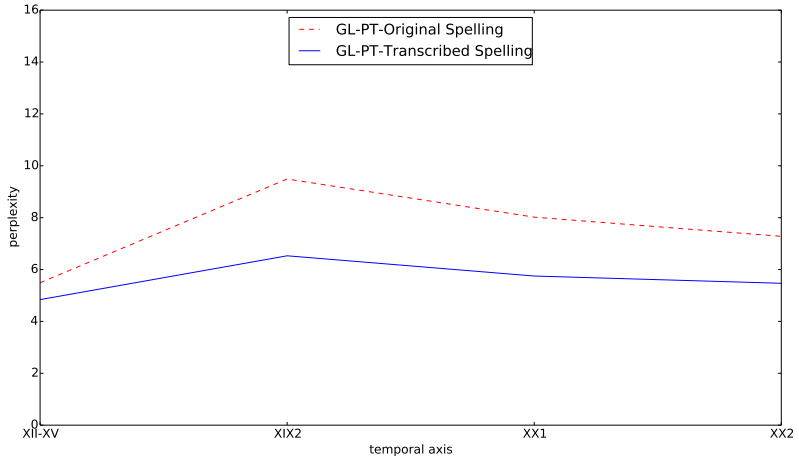


Figura V.6: Distância diacrónica interlinguística entre galego e português em OS e TS.

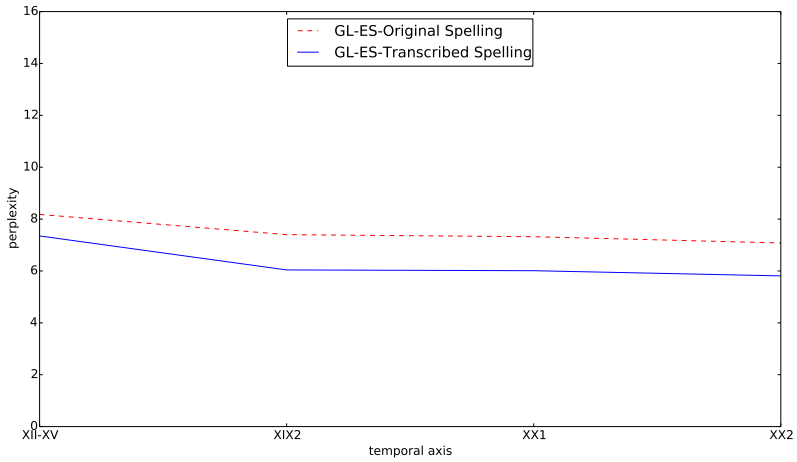


Figura V.7: Distância diacrónica interlinguística entre galego e espanhol em OS e TS.

(XII-XV e XVI-XVIII). Isto confirma que o galego e o português, no período medieval (conhecido como período galego-português) [Diez, 2008, Da Silva, 2018]), podem ser considerados como duas variedades da mesma língua.

No entanto, é precisamente quando a escrita em galego é retomada de forma significativa (XIX-2) que o galego e o português se distanciam mais, com uma PLD: 9.49 em OS e PLD: 6.53 em TS. Mais tarde, nos dois sub-períodos do século XX (XX-1 e XX-2), convergem progressivamente, com uma PLD: 8.02 OS e PLD: 5.75 (TS) na primeira metade do século XX e PLD: 7.28 OS e PLD: 5.47 (TS) na segunda metade do século XX.

Esta última distância, como se pode ver em Pichel et al. [2019b], é equivalente à distância que existe entre as variantes históricas do espanhol entre os períodos históricos XVI-XVIII e XIX-1 ou à distância em OS no período medieval com uma PLD: 5.49. Além disso, esta distância é ligeiramente inferior à distância entre o bósnio e o croata com uma PLD: 5.90.[4]

[4]Valor PLD calculado através da utilização do motor de busca `https://gramatica.usc.es/~gamallo/php/distance/`.

– **A convergência do galego com o espanhol é linear desde a segunda metade do século XIX**. O galego e o espanhol atingiram a distância máxima (PLD: 8.18 em OS e PLD: 7.35 em TS) na Idade Média (período XII-XV). Esta distância é inferior à distância actual entre o checo e o eslovaco com um PLD: 8.1 em (TS).[5]

  Contudo, a partir da segunda metade do século XIX, quando a escrita em galego foi retomada, ambas as línguas iniciaram um período de convergência progressiva, atingindo na segunda metade do século XX a sua distância mínima, com uma PLD: 7.08 em OS. Esta aproximação progressiva entre galego e espanhol pode ser explicada na tentativa de criar um padrão para o galego que rejeite os vulgarismos e hipergaleguismos para o registo culto e a progressiva escolaridade obrigatória em espanhol em que todos os galegos vivem, para além de ser o espanhol (castelhano) a língua oficial em toda a Espanha.

– **A ortografia é um importante factor de separação.** Uma das primeiras observações que vemos entre o galego, o português e o espanhol é que estas três variantes linguísticas românicas estão mais próximas se utilizarem uma ortografia comum entre elas. Contudo, se observarmos em detalhe a relação de cada par (galego-português, português-espanhol, galego-espanhol) a ortografia desempenha papéis historicamente diferentes na convergência ou divergência entre ambos.

  Assim, na relação entre galego e português, as diferentes ortografias medievais utilizadas por ambos, sendo muito próximas, não são um factor de divergência, mas são no caso da relação entre galego e espanhol.

  Posteriormente, o espanhol modificou significativamente a sua ortografia em relação a outras línguas românicas no final do século XVIII e início do século XIX, o que teve um enorme impacto nas diferentes ortografias utilizadas para o galego desde o período de XIX-2. Como afirma Monteagudo and Santamarina [1993]: *"in the early day of the Rexurdimento, written Galician ignored medieval and Portuguese spelling conventions, making*

[5]Valor PLD calculado através da utilização do motor de busca `https://gramatica.usc.es/~gamallo/php/distance/`

use of Spanish orthography, which was familiar to Galician writers".

Esta aproximação à ortografia em espanhol, pelo galego, é verificada na medida em que a distância entre galego e português (que utiliza uma ortografia diferente da espanhola) no período XIX-2 desce de PLD: 9.49 em OS para uma PLD: 6.53 em TS. No entanto, como esperado, a ortografia não é um factor relevante de separação no caso da relação entre galego e espanhol.

Mais tarde, no século XX, a ortografia continua a ser um factor relevante na separação entre galego e português. Assim, a PLD desce desde 8.02 em OS até 5.75 em TS na primeira metade do século XX e 7.28 em OS e 5.47 em TS na segunda metade do século XX. Pelo contrário, como o galego e o espanhol são escritos de forma muito semelhante, a ortografia não é um factor relevante para separar o galego do espanhol desde o período XIX-2. Como afirma Jones and Mooney [2017]: *"the use of Spanish orthographic conventions may help to distinguish Galician from Portuguese, to which it is linguistically more similar"*.

Por estas razões, podemos dizer que o galego se comporta como uma língua *Ausbau* em que a ortografia desempenha um papel importante na distância em relação ao português e ao espanhol. Isto é consistente com a afirmação de Kloss, Heinz [1967]: *"The process of* ausbau*, and the creation of* abstand*, involves establishing linguistic autonomy from related languages by* reshaping *the visual representation of the language while the linguistic structure of the language(s) remains, in principle, unchanged"*.

– **O galego aproxima-se do espanhol e do português desde o início do século XX**. Isto pode ser devido ao facto de, desde o período XX-1, o galego ter tido uma tendência para ser construído a partir de materiais diferentes provenientes do espanhol e do português, em maior ou menor medida. Assim, em Álvarez and Monteagudo [2005]: *"É dicir, na construción dun estándar de características semellantes ás do español e do portugués, asumindo a xerarquización social que a estandarización trae consigo"*. Por outras palavras, a estandardização do galego aproximou-o do espanhol e do português ao mesmo tempo.

– **O galego pode ser galego-português ou galego-espanhol tendo a ortografia um papel relevante nessa relação**. Fi-

nalmente, como conclusão final, observamos que a distância entre o galego e as outras duas línguas em TS no último subperíodo do século XX (XX-2) é equivalente à distância entre o bósnio e o croata, historicamente consideradas variantes da mesma língua ou línguas independentes [Gamallo et al., 2017a]. Além disso, o galego pode ser visto como uma variante muito próxima do espanhol se usar uma ortografia original muito próxima do espanhol e uma variante muito próxima do português se usar uma ortografia próxima do português, sendo a distância PLD em TS do galego com o espanhol e o português muito próximas. No primeiro caso, com PLD: 5.81 e no segundo com PLD: 5.47. Isto está de acordo com a hipótese controversa de Carvalho [1979], Calero [1981]: *"O galego ou é galego-português ou é galego-castelhano. Ou somos umha forma do sistema ocidental ou somos umha forma do sistema central. Nom há outra alternativa".*

V.3.3 Distância diacrónica interlinguística entre variedades diatópicas de português e espanhol

No Anexo IV o cálculo baseado em *perplexity* para a distância entre idiomas (PLD) evolui para medir a distância histórica entre as variantes diatópicas de duas línguas [Pichel et al., 2020b], confirmando a Hipótese H9.

Para este novo objectivo, adaptámos a metodologia para medir a distância histórica com PLD em variedades diatópicas de duas línguas: português europeu/português do Brasil e espanhol europeu/espanhol da Argentina. Além disso, criámos, por um lado, em *Carvalho* um subcorpus para os períodos da segunda metade do século XX e do século XXI para o português do Brasil e o espanhol da Argentina, e, por outro lado, estendemos o corpus de português europeu e de espanhol europeu até o século XXI. Ambos os sub-corpus estão também em ortografia original.

A seguir veremos as principais conclusões a que chegámos a partir do cálculo da distância diacrónica entre pares de variedades diatópicas do português e do espanhol em dois períodos históricos que podemos observar na Figura V.8 e na Figura V.9:

- **As variedades diatópicas do português e do espanhol estão distanciadas de forma muito semelhante.** Assim, ob-

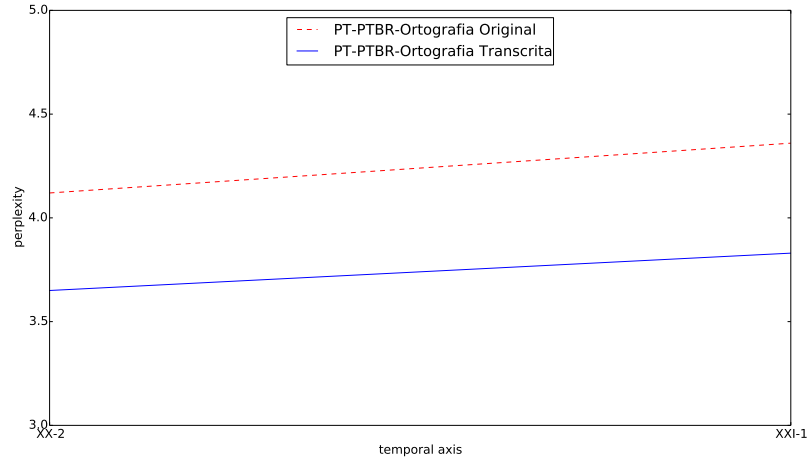


Figura V.8: Distância diacrónica interlinguística entre o português europeu e o português do Brasil em OS e TS.

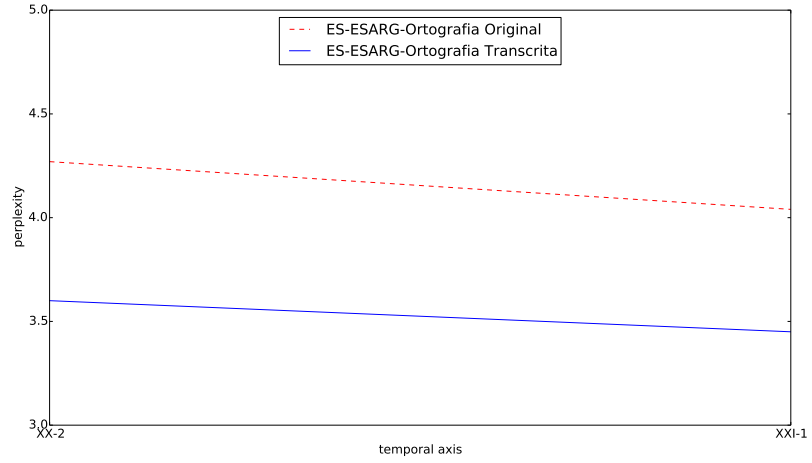


Figura V.9: Distância interlinguística entre espanhol europeu e espanhol da Argentina em OS e TS.

servamos que a distância entre as variedades diatópicas de português e espanhol é menor em PLD a 5, tanto em OS como em TS. Com base nos resultados em Gamallo et al. [2016], vemos que línguas consideradas línguas próximas por alguns ou variantes diatópicas por outros, como o bósnio e o croata, têm em TS uma distância bem maior, com PLD: 5.90.

- **O português do Brasil e o português europeu estão a afastar-se ligeiramente no século XXI**. O AO'90 foi apresentado como um factor de aproximação entre o português europeu e o português do Brasil. Talvez esta ligeira divergência mostre que Portugal e o Brasil funcionam de facto como sistemas culturais distintos, apesar do AO'90, que também tem sido aplicado lentamente e com muita resistência.

- **O espanhol da Argentina e o espanhol europeu estão a aproximar-se ligeiramente no século XXI**. Pelo contrário, no caso do espanhol europeu e do espanhol argentino, vemos que existe uma ligeira convergência no mesmo período (XXI-1), talvez devido aos esforços de coordenação entre as diferentes academias de língua espanhola e à existência de um maior intercâmbio material entre os sistemas culturais de Espanha e da Argentina.

- **A ortografia não desempenha um papel relevante na convergência ou divergência das variedades diatópicas de português e espanhol**. Quando calculamos o PLD em TS, observamos que diminui ligeiramente, mantendo a mesma tendência que em OS. Isto pode dever-se ao facto de estas variedades diatópicas serem escritas em ortografias próximas (português) ou ortografia indistinguível (espanhol).

Parte 2

Conclusões

VI. CAPÍTULO

Conclusões, contribuições e trabalho futuro

VI.1 Conclusões

Nas investigações realizadas para a nossa tese, definimos uma metodologia para o cálculo automático das distâncias entre línguas com base na métrica *perplexity* a partir de um corpus de diferentes línguas ou períodos de tempo. Verificámos também que papel desempenha a ortografia nessa distância.

Para o efeito, realizámos diferentes experiências, publicadas em revistas e conferências relevantes, que podem ser consultadas nos diferentes anexos.

Assim, no Anexo I, descrevemos as experiências realizadas em corpus actuais de 44 línguas europeias em ortografia original, que verificam que a métrica *perplexity* é uma métrica robusta para calcular a distância entre línguas.

No Anexo II, descreve-se a criação de um corpus diacrónico chamado *Carvalho* para português, espanhol e inglês em ortografias tão próximas quanto possível do original, e uma metodologia baseada em *perplexity* para o cálculo da distância entre os períodos históricos destas línguas.

No Anexo III, adaptámos a metodologia com base em *perplexity* a fim de medir a distância histórica entre duas ou mais línguas *Ausbau*, tal como o galego, o português e o espanhol. Neste caso, queríamos

verificar se *perplexity* era capaz de verificar as hipóteses dos linguistas sobre as distâncias entre línguas ou variantes linguísticas tão próximas e fazer outras observações.

Finalmente no Anexo IV, aproximámos ainda mais o foco para medir a distância entre os pares de variantes diatópicas de português e espanhol.

De todas as experiências realizadas e da discussão anterior podemos concluir que cumprimos todos os objectivos e verificámos todas as hipóteses levantadas no nosso trabalho, que detalhamos a seguir:

- A identificação automática das línguas e a distância entre as línguas estão intimamente relacionadas, o que corresponde às hipóteses H1 e H2.
- Definimos uma metodologia para calcular as distâncias entre línguas com *perplexity* que permite o cálculo das distâncias entre períodos históricos de línguas diferentes, línguas historicamente relacionadas e variantes diatópicas históricas das línguas, reconfirmando as hipóteses H1 e H2, além da hipótese H3.
- A ortografia é um factor que também actua sobre a distância entre as línguas e pode também aproximar ou afastar períodos históricos da mesma língua e mesmo variantes da mesma língua. Tal confirma a hipótese H4.
- Realizámos experiências entre línguas com corpora de tamanho suficiente, equilibrados entre ficção e não-ficção, para serem representativos. Isto permitiu-nos fazer comparações de distâncias entre línguas, períodos históricos de línguas ou variantes diatópicas de línguas, o que confirma a hipótese H5.
- *Perplexity* mede a distância sincrónica entre línguas, correlacionadas com as avaliações qualitativas dos especialistas. Além disso, é capaz de gerar novas observações sobre hipóteses minoritárias ou controversas. Tal confirma a hipótese H6.
- *Perplexity* mede a distância entre diferentes períodos históricos de uma mesma língua. Também se correlaciona com as avaliações qualitativas dos especialistas, e gera novas observações sobre hipóteses minoritárias ou controversas. Tal confirma a hipótese H7.

– *Perplexity* mede a distância histórica entre línguas relacionadas para determinar possíveis convergências/divergências históricas entre línguas ou variantes. Também se correlaciona com as avaliações qualitativas dos especialistas, gerando novas observações sobre hipóteses minoritárias ou controversas. Tal confirma a hipótese H8.

– *Perplexity* mede a distância histórica entre variantes diatópicas da mesma língua, assinalando convergências/divergências históricas entre elas. Tal confirma a hipótese H9.

– *Perplexity* mostra que não existe uma relação linear de convergência ou divergência entre línguas ou períodos históricos de línguas, mas sim períodos não lineares de convergência ou divergência histórica. Tal confirma a hipótese H10.

VI.2 Contribuições

Neste trabalho de tese fizemos contribuições em três campos: métodos de cálculo de distância entre línguas e variantes, compilação e disponibilização à comunidade científica de corpora sincrónicos e históricos e avaliação da evolução das línguas com base em dados.

VI.2.1 Distância entre línguas e variantes

O nosso maior contributo neste campo é ter proposto um método geral para quantificar automaticamente a distância actual entre línguas (sincrónico), entre períodos históricos (diacrónico) de uma ou mais línguas e entre variantes diatópicas de diferentes línguas.

O método chamado PLD é baseado em corpora escritos e utiliza a conhecida *perplexity* para quantificar essa distância. A sua avaliação qualitativa foi satisfatória (ver secção sobre *Discussão* no capítulo V).

O código para efectuar cálculos de distância a partir destes corpora escritos foi disponibilizado livremente e está ao serviço da comunidade de investigação para aprofundar dois aspectos da linguística histórica:

– O nosso método pode ser um *baseline* interessante para comparar os resultados com outras medidas. Entre elas destacamos

o aparecimento de novas propostas de medidas de distância em Gamallo et al. [2020] como a *ALD*, uma média entre quatro métricas diferentes, incluindo *PLD*, divergência Kullback-Leibler, Ranking-based distance e Cosine Distance Average. *ALD* foi aplicada para medir a distância entre línguas isoladas na Europa.

– Contribuição de novas observações provenientes de corpora históricos que podem produzir novos dados para novas hipóteses, para além de verificar hipóteses minoritárias ou mesmo hipóteses controversas.

VI.2.2 Compilação de corpora sincrónicos e históricos

Outro contributo importante é o corpus histórico em ortografia original *Carvalho*, que contém línguas diferentes e é descarregável, excepto para o galego por razões de direitos de autor.

O corpus *Carvalho* contém: Carvalho-PT-PT (português europeu) e Carvalho-PT-BR (português do Brasil) para o português; Carvalho-ES-ES (espanhol europeu) e Carvalho-ES-AR (espanhol da Argentina) para o espanhol; Carvalho-EN-UK (inglês britânico) para o inglês e Carvalho-GL para o galego.

O corpus *Carvalho* está dividido em período medieval (XII-XV), período renascentista (XVI-XVIII) e subdivididos em períodos de 50 anos os séculos XIX e XX para todos os idiomas (português, espanhol e inglês), excepto português do Brasil e espanhol da Argentina com dois períodos, segunda metade do século XX e século XXI. No caso do galego, está dividido em períodos em que existe corpus suficiente para as nossas experiências: XII-XV, XIX-2, XX-1 e XX-2. Cada um destes períodos contém 1.5M de palavras, com uma representatividade de 50% de textos de ficção e 50% de textos de não-ficção.

VI.2.3 Avaliação da evolução das línguas

A tese faz também contribuições relevantes baseadas em dados sobre questões relacionadas com a linguística histórica, que são detalhadas de seguida:

– A não-linearidade na convergência ou divergência entre línguas. Assim, através das nossas investigações, verificámos que quando

duas línguas ou variantes linguísticas divergem ou convergem, esta relação não é necessariamente mantida de forma linear ao longo do tempo, mas que podem existir períodos diferentes de divergência e convergência que não são necessariamente lineares. Embora esta avaliação não possa ser conclusiva, fornece dados adicionais para estudos nesta área.

– Por outro lado, verificámos que a ortografia é um factor relevante na distância entre as línguas, ajudando por vezes a aproximá-las e por vezes a separá-las. Isto também se aplica a períodos históricos da mesma língua ou a variantes diatópicas da língua.

VI.2.4 Aplicação do método a outros campos

Finalmente, este método pode ser aplicado a outras áreas. Assim, vimos que tem sido aplicada a campos relacionados, tais como a tradução automática [Barrault et al., 2019], a sociolinguística [Chavula and Suleman, 2020] ou a sociologia [Anna and Weller, 2020].

VI.3 Trabalho futuro

Em relação ao trabalho futuro, acreditamos que esta investigação sobre métodos de cálculo da distância entre línguas com PLD pode abrir novas linhas de investigação no campo da linguística computacional, sociolinguística ou linguística histórica.

O nosso primeiro objectivo a curto prazo é completar e melhorar a medida PLD, complementando-a com diferentes modelos de língua (por exemplo, n-gramas calculados desde palavras linguísticas relevantes) e regras fonológicas mais complexas que modifiquem a ortografia. Também se pode estudar o uso de *embeddings* (contextualizados).

A fiabilidade da medida PLD também precisaria de ser estudada mais aprofundadamente e a sua relação com a qualidade/quantidade dos dados recolhidos. Poderia ser feita uma avaliação semelhante à de *cross-validation* (*leave-one-out* por exemplo), calculando os desvios e melhorando a fiabilidade da medida.

Finalmente, queremos abrir uma linha de investigação com novas métricas que alarguem as observações que a PLD permite, tais como

a divergência de *Kullback-Leibler* (KLD). Esta divergência Kullback-Leibler [Kullback and Leibler, 1951] mede o quanto duas distribuições diferem. Assim, podemos usá-la para medir quão diferente é uma distribuição de probabilidade (por exemplo, o modelo de língua da língua de origem) de uma distribuição de probabilidade de referência (por exemplo, o modelo de língua da língua de destino).

Outros objectivos a médio prazo estão relacionados com diferentes campos e são descritos a seguir:

- No domínio da tradução automática, gostaríamos de investigar a relação entre a distância entre línguas e a estimativa da qualidade da tradução automática (*quality estimation*) [Specia et al., 2018, Han et al., 2013].
- No domínio da identificação automática da língua, utilizaremos PLD juntamente com outras métricas, tais como ALD [Gamallo et al., 2020] para melhorar a precisão nas línguas *Ausbau* e variedades estreitamente relacionadas.
- No campo da linguística histórica, com base nestes resultados, gostaríamos, por um lado, de utilizar PLD para calcular a distância diacrónica em outras línguas e variedades de línguas não europeias. Por outro lado, gostaríamos de aplicar a métrica PLD a casos semelhantes ao galego, a fim de fornecer novas observações na classificação filogenética das línguas (línguas independentes ou variantes da mesma língua). Entre estes casos podemos destacar ao sérvio, bósnio e croata; flamengo e neerlandês, moldavo e romeno, etc. Também queremos ver o papel que tem a ortografia nestes casos.
- No domínio da sociolinguística, queremos aplicar PLD às mesmas línguas e variedades diatópicas em estudo nesta tese, mas com um corpus de linguagem popular e registos diversos. Estes corpora, que irão aumentar o corpus *Carvalho*, podem ser construídos a partir de redes sociais (p.e.: Twitter ou Instagram) e comentários em plataformas digitais (p.e: Tripadvisor, AirBnB, Booking, etc.).
- Finalmente, no campo da linguística histórica aplicada ao galego, devido à falta de corpora necessários para as nossas experiências entre os séculos XVI e XIX-1, queremos investigar se através de

incorporar nesses períodos os corpora em português ou em espanhol, existe uma linearidade histórica no galego entre a Idade Média e a segunda metade do século XIX-2 com uma ou outra língua. As conclusões destas experiências poderiam fornecer novos dados para a análise da evolução da língua galega e mesmo oferecer novos dados para avaliar a controversa hipótese do linguista português Rodrigues Lapa: *"Nada mais resta senão admitir que, sendo o português literário actual a forma que teria o galego se o não tivessem desviado do caminho próprio, este aceite uma língua que lhe é brindada numa salva de prata."* [Lapa, 1973].

Parte 3

Anexos

Artigos publicados e descrição dos corpora

I Medidas de distâncias entre línguas: comparação e avaliação em corpus sincrónicos

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

From Language Identification to Language Distance

Pablo Gamallo

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
University of Santiago de Compostela, Galiza
pablo.gamallo@usc.es

José Ramom Pichel

Imaxin|Software, Galiza
jramompichel@imaxin.com

Iñaki Alegria

IXA Nlp Group, UPV/EHU, Basque Country
i.alegria@ehu.eus

Abstract

In this paper, we define two quantitative distances to measure how far apart two languages are. The distance measure that we have identified as more accurate is based on the *perplexity* of n-gram models extracted from text corpora. An experiment to compare forty-four European languages has been performed. For this purpose, we computed the distances for all the possible language pairs and built a network whose nodes are languages and edges are distances. The network we have built on the basis of linguistic distances represents the current map of similarities and divergences among the main languages of Europe.

Keywords: Language Distance, N-Gram Models, Perplexity, Corpus-Based Linguistics, Natural Language Processing, Language Identification

1. Introduction

In this article, we deal with the concept of *language distance*, which refers to how different one language or variety is from another. Even though there is

*Corresponding author
Email address: pablo.gamallo@usc.es (Pablo Gamallo)

no well-established measure to quantify the distance between two languages [1], some specific linguistic work relies heavily on the use of this concept, namely in phylogenetic studies within historical linguistics [2, 3], in dialectology [4], or in studies about learning additional languages within the field of second language acquisition [5]. The prevailing view, however, is that linguistic distance cannot be measured since two languages may differ in many linguistic aspects, e.g. phonetics, written form, morphology, syntax, and so on. Quantifying all these aspects by reducing them to a single distance score is a difficult task which is far from being fulfilled or at least appropriately addressed, perhaps as it has not yet been a priority in the field of Natural language Processing (NLP).

The concept of language distance seems to be related to the process of language identification. In fact, language distance and language identification are two sides of the same coin. The more difficult the identification of differences between two languages is, the shorter the distance between them. Language identification was one of the first natural language processing problems for which a statistical approach was used and it is now considered as an (almost) solved problem except for complex tasks such as similar variety discrimination or short text classification. The best language identification systems are based on n-gram models of characters extracted from text corpora.

The main objective of our work is to define a linguistic distance between two languages by considering character-based n-gram models, in a similar way to traditional language identification strategies. Character n-grams not only encode lexical and morphological information, but also phonological features since 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)[1]. This 7-gram might

[1]The stress accent (e.g. *ión*) has been removed to simplify language encoding.

be considered as an instance of the generic pattern *"noun-prep-noun"* since *ion* is a noun suffix and *de* a very frequent preposition (translated as *of* or *from* in English) introducing prepositional phrases. So, models built from corpora and based on long character n-grams are complex linguist artifacts provided with linguistic features at different levels, including phonological, morphological, lexical, and even (very basic) syntactic information. We must point out that our study is aimed at comparing not a continuum of dialectal varieties, but well-defined written standards. These are *standardized varieties* including not only standards that are distinctly separate from any other language (*Abstand* languages or languages by distance), but also cultural and political constructs known as *Ausbau* (elaboration) languages. The latter are called *elaboration* languages because their distance to each other has been elaborated historically even though they are mutually intelligible [6].

In order to compute language distance, two specific metrics will be proposed. Firstly, we measure the *perplexity* of a n-gram model on a test text. Perplexity is defined as the inverse probability of the test text given the model. Most of the best systems for language identification use probability-based metrics with n-gram models. Secondly, we also use a *ranked-based* method that ranks n-grams according to frequency. N-grams with highest frequencies are retained and the rest are discarded. This gives us pruned character n-gram models which are used for defining the distance between languages. These two metrics were chosen because they represent two well-known families of language identification strategies: those that classify languages according to n-gram probabilities, and those relying on ranked lists of n-grams.

The two metrics will be tested in different experimental setups. We start by testing their performance in a language identification task, and then, we use them to measure the distance between European languages. The latter experiment will allow us to draw a diagram showing the linguistic distance among most European languages. The diagram will be derived from a 2D-matrix of languages and their relationship to each other.

The remainder of the article is organized as follows. Section 2 introduces

the works using the notion of language distance in historical linguistics, as well as the main methods used in language identification. Following this, Section 3 defines two distance measures based on n-grams of characters. Two experiments are reported in Section 4: the first one uses our distance measures for the difficult task of identifying similar languages and varieties, and the second one applies them for building a network of the main languages of Europe. Finally, conclusions are drawn in Section 5.

2. Related Work

Linguistic distance has been measured and defined from different perspectives using different methods. Many of them compare lists of words to find phylogenetic links, while there are few corpus-based approaches from a synchronic point of view.

2.1. Phylogenetics and Lexicostatistics

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 [3]. In order to automatically build such a phylogenetic tree, many researchers make use of what they call *lexicostatistics*, which is an approach of comparative linguistics that involves quantitative comparison of lexical cognates [7, 8, 9, 2, 3]. More precisely, these computational studies are based on cross-lingual word lists (e.g. Swadesh list [10] or ASJP database [11]) to measure distances from the percentage of shared cognates, which are words with a common historical origin. Given a standardized word list, the distance between a pair of languages is defined by considering the cognates they share. More precisely, as described by Wichmann [12], the basic lexicostatistical technique defined by Swadesh consists of the following steps: (1) a cross-lingual word list is created, (2) cognates are identified, (3) the percentage of shared cognates is computed for each pair of languages to produce a pairwise inter-language distance matrix, and (5) the lexical distances

are transformed into separation times: the more distant two languages are, the more time is needed to find a common ancestor. This last step is one of the main objectives in *glottochronology*.

Other related work proposed an automated method which uses Levenshtein distances among words in a cross-lingual list [2]. Unlike lexicostatistical strategy, this method does not aim to distinguish cognates from non-cognates. The global distance between two languages is computed by considering a normalized Levenshtein distance between words and then finding the average of all such distances contained in the list.

A slightly different strategy is based on more standard supervised machine learning approaches. The input to a phylogenetic analysis is generally a data matrix, where the rows represent the given languages, and the columns represent different linguistic features (also called *characters*) by which the languages are described [13]. Features need not be lexical; they can also be syntactic and phonological. Some of these approaches use Bayesian inference to classify new data on the basis of the language models coded in the data matrix [14].

Computational methods taken from computational phylogenetics have been applied not only on lists of lexical units but also on phonetic data [7]. And they have been used to explore the origins of Indo-European languages [15, 16], Austronesian language groups [17, 16], Bantu languages [18], as well as the subfamily of Dravidian languages [19].

In sum, computational phylogenetics use cross-lingual lists to compute string or/and phonological distances among words, which are in turn used to measure distances among languages. These distances are then submitted to tree-building or clustering algorithms for the purpose of generating phylogenetic trees or clusters showing historical relationships among languages [20]. An excellent survey explaining the different types of phylogenetic strategies is reported in Wichmann [12].

2.2. Distributional-Based Approaches

To compare different languages, very recent approaches construct complex language models not from word lists, but from large cross-lingual and parallel corpora [21, 22, 23]. 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. The works by Liu and Cong [21] and [22] were performed on a relatively small number of languages. More precisely, Liu and Cong [21] compared fourteen languages and Gao et al. [22] studied merely six languages. By contrast, Asgari and Mofrad [23] performed language comparison on fifty natural languages from different linguistic families, including Indo-European (Germanic, Romance, Slavic, Indo-Iranian), Austronesian, Sino-Tibetan, Altaic, Uralic, and Afro-Asiatic. The authors built the language models for each language from a collection of sentence-aligned parallel corpora. The corpora used is the Bible Translations Project described in Christodoulopoulos et al. [24]. The results of this large-scale language comparison are, however, not very promising, since the similarity measure gives rise to several counter-intuitive findings. For instance, Norwegian and Hebrew, belonging to two different language families (Indo-European and Semitic), are wrongly grouped together. The system also separates in different clusters the two main languages of the Finno-Permian family: Estonian is clustered with Arabic and Korean while Finish is grouped with Icelandic, an Indo-European language.

Another limitation of the distributional-based approaches is that they require parallel corpora to build the models to be compared, and this kind of data is not easily available for many pairs of languages.

2.3. Language Identification

Two specific tasks of language identification have attracted a lot of research attention in recent years, namely discriminating among closely related languages [25] and language detection on noisy short texts such as tweets [26, 27].

The Discriminating between Similar Languages (DSL) workshop [28, 29, 25] is a shared task where participants are asked to train systems to discriminate between similar languages, language varieties, and dialects. In the three editions organized so far, most of the best systems were based on models built with high-order character n-grams ($>= 5$) using traditional supervised learning methods such as SVM, logistic regression, or Bayesian classifiers. By contrast, deep learning approaches based on neural algorithms did not perform very well.

TweetLID [30, 27] is another shared task aimed at comparing language detection systems tested on tweets written in the 5 most spoken languages from the Iberian Peninsula (Basque, Catalan, Galician/Portuguese, and Spanish), and English. Some of the target languages are closely related: e.g. Spanish and Galician or Spanish and Catalan, and there are even varieties of the same language in two different spelling rules, e.g. Portuguese and Galician. So the systems are tested, not only on noisy short texts (tweets), but also on a set of texts written in very similar languages/varieties. As in DSL Shared Task, the best systems were also based on character n-grams and traditional classifiers.

In addition to n-gram models, other traditional approach with satisfactory results in language identification is that relying on ranked n-grams [31, 32]. This approach relies on the observation that the most frequent n-grams are almost always highly correlated with the language. The rank-based measure sums up the differences in rankings of the n-grams found in the test data as compared to the training data. Rank-based systems seem to be stable across different domains and perform reasonably well on out-of-domain tests [26, 33]. Ranking-based methods have also been applied successfully in machine learning to order classification algorithms [34].

Given that corpus-based n-grams are still the best way of building language models for language identification and classification, we will use them for quantifying the distance between languages, which is a task very similar to language identification.

3. Measures for Computing Language Distance

We propose defining language distances using n-grams extracted from text corpora, in a very similar way as linguistic identification systems learn their language models. More precisely, two different n-gram-based coefficients to measure language distance are proposed: *perplexity* and *ranking*.

3.1. Perplexity

The most widely used evaluation metric for language models is the perplexity of test data. In language modeling, perplexity is frequently used as a quality measure for language models built with n-grams extracted from text corpora [35, 36]. It has also been used in very specific tasks, such as to classify between formal and colloquial tweets [37].

Perplexity is a measure of how well a model fit the test data. More formally, the perplexity (called PP for short) of a language model on a test set is the inverse probability of the test set. For a test set 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 is defined by comparing the n-grams of a text in one language with the n-gram model trained for the

other language. Then, the perplexity of the test text CH in language $L2$, given the language model LM of language $L1$, can be used to define the distance, $Dist_{perp}$, between $L1$ and $L2$:

$$Dist_{perp}(L1, L2) = PP(CH_{L2}, LM_{L1}) \tag{3}$$

The lower the perplexity of CH_{L2} given LM_{L1}, the lower the distance between languages $L1$ and $L2$. The distance $Dist_{perp}$ is an asymetric measure.

3.2. Ranking

The ranking-based distance derives from the observation that, for each language, there is a set of sequence of characters that make up a large portion of any text and their presence is to be expected as word distribution follows Zipf's law. Like in Cavnar and Trenkle's method [31], we used pruned n-grams profiles of two languages to be compared. N-grams are ranked according to frequency in a training corpus, and those with highest frequencies are selected while the rest are discarded. This gives us the pruned character n-grams profile for each language. A *language profile* is thus the ranked list of the most frequent n-grams in the training corpus. Unlike n-gram language models, language profiles do not make use of prior probabilities but simply of ranked lists.

The ranking-based distance between two languages is obtained by comparing the ranked lists of the two languages. It takes two n-gram profiles and calculates a simple rank-order statistic based on an "out-of-place" measure. This measure determines how far out of place an n-gram in one profile is from its place in the other profile [31]. More precisely, given the ranked profiles $Rank_{L1}$ and $Rank_{L2}$ of languages $L1$ and $L2$, respectively, $Dist_{rank}$ is computed as follows:

$$Dist_{rank}(L1, L2) = \sum_{\substack{i=1 \\ gr_i \in Rank_{L1}}}^{K} abs(Rank_{L1}(gr_i) - Rank_{L2}(gr_i)) \tag{4}$$

where $Rank_{L1}(gr)$ is the rank of a specific n-gram, gr_i, in the profile of $L1$, and $Rank_{L2}(gr_i)$ is the rank of the same n-gram in the profile of $L2$. Notice that

the measure only considers those n-grams appearing in the profile of $L1$, which might also appear in that of $L2$. For those cases where the n-gram is not in the profile of $L2$, subtraction of zero is not a good solution since it gives low values for very frequent n-grams appearing only in $L1$. In such a case, we apply a smoothing technique which consists of subtracting the rank of the n-gram in $L1$ from the total size of the profile: $K - Rank_{L1}(gr_i)$.

The range of this measure is from 0 (identical profiles) to K^2 (entirely different ones). Like $Dist_{perp}$, the distance $Dist_{rank}$ is an asymmetric measure.

4. Experiments

Our main objective is to use the language distance metrics defined above to build a current map of the European languages (Subsection 4.2). However, first we will evaluate the two metrics by applying them on the standard language identification task (Subsection 4.1).

4.1. Discrimination between Similar Varieties

The two distance metrics, $Dist_{perp}$ and $Dist_{rank}$, were used to build two language detection systems which were evaluated against the gold standard provided by the Discriminating Similar Languages Shared Task 2016 [25, 38]. The objective is to compare our methods with the participant systems at the Shared Task, and observe how they behave when they are applied on the difficult task of discriminating between very closely related languages or similar varieties.

The State-of-the-art language identification systems perform very well when discriminating between unrelated languages on standard datasets. Yet, this is not a solved problem, and there are a number of scenarios in which language identification has proven to be a very challenging task, especially in the case of very closely related languages or varieties [29]. This is the scenario in which we are evaluating the systems based on our two distance metrics.

Tables 1, 2 and 3 show the accuracy obtained by our two strategies (in bold) on the three tests of DSL Shared Task: test A consists of journal news as

the training data used to build the language models (in-domain dataset), while tests B1 and B2 are constituted by tweets (out-domain dataset). The tables also contain three representative scores for each test: the best, the median, and the lowest accuracies achieved by the participants to the shared task. We specify the position of each system between brackets. This allows us to compare our techniques with the systems that participated to the DSL Shared Task 2016.

In test A (Table 1), our perplexity-based strategy would reach the second position, very close to the best system [39]. By contrast, the rank-based method would be the last one on this task. However, this system is very stable across domains, since it reaches similar scores in out-domain tests (see Tables 2 and 3), where its accuracy is now above the median. The accuracy of the perplexity system slightly decreases in the out-domain tests but it is still clearly above the median, being in total one of the best three systems in the shared task. Most systems yield mixed results across domains. For instance, the best system on test A is the 5th on tests B1 and B2, whereas the second one on test A is the 12th on tests B1 and B2.[2]

The results of these experiments show that our distance-based strategies, even though they were not primarily conceived for the task of language detection, are able to reach very competitive scores. More precisely, the perplexity-based distance is very close to the state-of-the-art measures in the specific task of identifying similar varieties.

4.2. Distance among the Languages of Europe

In the following experiment, we use our distance metrics to build up a network linking forty-four European languages according to their current linguistic distances. This is a more natural application for the two metrics defined above. In this case, instead of a quantitative evaluation, we will provide a visual diagram and a qualitative analysis of the results.

[2]The best system [40] on test B1 is also the best system on B2.

Systems	**Accuracy**
Best (1)	0.8938
Perplexity (2)	**0.8926**
Median (9)	0.8779
Lowest (18)	0.8240
Rank (19)	0.7940

Table 1: Results for test A in DSL Shared Task 2016.

Systems	**Accuracy**
Best (1)	0.920
Perplexity (5)	**0.884**
Rank (7)	0.804
Median (9)	0.688
Last (18)	0.530

Table 2: Results for test B1 in DSL Shared Task 2016.

Systems	**Accuracy**
Best (1)	0.878
Perplexity (6)	**0.820**
Rank (7)	0.762
Median (9)	0.698
Last (18)	0.554

Table 3: Results for test B2 in DSL Shared Task 2016.

4.2.1. Comparable Corpora

The goal of the current experiment is to compare forty-four language models. In order to make them comparable, the texts from which they are generated should belong to similar domains and genres. Thus, we trained the models from *comparable corpora*, that is, from collection of documents in several languages which are not translations of each other, but which share the same genre and/or domain [41, 42].

Two different comparable corpora for the 44 targeted languages were built.

The first corpora was built using the BootCat strategy defined in Baroni and Bernardini [43] and the corresponding Web tool[3] described in Baroni et al. [44]. BootCat is a method to automatically generate a corpus. It starts from a set of seed words which are sent as queries to a search engine. The resulting pages which are at the top of the search engine's hits pages are then retrieved and used to build a corpus [44]. To generate our BootCat comparable corpus, we used the same seed words (translated by means of Google Translate[4]) for the forty-four languages. Given a query in a particular language, most of the documents returned by the system are in the target language even though some of the seed words of the query were not well translated. The final corpus was manually revised and odd pages returned by the search engine were removed.

Following this, we divided the texts generated for each language in two parts: training and test corpora. We followed the same procedure for all languages in order to have the same size: the training corpus consists of a selection of $\sim 120k$ tokens while the test is three times smaller: $\sim 40k$.

The second comparable corpus was derived from different versions of the Bible. Recently, a parallel corpus based on 100 translations of the Bible has been created in XML format [24]. As this corpus does not cover all the European languages, we used additional sources[5] to fill out the same forty-four languages of

[3] WebBootCat is available at `https://the.sketchengine.co.uk`

[4] `https://translate.google.com`

[5] `https://www.bible.com/`

the BootCat corpus. The train and test parts were created in the same manner as previously, except for those languages (e.g. Gaelic) whose Bible version is just a partial translation with few chapters. In those cases, the language is kept in the list even though the size of the training and test corpora does not reach the number of tokens we have established.

All languages were transliterated to the Latin script and normalized using a generic orthography. 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. The encoding is thus close to a phonological one.

Finally, we generated 7-gram models for all languages, which are the input of the language distances.

4.2.2. Building Language-to-Language Matrices

By applying the two distances, $Dist_{perp}$ and $Dist_{rank}$, on the language models (created from both the Web and the Bible corpora), we obtained four 44x44 matrices, each one derived from a distance-corpus strategy: *perp-web*, *perp-bible*, *rank-web*, and *rank-bible*. Since the two distance metrics are asymmetric, each matrix consists of 1936 different values.

We measured the similarity between the four distance-corpus methods by computing the Spearman correlation of the values they generated. Given two strategies, we compare whether their distance values are ranked in a similar manner. Table 4 shows the Spearman coefficient between each pair of methods. We observe that there is strong correlation (75.481) between the two methods based on perplexity, *perp-web* and *perp-bible*, even though they are applied on two very different corpora. When the two distances are applied on the same corpus, the correlation is moderate (65.087, 57.386). Not surprisingly, the correlation is lower (46.056, 33,934) if the two compared strategies are completely different. However, the correlation between the two rank-based strategies is quite weak: 46.256. It follows that, in this experiment, perplexity seems to be

	perp-web	perp-bible	rank-web	rank-bible
perp-web	1	**75.481**	65.087	46.056
perp-bible		1	33,934	57.386
rank-web			1	46.256
rank-bible				1

Table 4: Spearman coefficient between pairs of methods

more stable across domains than the ranking distance.

4.2.3. Language Interaction Network

As previously mentioned, in most works on historical linguistics the distance values among languages are computed from lists of words with a great stability in terms of form/meaning change. The inter-language distances are then supplied to hierarchical clustering algorithms to infer a tree structure for the set of languages. Hierarchical clusters and trees are intended to represent language families and phylogenetic evolution from a diachronic perspective. However, in our work, language distance is not computed from pre-defined lists of stable and universal vocabulary, but from text corpora containing a great variety of linguistic phenomena including loan and foreign words. So, the language distance we have defined intends to measure interactions among languages from a synchronic perspective. The most suitable representation for this type of data is not a hierarchical tree but rather a network showing language interactions.

To create a visual language network, we need to identify true interactions between languages. Given a language and a list of languages ranked by their distance to the first one, we are required to distinguish between those that are actually related (by an arch in the network) to the given language and those that are so far that can be considered as unrelated. For this purpose, we select languages (nodes) and interactions (arcs) from each language-to-language matrix according to a set of filters and requirements (i.e. conditions). More precisely, given a target language, we create an arc with another language if their distance fulfills at least one of the two following conditions:

- It is lower than a minimum score.
- It is lower than a maximum score and is one of the two closest distances to the target language.

To set the optimum values of the two thresholds (minimum and maximum), we built a *gold standard* dataset consisting of 45 well-known language interactions annotated by a linguist who took into account the classification reported in Ethnologue [45]. Only interactions between languages by elaboration (*Ausbau* languages) were considered since they are clearly related. For instance, two examples of manually annotated interactions are the following:

Portuguese	Galician	1
Galician	Spanish	2

The first row means that Galician is the closest language, rank 1, to Portuguese. The second row means that Spanish should be among the 2 closest languages to Galician, since this language is between Portuguese and Spanish. The gold standard dataset only contains language relationships that are well established in comparative linguistics. It is used as a reference test to evaluate the accuracy of all possible networks built from the four language-to-language matrices by using different thresholds. The threshold values giving rise to the highest accuracy are considered to build the best networks. In the end, we select the best network for each one of the four matrices. Table 5 shows the highest accuracy reached by each network (they are called by the name of the method used to create their original matrix). The last column shows the minimum and maximum values that maximize the accuracy. Table 6 shows a sample of languages with their two most similar languages and their distance.[6] The sample was extracted from the *perp-web* network.

[6]The best network configuration was obtained by removing Romance languages from the ranked list associated to non-Romance ones. Given the strong Latin influence over many European languages, the distance between many non-Romance languages and those derived from Latin tend to be short.

networks	accuracy	thresholds
perp-web	**.85**	min=30, max=100
perp-bible	**.85**	min=50, max=200
rank-web	.825	min=5, max=10
rank-bible	.825	min=5, max=10

Table 5: Accuracy reached by the four language networks with the best *max* and *min* thresholds for each one (column 3).

To visualize language networks, we use Cytoscape, an open-source software designed to simulate biochemical reactions and molecular interactions [46]. Languages are attracted and disassociated in a similar way as to how molecules interact with each other. Figure 1 is a network graph, with languages represented as nodes and inter-language interactions represented as links, that is, edges or arcs, between nodes. The length of each arc is a complex function that considers both the distance score between the two linked languages and the number of common nodes to which they are also linked [46].

4.2.4. Analysis

Figure 1 shows that groups of languages having short distances and several internal arcs (only shared by the nodes of the group) tend to form a language family or sub-family: e.g. Romance, Slavic, Germanic, Celtic, Finno-Permian, or Turkic languages. The two groups with strongest internal cohesion (i.e. those having more internal links and shortest distances) are Romance and Slavic. However, Romance languages have a central position in the network since their elements are more connected to external nodes than the Slavic languages. The centrality of Romance language is explained by the fact that most languages have borrowed morphemes and lexical units from Latin in the past, and many neologisms from English nowadays. Notice that a significant portion of English vocabulary (about 56%) comes from Romance languages, a portion of these borrowings come directly from Latin (15%) and another portion through French (41%) [47]. This makes English a special language between Romance

target language	**closest languages**	**distance**
bosnian	croatian	5
bosnian	slovene	8
bulgarian	macedonian	15
bulgarian	serbian	20
catalan	spanish	8
catalan	galician	10
croatian	bosnian	7
croatian	serbian	11
czech	slovak	9
czech	slovene	21
english	french	16
english	dutch	31
french	catalan	14
french	spanish	15
georgian	basque	37
georgian	serbian	47
irish	gaelic	9
irish	english	33
maltese	italian	24
maltese	english	25
portuguese	galician	6
portuguese	spanish	8
serbian	croatian	13
serbian	bosnian	13
spanish	galician	6
spanish	portuguese	8
swedish	danish	12
swedish	norwegian	13
turkish	azeri	20
turkish	english	46

Table 6: Sample of some languages extracted from the *perp-web* network. Only their two closest languages are shown (second column), as well as the distance score between each pair (third column).

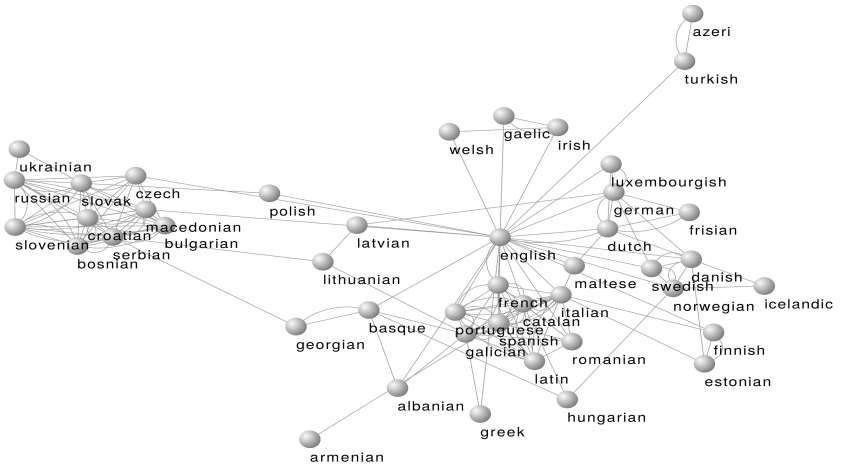


Figure 1: Network of languages spoken in Europe. It has been built using the perplexity-based distance and the Web corpus (*perp-web* strategy).

and Germanic languages, as we can observe in Figure 1. Moreover, it has many interactions with other languages from different families. English turns out to be the core of the map since it is the node with more connections to different sub-areas of the network.

The figure also shows us other interesting cases. Maltese, which is an Arabic language written in Latin alphabet, is interconnected with both English, the other national language in Malta, and Italian, probably because of its close geographical and cultural proximity.

Basque, a non-Indo-European language spoken between Spain and France, is identified by our distance measure as the closest language to Georgian (anyway the distance is quite high as can be observed in Table 6), belonging to the non-Indo-European Kartvelian family indigenous to the Caucasus. In fact, both languages are mutually related because Georgian is also identified as the closest non-Romance language to Basque, which is also strongly connected by our distance to Romance languages probably because of the great lexical influence of

Latin and Spanish. Some controversial comparative-historical and typological approaches have tried to find a Basque-Caucasian connection [48]. However, according to other authors, the case for a link remains unproven, or even, they firmly rejected it [49].

It is also interesting to note that, in our network, Hungarian does not have any connections to Finnish and Estonian. Even if most historical linguists situate Hungarian as a member of the Uralic/Finno-Ugric family, it is also assumed that Hungarian is very detached from the Finno-Permic sub-family (Finnish, Estonian). Similarly, Figure 1 also shows that Polish and the two Baltic languages (Lithuanian and Latvian), even though they belong to the Slavic family, are very far from the core of Slavic languages.

Finally, notice the network does not point at the presence of the Indo-European family. All languages, Indo-European or non-Indo-European, are somehow related either to the members of the family of Romance language or to English. As previously mentioned, our work does not intend to prove the existence of language families and historical relationships, but rather to show the existence of strong links and current interaction from a synchronic perspective.

5. Conclusions

To the best of our knowledge, this is the first time that models and methods from Language Identification have been applied to quantify the distance between languages. Basic n-gram models of characters extracted from text corpora can be used, not only for classifying languages or varieties as in the traditional task of language identification, but also for measuring the distance between language pairs in a global and quantitative way. We have shown that perplexity is an effective way of comparing models, but certainly not the only way. Other strategies, such as ranking-based methods can also be applied on the task of defining a distance measure working on n-grams.

We performed language comparison for forty-four European languages on the basis of two comparable corpora. We calculated the distances of $44**2$ language

pairs and built a network that represents the current map of similarities and divergences among the main languages of Europe.

In many cases languages within the same family or sub-family have low distances as expected, but in some cases there are higher distances than one could expect for languages that are genetically related (e.g. Hungarian and Finish). The contrary is also true; we find low distances, as in the case of Maltese and Italian, for languages that are not related by phylogenetic links. This suggests that our quantitative measure can have applications applications not only on historical linguistics and the classification of language and language varieties, but also on NLP tasks such as machine translation, which requires knowing how close, or far apart, two languages are. This way, the choice of a specific machine translation strategy (e.g. rule-based, SMT, or neural-based) might rely on the distance between the source and target languages.

Finally, it is worth pointing out that our corpus-based strategy is just one more method to compute language distance, which should be seen as a complementary strategy to the existing ones. In particular, corpus-based n-grams might be seen as an additional linguistic source that complements the Swadesh list (and similar closed resources) used in phylogenetics and lexicostatistics. Unlike linguistic phylogenetics, which is focused on diachronic relationships, a n-gram method based on comparable corpora aims at relating languages from a synchronic perspective. The strategy defined in this article is an attempt to adapt the well-known and well-succeeded algorithms used in language identification to compute language distance. However, given that this is a complex and multidimensional task, further methods and strategies will be required to cover all the different aspects of languages. For instance, the use of delexicalized parsers trained and tested with different languages might be an interesting technique to compute the syntactic distance among them [50]. A more global strategy covering more linguistic aspects would be the use of the same technique in machine translation. Evaluating the translation quality of different target languages given the same source and the same models might provide us with a new quantitative metric for measuring the distance among languages.

Corpora and resulting datasets are freely available.[7]

Acknowledgements

We would like to thank the linguist Marta Muñoz-González for her valuable help in building and cleaning the corpora as well as in setting the gold reference. This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), TADEEP (MINECO, ref:TIN2015-70214-P), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

[1] J. Nerbonne, E. Hinrichs, Linguistic distances, in: Proceedings of the Workshop on Linguistic Distances, LD'06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 1–6.
URL http://dl.acm.org/citation.cfm?id=1641976.1641977

[2] F. Petroni, M. Serva, Measures of lexical distance between languages, Physica A: Statistical Mechanics and its Applications 389 (11) (2010) 2280–2283.
URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:389:y:2010:i:11:p:2280-2283

[3] F. Barbançon, S. Evans, L. Nakhleh, D. Ringe, T. Warnow, An experimental study comparing linguistic phylogenetic reconstruction methods, Diachronica 30 (2013) 143–170.

[4] J. Nerbonne, W. Heeringa, Measuring dialect distance phonetically, in: Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology, 1997, pp. 11–18.

[7] http://fegalaz.usc.es/~gamallo/resources/europe_languages.tar.gz

[5] B. Chiswick, P. Miller, Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages, Discussion papers, IZA, 2004.
URL https://books.google.es/books?id=nebHnQEACAAJ

[6] H. Kloss, Abstand languages and ausbau languages, Anthropological Linguistics 9 (7) (1967) 29–41.

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

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

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

[10] M. Swadesh, Lexicostatistic dating of prehistoric ethnic contacts, in: Proceedings of the American Philosophical Society 96, 1952, pp. 452–463.

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

[12] S. Wichmann, Genealogical classification in historical linguistics, in: M. Aronoff (Ed.), Oxford Research Encyclopedias of Linguistics, Oxford University Press, 2017.

[13] J. Nichols, T. J. Warnow, Tutorial on computational linguistic phylogeny., Language and Linguistics Compass 2 (5) (2008) 760–820.
URL http://dblp.uni-trier.de/db/journals/llc/llc2.html#NicholsW08

[14] L. D. Michael, A bayesian phylogenetic classification of tupí-guaraní, LIAMES 15.

[15] R. Gray, Q. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin., Nature.
URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&uid=14647380&cmd=showdetailview&indexed=google

[16] F. Petroni, M. Serva, Language distance and tree reconstruction, Journal of Statistical Mechanics: Theory and Experiment 2008 (08) (2008) P08012.
URL http://stacks.iop.org/1742-5468/2008/i=08/a=P08012

[17] R. D. Gray, F. M. Jordan, Language trees support the express-train sequence of austronesian expansion, Nature 405 (6790) (2000) 1052–1055. doi:10.1038/35016575.
URL http://dx.doi.org/10.1038/35016575

[18] C. J. Holden, R. D. Gray, Rapid radiation, borrowing and dialect continua in the bantu languages, in: P. Forster, C. Renfrew (Eds.), Phylogenetic Methods and the Prehistory of Languages, 2006, Ch. 2.
URL http://groups.lis.illinois.edu/amag/langev/paper/holden06phylogeneticMethods.html

[19] T. Rama, S. Kolachina, B. Lakshmi Bai, Quantitative methods for phylogenetic inference in historical linguistics: An experimental case study of south central dravidian, Indian Linguistics 70.

[20] S. Wichmann, E. W. Holman, D. Bakker, C. H. Brown, Evaluating linguistic distance measures, Physica A: Statistical Mechanics and its Applications 389 (17) (2010) 3632–3639.
URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:389:y:2010:i:17:p:3632-3639

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

[22] Y. Gao, W. Liang, Y. Shi, Q. Huang, Comparison of directed and weighted co-occurrence networks of six languages, Physica A: Statistical Mechanics and its Applications 393 (C) (2014) 579–589.
URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:393:y:2014:i:c:p:579-589

[23] E. Asgari, M. R. K. Mofrad, 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, San Diego, California, 2016, pp. 65–74.
URL http://arxiv.org/abs/1604.08561

[24] C. Christodoulopoulos, M. Steedman, A massively parallel corpus: the bible in 100 languages, Language Resources and Evaluation 49 (2) (2015) 375–395. doi:10.1007/s10579-014-9287-y.
URL http://dx.doi.org/10.1007/s10579-014-9287-y

[25] S. Malmasi, M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, 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), Osaka, Japan, 2016.

[26] P. Gamallo, S. Sotelo, J. R. Pichel, Comparing ranking-based and naive bayes approaches to language detection on tweets, in: Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014, Girona, Spain, 2014.

[27] A. Zubiaga, I. S. Vicente, P. Gamallo, J. R. Pichel, I. Alegria, N. Aranberri, A. Ezeiza, V. Fresno, Tweetlid: a benchmark for tweet language identification, Language Resources and Evaluation (2015) 1–38doi:10.1007/s10579-015-9317-4.
URL http://dx.doi.org/10.1007/s10579-015-9317-4

[28] M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, A report on the dsl shared task 2014, in: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial), Dublin, Ireland, 2014, pp. 58–67.

[29] M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, P. Nakov, Overview of the dsl shared task 2015, in: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), Hissar, Bulgaria, 2015, pp. 1–9.

[30] A. Zubiaga, I. S. Vicente, P. Gamallo, J. R. Pichel, I. naki Alegria, N. Aranberri, A. Ezeiza, V. Fresno, Overview of tweetlid: Tweet language identification at sepln 2014, in: TweetLID - SEPLN 2014, Girona, Spain, 2014.

[31] W. B. Cavnar, J. M. Trenkle, N-gram-based text categorization, in: Proceedings of the Third Symposium on Document Analysis and Information Retrieval, Las Vegas, USA, 1994.

[32] R. Cordoba, L. F. D'Haro, F. Fernandez-Martinez, J. Macias-Guarasa, J. Ferreiros, Language identification based on n-gram frequency ranking, in: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 2007, pp. 27–31.

[33] P. Gamallo, I. Alegria, J. R. Pichel, M. Agirrezabal, Comparing two basic methods for discriminating between similar languages and varieties, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016.

[34] P. Brazdil, C. Soares, A comparison of ranking methods for classification algorithm selection, in: Machine Learning: ECML 2000, 11th European Conference on Machine Learning, Barcelona, Catalonia, Spain, May 31 - June 2, 2000, pp. 63–74. doi:10.1007/3-540-45164-1_8.
URL http://dx.doi.org/10.1007/3-540-45164-1_8

[35] S. F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, in: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL '96, Association for Computational Linguistics, Stroudsburg, PA, USA, 1996, pp. 310–318. doi:10.3115/981863.981904.
URL http://dx.doi.org/10.3115/981863.981904

[36] R. Sennrich, 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, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 539–549.
URL http://dl.acm.org/citation.cfm?id=2380816.2380881

[37] M. González, An analysis of twitter corpora and the differences between formal and colloquial tweets, in: Proceedings of the Tweet Translation Workshop 2015, 2015, pp. 1–7.

[38] C. Goutte, S. Léger, S. Malmasi, M. Zampieri, Discriminating Similar Languages: Evaluations and Explorations, in: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016.

[39] Çağrı Çöltekin, T. Rama, Discriminating Similar Languages with Linear SVMs and Neural Networks, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016.

[40] B. D. Ayah Zirikly, M. Diab, The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016.

[41] X. Saralegi, I. S. Vicente, A. Gurrutxaga, Automatic generation of bilingual lexicons from comparable corpora in a popular science domain, in: LREC 2008 Workshop on Building and Using Comparable Corpora, 2008.

[42] P. Gamallo, J. R. Pichel, Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora, in: CICLING, LNCS, Vol. 6008, Springer-Verlag, Iasi, Romania, 2010, pp. 473–483.

[43] M. Baroni, S. Bernardini, Bootcat: Bootstrapping corpora and terms from the web, in: In Proceedings of LREC 2004, 2004, pp. 1313–1316.

[44] M. Baroni, A. Kilgarriff, J. Pomikálek, P. Rychlý, Webbootcat: a web tool for instant corpora, in: C. O. Elisa Corino, Carla Marello (Ed.), Proceedings of the 12th EURALEX International Congress, Edizioni dell'Orso, Torino, Italy, 2006, pp. 123–131.

[45] R. G. Gordon, J. Dallas, Ethnologue: Languages of the world (15th edn), IL International, 2005.

[46] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks., Genome Research 13 (2003) 2498–2504.

[47] J. M. Williams, In Origins of the English Language, The Free Press, 1975.

[48] N. Sturua, On the basque-caucasian hypothesis, Studia Linguistica 45 (1–2) (1991) 164–175.

[49] R. L. Trask, The History of Basque, Psychology Pres, 1997.

[50] J. Tiedemann, Cross-lingual dependency parsing for closely related languages, in: VarDial 2017, Valencia, Spain, 2017.

II Distância diacrónica intralinguística: Aplicação ao português, espanhol e inglês

– J.R. Pichel, P. Gamallo, I. Alegria. 2018. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), 145–155.

– J.R. Pichel, P. Gamallo, I. Alegria. 2019. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. Natural Language Engineering, 1-22.

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

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.

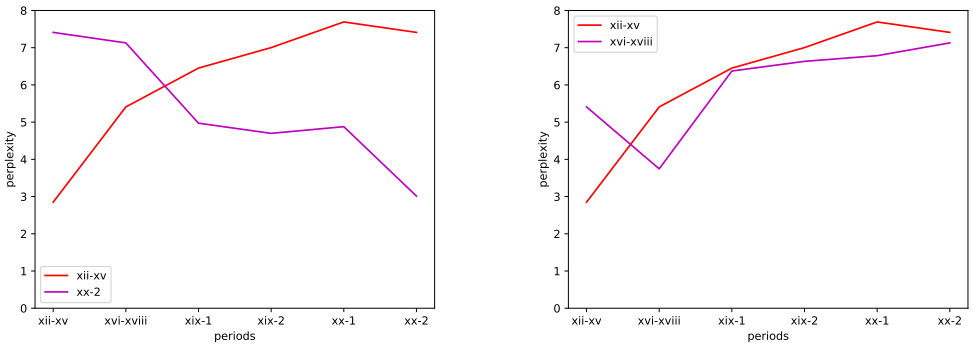
5.2.1 PLD with original spelling

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.

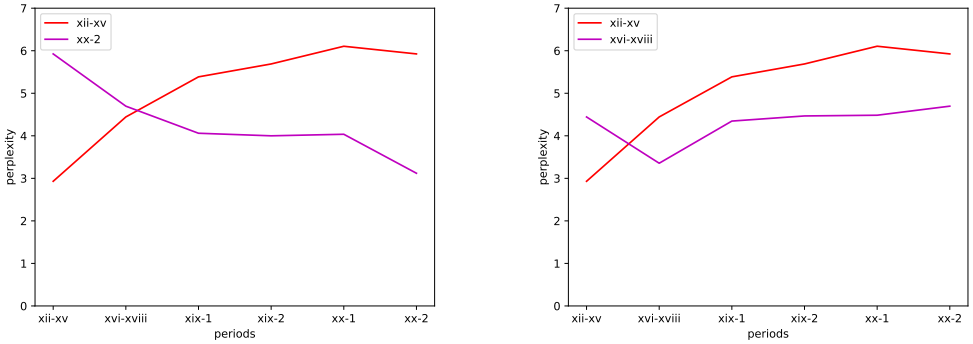
Figure 1(a) plots how XII-XVth diverges from all the periods in a regular basis: there is an almost linear growth from 4.48 for XVI-XVIII (the closest PLD distance), up to 7.69 for XX-1 (the furthest one), even though the distance grows smoothly from XIX-1 and decreases slightly in XX-2. The same pattern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

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 29, 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 (Millar and Trask, 2015), 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 (Petroni and Serva, 2010), in dialectology (Nerbonne and Heeringa, 1997a), in language identification (Malmasi et al., 2016), and in studies about learning additional languages within the field of second language acquisition (Chiswick and Miller, 2004).

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 (Gamallo et al., 2017a). 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 (Gamallo et al., 2016), to measure the distance between languages (Gamallo et al., 2017b), and to quantify the diachronic distance in a language (Pichel et al., 2018). 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 Gamallo et al. (2017a) 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 (Wieling and Nerbonne, 2015). 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" (Cavnar et al., 1994) which is one of the first papers to use n-grams for Language Identification or “Statistical Identification of Language” (Dunning, 1994).

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). 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 (Kloss, 1967) (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 (Malmasi et al., 2016; Zampieri et al., 2018; Kroon et al., 2018), and also for language detection on noisy short texts such as tweets (Gamallo et al., 2014; Zubiaga et al., 2015)

In recent years 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 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 (Zampieri et al., 2018). In the German Dialect Identification task in 2016 the best language identification systems were based on n-gram models (Malmasi et al., 2016). 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 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 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" (Borin, 2013, 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 (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 measure distances using the percentage of shared cognates. Among these studies, Kolipakam et al. (2018), List et al. (2018) and Satterthwaite-Phillips (2011) can be highlighted. 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, to build language taxonomies comparing lexical forms. The method only compares words language-internally and never cross-linguistically. Finally, Satterthwaite-Phillips (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-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 (Gamallo et al., 2017a).

2.3 Language distance

To measure language distances, there were first approaches such as those of Nerbonne and Heeringa (1997b) and Kondrak (2005) 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" (Singh and Surana, 2007).

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 (Liu and Cong, 2013; Gao et al., 2014; Asgari and Mofrad, 2016).

Recently, Degaetano-Ortlieb et al. (2016) have presented an information-theoretic approach based on entropy to investigate diachronic change in scientific English, Rama et al. (2015) have used cross-entropy to measure distances and Degaetano-Ortlieb and Teich (2018) 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 Lai et al. (2018)) 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 (González, 2015), to identify varieties of very related languages (Gamallo et al., 2016), or to measure distances among languages (Gamallo et al., 2017a).

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 Gamallo et al. (2017b). 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 diachronic 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 Klarer (2013): "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 Degaetano-Ortlieb and Teich (2018) or the method of identification of stages done by Th. Gries and Hilpert (2008). 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 Biber (1993): "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 (Rissanen et al., 1993): "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

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*

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.

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 (Rissanen et al., 1993). 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

[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

studies	"A history of the English language" (Baugh and Cable, 1993), "The Short Oxford History of English Literature" (Sanders, 1994), "The Story of English: How the English Language conquered the World" (Gooden, 2009), "The history of English" (Mastin, 2011), "The historical development of the English spelling system" (Jurić, 2013), "An Historical Study of English Function, form and change" (Smith, 2003)
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.*

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 partitions, 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 orthogra-

[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	History of Portuguese Language (Teyssier, 1982), Historical Phonology and Morphology of the Portuguese Language (Williams, 1962), *História da Literatura Portuguesa* (History of Portuguese Literature) (Saraiva, 2001), *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)
sources	Tycho Brahe corpus[9] (Galves and Faria, 2010), Colonia[10] (Zampieri, 2017), *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) (Xavier et al., 1994), 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.*

phy 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	"Historia de la lengua española" (History of Spanish Language) by Rafael Lapesa (Lapesa and Pidal, 1942), "Los 1001 años de la Lengua española" (1001 years of Spanish Language) by Antonio Alatorre (Alatorre, 2002)
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.*

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 (Del Valle, 2013). However, none of the varied orthographies used until the 18th century crystallized. It was only after the reforms of the Royal Academy (RAE) in 1741 that the process of standardization of the written system was actually consolidated as a result of the removal by the RAE of common spelling with other Romance languages such as "ss", "ç" and latinisms (Alatorre, 2002). Thus, a medieval text can be written like this *"dios llamo a moysen dela tienda del paramjento y dixole fabla con los fijos de israel y diles todo onbre de vos que diere ofrenda a dios de ganados esto es de buyes o de ovejas o fazer 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

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

	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.*

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.

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.

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).

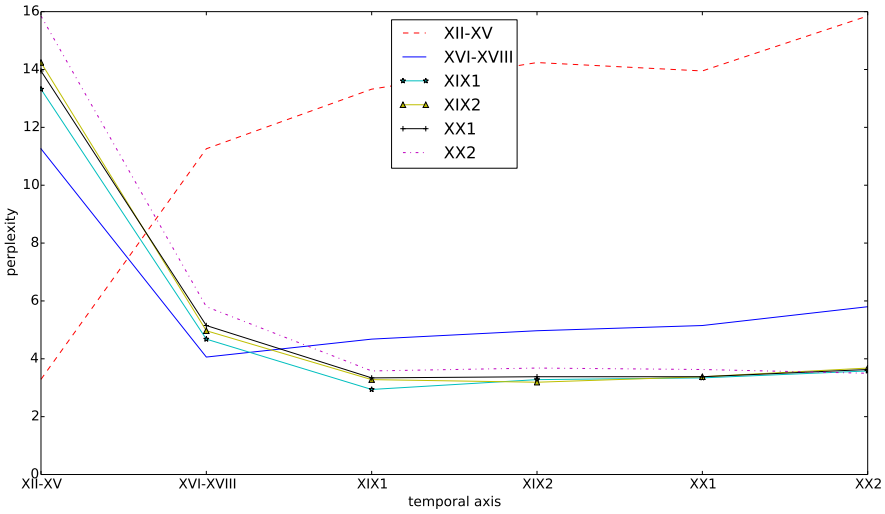
In the case of Figure 1(a) it can be seen more clearly how the medieval period (XII-XV) is progressively separated from the other periods of English, the distance being very large with respect to all periods. In the case of the XV-XVIII period, the distance with regard to the medieval period is much 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

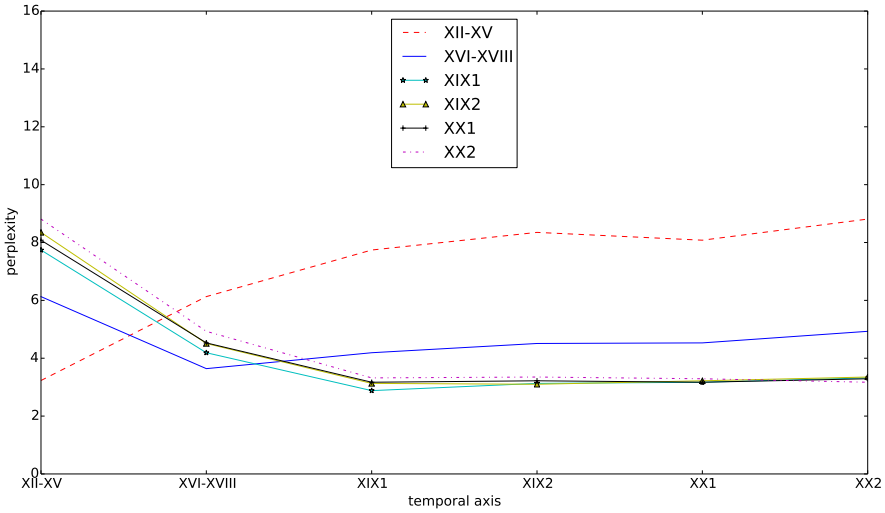
In a second experiment, we have converted the Carvalho-EN-UK corpus into a new common spelling: Carvalho-EN-UK_norm. After this transformation, the same experiment as for Carvalho-EN-UK has been performed. The same will be done for Portuguese (Carvalho-PT-PT and Carvalho-PT-PT_norm) and Spanish (Carvalho-ES-ES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



(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.

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: 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.

	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.)*

5.1.3 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 (Gamallo et al., 2017b), 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

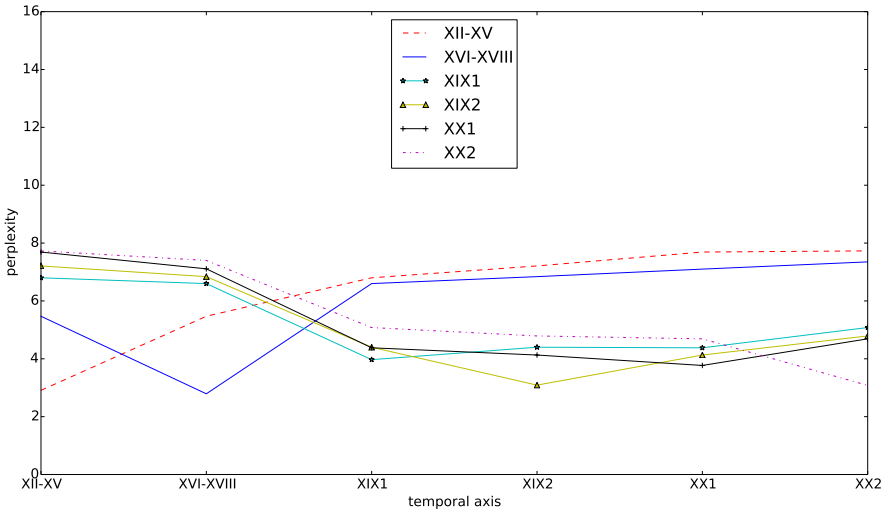
The same two experiments will be performed for Portuguese: the first one consists in applying PLD measure on a Portuguese historical corpus (Carvalho-PT-PT) keeping the original spelling to all and between all historical periods. In the second experiment, we apply the same PLD measure to the same historical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

5.2.1 PLD with original spelling

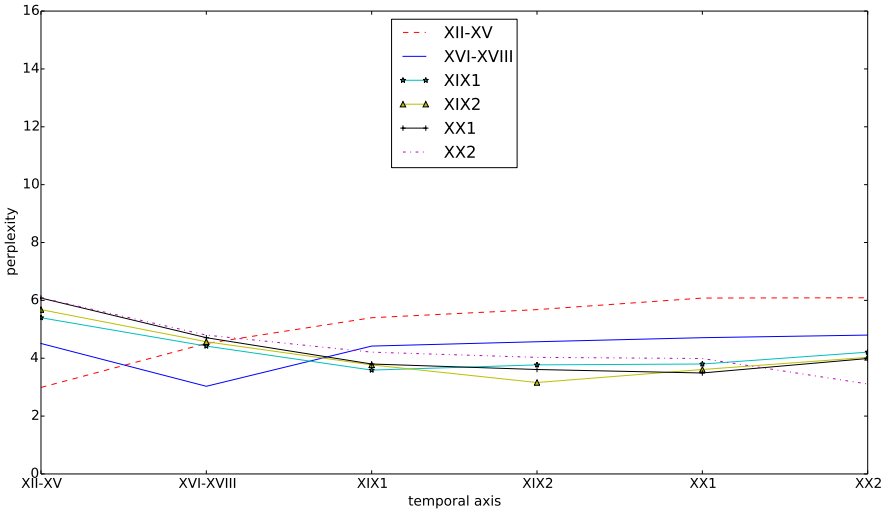
We can observe in Table 10 the following phenomena:

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



(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.

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.

	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)*

	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.)*

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.

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 Gamallo et al. (2017b), 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

	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)*

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 Gamallo et al. (2017b) 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.

5.3.1 PLD with original spelling

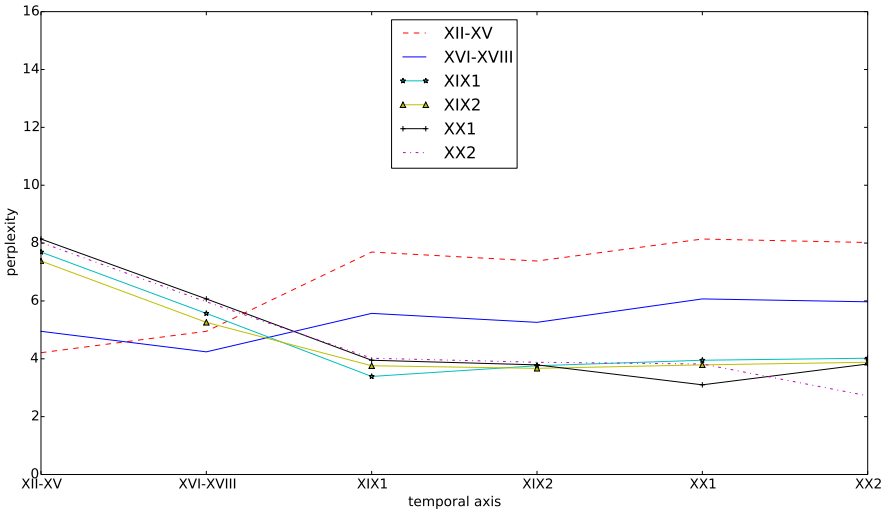
Table 12 shows that the Renaissance and Medieval periods are quite similar: only 4.95 between the two periods. However, between the medieval period and the first half of the 19th century (XIX-1), the PLD reaches 7.69, increasing progressively in later periods and reaching the maximum in the second half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

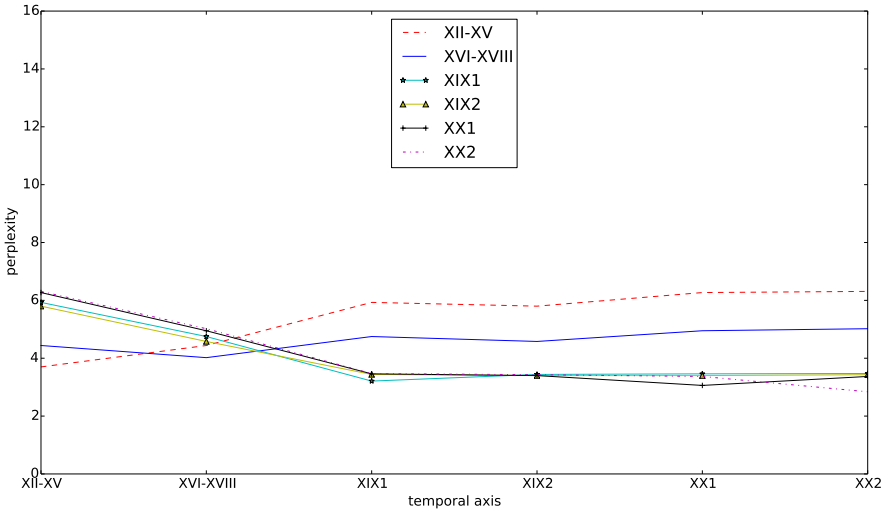
In Figure 3(a), it can be clearly seen that the first half of the nineteenth century (XIX-1) is almost equally distant from the medieval period as from the Renaissance period. Finally, the XIX-1 period is almost linearly distanced from the rest of recent periods: XIX-2, XX-1 and XX-2.

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).



(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.

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).

	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.)*

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 Gamallo et al. (2017b), 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 (Gamallo et al., 2017b) and diachronic language distance applied to a language (Pichel et al., 2018), 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

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

relationship between orthography and distance between languages such as Jágrová et al. (2019), Stenger et al. (2017), and Jágrová et al. (2016).

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 (Specia et al., 2018).

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 Pechenick et al. (2015) which studies how to validate corpora for analysis of cultural and linguistic evolution, the research performed by Bochkarev et al. (2014) 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, Degaetano-Ortlieb and Teich (2018) measure the diachronic change at the lexical and grammatical level in scientific writing and Barron et al. (2018) 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 (Álvaro Iriarte et al., 2018).

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.

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

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.

III Distância diacrónica interlinguística entre línguas próximas: aplicação ao galego, português e espanhol

– J.R. Pichel, P. Gamallo, I. Alegria. 2019. Cross-lingual Diachronic Distance: Application to Portuguese and Spanish. Procesamiento del Lenguaje Natural 63 (2019): 77-84.

– J.R. Pichel, P. Gamallo, I. Alegria, M. Neves. 2020. A Methodology to Measure the Diachronic Language Distance between Three Languages Based on Perplexity. Journal of Quantitative Linguistics (2020): 1-31.

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

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).

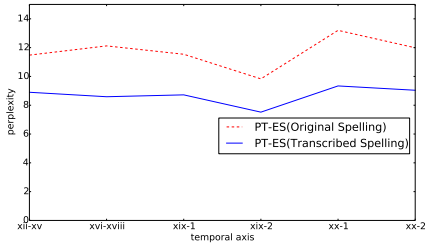


Figure 1: Cross-lingual diachronic distance between Spanish and Portuguese through time axis in OS and TS.

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*.

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

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

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

ARTICLE HISTORY
Compiled February 3, 2020

ABSTRACT

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

KEYWORDS
Language Distance; Historical Linguistics; Perplexity

1. Introduction

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

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

In our previous research, we created perplexity-based methodologies to measure the synchronic distance between European *Abstand* and *Ausbau* languages (Gamallo, Pichel, and Alegria, 2017), to quantify the intralingual diachronic language distance between three languages, one *Abstand* (English) in relation to the others, which have

an *Ausbau* relationship (Portuguese and Spanish) (Pichel, Gamallo, and Alegria, 2018, 2019b), and finally to measure the cross-lingual diachronic distance between two historical *Ausbau* languages: Portuguese and Spanish (Pichel, Gamallo, and Alegria, 2019a).

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

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

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

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

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

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

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

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

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

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

2. Related work

2.1. Language Distance

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

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

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

2.1.1. Linguistic Phylogenetics methodologies

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

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

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

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

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

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

2.1.2. Corpus-driven methodologies

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

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

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

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

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

2.2. Sociolinguistics

Languages are tools of communication between people and, as such, they are conditioned by the human societies where they are used. These societies are in continuous

evolution, which affects their language or languages in different ways. Holmes and Wilson (2017) claim: "Language varies in three major ways which are interestingly interrelated – over time, in physical space and socially. Language change – variation over time – has its origins in spatial (or regional) and social variation". Sociolinguistics is focused on the relationship between societies and languages.

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

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

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

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

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

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

3. Materials and Methods

3.1. Corpus

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

close to) the original texts.

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

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

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

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

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

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

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

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

3.1.1. Galician Corpus

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

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

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

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

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

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

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

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

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

3.1.2. Portuguese Corpus

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

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

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

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

[3] `https://ilg.usc.es/tmilg/`
[4] `https://ilg.usc.es/TILG/`

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

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

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

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

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

3.1.3. Spanish Corpus

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

Since the time of Alfonso X, in Spain, there was a desire to harmonize spelling and create a single standard. However, only after the creation of the Real Academia Española in 1713 and the orthographic standard in 1741 (Lapesa and Pidal, 1942) a

standardized spelling began to spread. The Spanish spelling standard didn't include solutions that are still used in the rest of the Romance languages, such as "ss", "ç" and latinisms (Alatorre, 2002).

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

3.2. Methodology

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

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

This methodology is unsupervised as no annotated text is required.

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

3.2.1. Perplexity-Based Measurement

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2.2. Task Description

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

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

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

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

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

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

4. Results

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

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

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

4.1. Intralingual Diachronic Distance for Galician

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

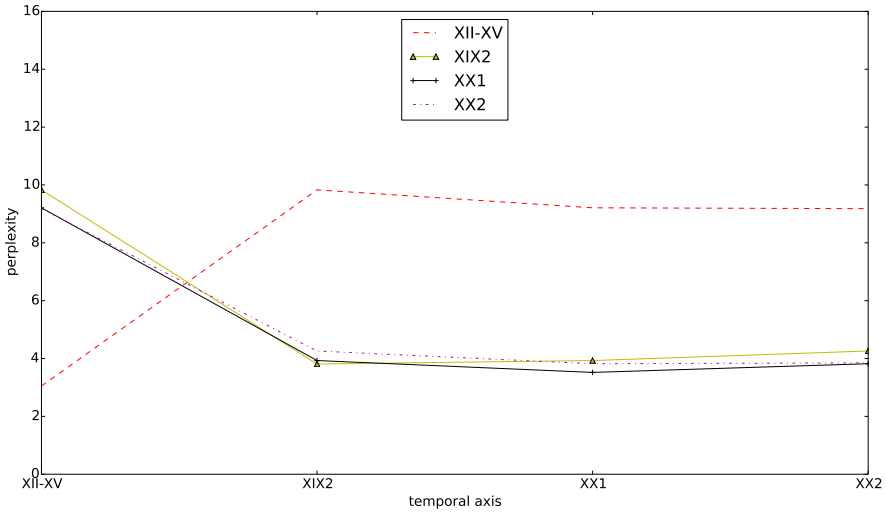
The PLD values in both OS and TS show that Galician in the Middle Ages (XII-XV) shows a significant distance from the period when the Galician language started being written again in an extensive way (XIX-2). This distance decreases progressively in the following subperiods of the 20th century (XX-1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

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

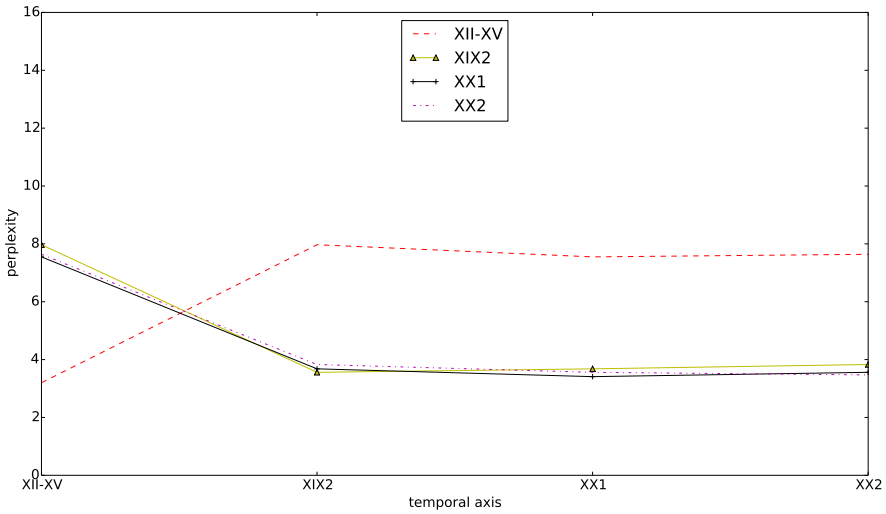
Then, the PLD distance between XII-XV and XX closes a little (9.21 and 9.18 in

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

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



(a) Original spelling



(b) Transcribed spelling

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

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

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

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

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

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

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

4.2. Cross-lingual Diachronic Distance

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

4.2.1. Portuguese-Spanish

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

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

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

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

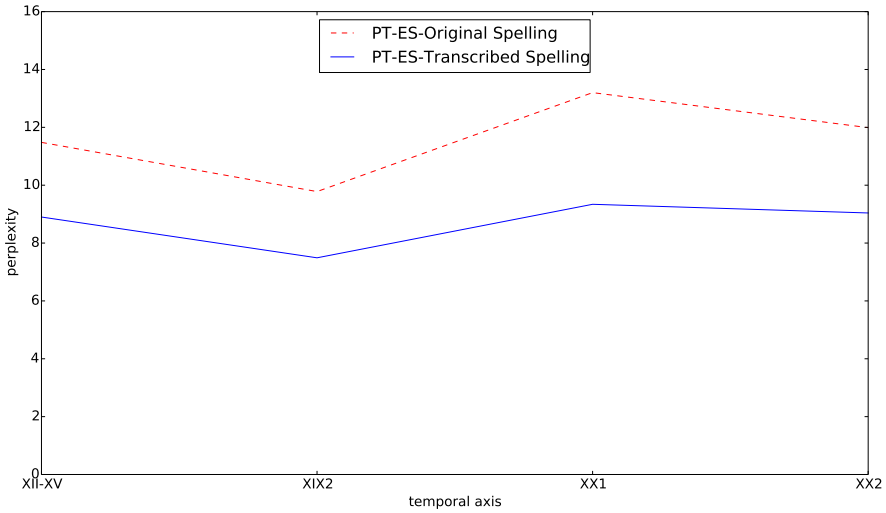


Figure 2. Cross-lingual diachronic distance between Portuguese and Spanish through time axis in OS and TS.

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

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

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

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

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

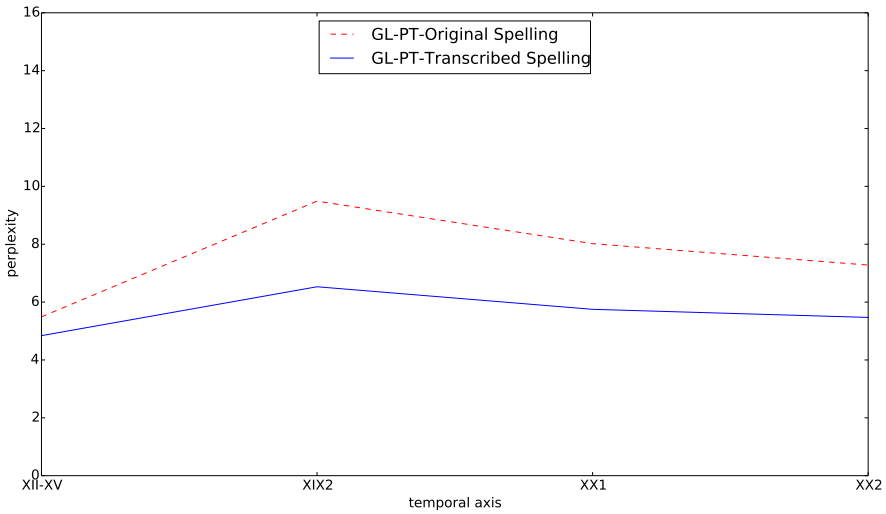


Figure 3. Cross-lingual diachronic distance between Galician and Portuguese through time axis in OS and TS.

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

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

4.2.2. Galician-Portuguese

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

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

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

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

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

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

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

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

4.2.3. Galician-Spanish

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

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

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

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

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

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

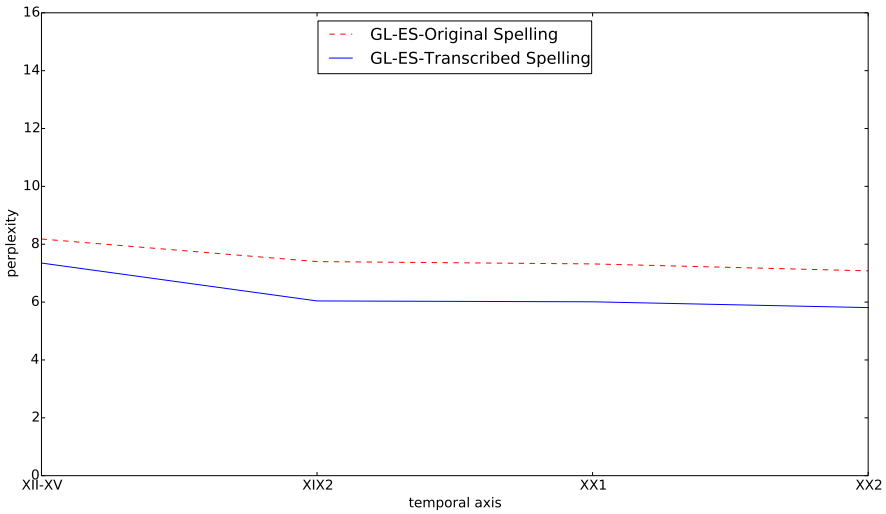


Figure 4. Cross-lingual diachronic distance between Galician and Spanish through time axis in OS and TS.

language, as we mentioned in Section 4.1.

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

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

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

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

5. Discussion

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

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

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

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

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

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

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

5.1. Final Discussion

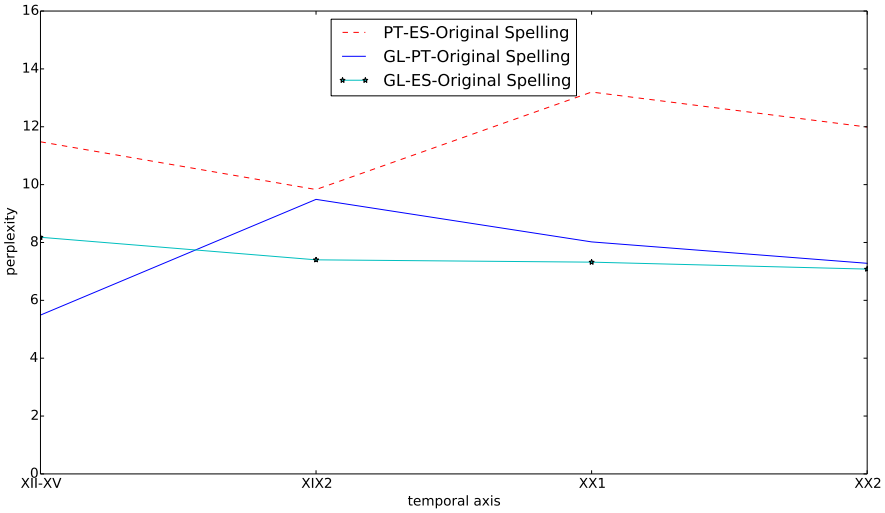
Regarding the *IntraDiaDist* in Galician, we observe the two following facts:

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

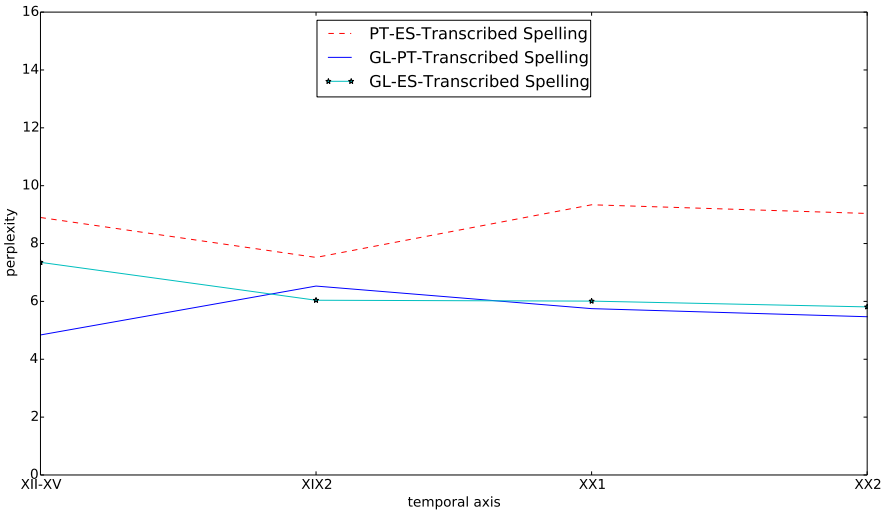
(2) Secondly, we observe that all recent periods (XIX-2, XX-1 and XX-2) are close to each other. In fact, their PLD values are lower than equivalent PLD values found when measuring the distance between different periods both in Portuguese and in Spanish (Pichel et al., 2019b). This observation seems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

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

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



(a) Original spelling



(b) Transcribed spelling

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

calmly absorbed". Our experiments support this hypothesis.

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

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

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

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

5.2. Further work

Based on these results, we would like to apply PLD to measure the distance in polycentric languages such as Portuguese (European and Brazilian Portuguese) and Spanish (European and Latin American Spanish). It is also our aim to measure distances in a diachronic perspective (i.e. did the distance between Argentinean Spanish and European Spanish increase or decrease during their history?, Is the distance between Brazilian Portuguese and European Portuguese greater, lesser, or equal to the distance between Argentinean Spanish and European Spanish?).

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

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

Acknowledgement(s)

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

Funding

This work has been partially supported by TelePares (MINECO, ref:FFI2014-51978-C2-1-R) and DOMINO (Neural Machine Translation, in DOMaIn, and NO supervised.MCIU/AEI/FEDER-UE PGC-2018-102041-B-I00) 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

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

Alecha, Esteve Valls and Manuel González González. 2016. Variación e distancia lingüística na

romania antiqua: unha contribución dialectométrica ao debate sobre o grao de individuación da lingua galega. *Estudos de Lingüística Galega* 8:229–246.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Acoustics, Speech, and Signal Processing, 2005., vol. 1, pages I–833. IEEE.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hinkka, Atte et al. 2018. Data-driven language typology .

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

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

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

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

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

Kessler, Brett. 1995. Computational dialectology in irish gaelic. In *Proceedings of the seventh*

conference on European chapter of the Association for Computational Linguistics, pages 60–66. Morgan Kaufmann Publishers Inc.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nordhoff, Sebastian and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop*

on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011).

Passerini, Thiago Zilio et al. 2019. Ocultação de paternidade ou filiação ilegítima? o lugar do galego na origem da língua portuguesa em textos dos séculos xvi e xix .

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

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

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

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

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

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

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

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

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

Purver, Matthew. 2014. A simple baseline for discriminating similar languages. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

Serva, Maurizio and Filippo Petroni. 2008. Indo-european languages tree by levenshtein dis-

tance. *EPL (Europhysics Letters)* 81(6):68005.

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

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

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

Teyssier, Paul. 1982. História da língua portuguesa .

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

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

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

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

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

Villa, Laura. 2013. *The officialization of Spanish in mid-nineteenth-century Spain: the Academy's authority*, page 93–105. Cambridge University Press.

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

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

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

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

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

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

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

IV Distância diacrónica interlinguística entre variedades diatópicas de línguas: Aplicação ao português e ao espanhol

– J.R. Pichel, P. Gamallo, M. Neves, I. Alegria. 2020. Distância diacrónica automática entre variantes diatópicas do português e do espanhol. Linguamática 12, no. 1 (2020): 117-126.

Proposta recebida em 15 de Outubro 2017 e aceite para publicação em 3 de Dezembro 2017.

Distância diacrónica automática entre variantes diatópicas do português e do espanhol

Automatic diachronic distance between diatopic variants of Portuguese and Spanish

José Ramom Pichel
imaxin software
jramompichel@imaxin.com

Pablo Gamallo
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

Marco Neves
Universidade Nova de Lisboa
mfneves@fcsh.unl.pt

Iñaki Alegria
Universidade do País Basco (EHU/UPV)
i.alegria@ehu.eus

Resumo

O objetivo deste trabalho é aplicar uma metodologia baseada na perplexidade, para calcular automaticamente a distância interlinguística entre diferentes períodos históricos de variantes diatópicas de idiomas. Esta metodologia aplica-se a um corpus construído *adhoc* em ortografia original, numa base equilibrada de ficção e não-ficção, que mede a distância histórica entre o português europeu e do Brasil, por um lado, e o espanhol europeu e o da Argentina, por outro. Os resultados mostram distâncias muito próximas em ortografia original e transcrita automaticamente, entre as variedades diatópicas do português e do espanhol, com ligeiras convergências/divergências desde meados do século XX até hoje. É de salientar que o método não é supervisionado e pode ser aplicado a outras variedades diatópicas de línguas.

Palavras chave

distância linguística, linguística diacrónica, perplexidade

Abstract

The objective of this work is to apply a perplexity-based methodology to automatically calculate the cross-lingual distance between different historical periods of diatopic language variants. This methodology applies to an *adhoc* constructed corpus in original spelling, on a balanced basis of fiction and non-fiction, which measures the historical distance between European and Brazilian Portuguese on the one hand, and European and Argentinian Spanish on the other. The results show very close distances, both in original spelling and automatically transcribed spelling, between the diatopic varieties of Portuguese and Spanish, with slight convergences/divergences from the middle of the 20th century until today. It should be noted that the method is not supervised and can be applied to other diatopic varieties of languages.

Keywords

language distance, diachronic linguistics, perplexity

1 Introdução

Os idiomas e as suas variedades diatópicas mudam constantemente ao longo da história (Millar, Robert McColl and Trask, Larry, 2015), pelo que medir esta distância de forma automática é um desafio.

Historicamente houve diferentes abordagens para calcular esta distância, nomeadamente com base nos estudos filogenéticos no âmbito da Linguística Histórica (Petroni & Serva, 2010), da dialectologia (Nerbonne & Heeringa, 1997a), do campo da aquisição de segunda língua (Chiswick & Miller, 2004), ou da identificação automática da língua (Malmasi et al., 2016).

Para Gamallo et al. (2016), o conceito de distância linguística está intimamente relacionado com o processo de identificação automática da língua. Na verdade, quanto mais difícil for a identificação das diferenças entre duas línguas ou variedades linguísticas, menos distância existe entre elas.

Com este fim, os melhores sistemas de identificação automática de línguas baseiam-se em modelos de n-gramas de caracteres extraídos de corpora textuais (Malmasi et al., 2016). Os n-gramas de caracteres não só codificam informação léxica e morfológica, mas também características fonológicas, uma vez que os sistemas fonográficos escritos estão relacionados com a forma como as línguas eram pronunciadas no passado.

Tendo isto em mente, o objectivo principal do presente artigo é aplicar uma metodologia para medir a distância diacrónica entre duas variedades diatópicas do português e duas do espanhol. Para isso utilizaremos modelos de n-gramas de caracteres obtidos a partir de corpus histórico construído *adhoc*, e a métrica chamada *Perplexity Language Distance* (PLD), baseada na perplexidade e definida em Pichel et al. (2019a). A distância automática entre variedades diacrónicas de línguas diferentes será referida de forma abreviada como *CrossDiaDist*.

A nossa metodologia orientada por corpus não é supervisionada e, portanto, só necessitamos de corpora históricos em bruto. Os textos sobre os quais realizamos as experiências de distância linguística preservam a ortografia original; também calculamos essa distância entre esses mesmos textos transliterados para uma ortografia comum às duas línguas. Um trabalho similar de transcrição foi realizado em Simões et al. (2012), com o objectivo de modernizar ortograficamente versões antigas de palavras em português num dicionário.

De agora em diante, usaremos as siglas *OS* para ortografia original e *TS* para ortografia transcrita.

Em resumo, o nosso objetivo é tentar verificar se as duas variedades de línguas têm uma *CrossDiaDist* estável ou se, pelo contrário, têm períodos convergentes e/ou divergentes. Além disso, tentamos também medir até que ponto a ortografia desempenha um papel nesta *CrossDiaDist* entre variedades diatópicas nos períodos históricos estudados.

O artigo está organizado da seguinte forma: em primeiro lugar, descrevemos alguns estudos sobre distância automática entre línguas com diferentes abordagens na Secção 2. Depois, descrevemos o corpus usado e o conceito de perplexidade na Secção 3. Posteriormente, na Secção 4, apresentamos a metodologia, baseada na perplexidade, a aplicar ao corpus diacrónico. Por fim, na Secção 5, apresentamos e discutimos os resultados, comentando as conclusões e o trabalho futuro na Secção 6.

2 Trabalho relacionado

Para medir a proximidade ou distanciamento entre línguas ou variedades diatópicas de línguas, existem diferentes abordagens: identificação automática de línguas, filogenética e cálculo da distância automática entre línguas.

2.1 Identificação automática de línguas

A identificação automática de línguas é um campo da linguística computacional ainda com desafios por resolver, tais como a diferenciação automática de línguas muito próximas (por exemplo, checo e eslovaco, croata e bósnio) ou variedades diatópicas na mesma língua (por exemplo, espanhol argentino e espanhol europeu, português de Angola e português de Portugal).

Para esta identificação de línguas têm sido usadas diferentes abordagens: dicionários baseados em listas de palavras e heurísticas (ortografia, morfologia, características sintácticas) ou abordagens estatísticas baseadas em modelos de língua (nomeadamente, n-gramas de caracteres ou n-gramas de palavras) a partir de corpora.

Estes últimos, especialmente os baseados em n-gramas de caracteres, costumam ser os melhores sistemas de identificação linguística (Malmasi et al., 2016). A razão provável é que os n-gramas de caracteres não só codificam informações lexicais e morfológicas, mas também características fonológicas, uma vez que os sistemas fonográficos escritos estão relacionados com a forma como as línguas eram pronunciadas no passado. Se os n-gramas forem longos (por exemplo, $\geq$6-gramas), também codificam relações sintácticas, pois podem representar o fim de uma palavra e o início da próxima numa sequência. Também podemos destacar, no que toca à identificação eficiente de idiomas próximos, o trabalho de Tiedemann & Ljubešić (2012) baseado em n-gramas de palavras utilizando *blacklists*.

Entre os estudos mais relevantes e pioneiros devemos destacar os artigos de Cavnar et al. (1994) e Dunning (1994), que são os primeiros trabalhos a usar n-gramas para identificação automática de línguas.

Também existem trabalhos para classificar línguas próximas ou variedades diatópicas (Malmasi et al., 2016; Zampieri et al., 2018; Kroon et al., 2018), e também para a detecção de línguas em textos curtos e com muito ruído como tweets (Gamallo et al., 2014; Zubiaga et al., 2015, 2016).

Finalmente, existem abordagens relacionadas com a aprendizagem profunda (*deep learning* (Lopez-Moreno et al., 2014; Gonzalez-Dominguez et al., 2014). Na *Evaluation Campaign* mais recente organizada no Workshop on Natural Language Processing for Similar Languages, Varieties and Dialects (VarDial-2019), confirma-se que as abordagens mais sofisticadas baseadas em aprendizagem profunda e vectores contextuais não melhoram os resultados das estratégias mais tradicionais com modelos de n-gramas de caracte-

res e classificadores de tipo *Naive Bayes* ou *Support Vector Machine* (Zampieri et al., 2019).

2.2 Filogenética

Na filogenética, para calcular a distância ou proximidade entre línguas, a estratégia consiste em classificar as línguas através da construção de uma árvore enraizada que descreve a história evolutiva de um conjunto de línguas ou variedades relacionadas.

Para isso existem diferentes metodologias, como as baseadas em comparar cognatos lexicais, ou seja, palavras que têm uma origem histórica comum (Nakhleh et al., 2005; Holman et al., 2008; Bakker et al., 2009; Petroni & Serva, 2010; Barbançon et al., 2013). Também existem aproximações lexico-estatísticas baseadas em listas de palavras em vários idiomas, por exemplo, Swadesh list (Swadesh, 1952) ou a base de dados ASJP (Brown et al., 2008), que medem automaticamente distâncias usando a percentagem de cognatos compartilhados. Também a distância Levenshtein entre as palavras numa lista cross-lingual (Yujian & Bo, 2007) é uma das métricas mais comuns usadas neste campo (Petroni & Serva, 2010). Finalmente, também usando uma distância baseada na perplexidade, Gamallo et al. (2017a) construíram uma rede que representa o mapa actual de semelhanças e divergências entre as principais línguas da Europa.

2.3 Distância entre idiomas

Inicialmente houve abordagens como as de Nerbonne & Heeringa (1997b) e Kondrak (2005) a partir da comparação entre formas fonéticas de idiomas, "mas alguns pesquisadores têm argumentado contra a possibilidade de obter resultados significativos a partir da comparação entre formas fonéticas de idiomas", (Singh & Surana, 2007).

Em tempos recentes o cálculo da distâncias entre línguas baseiam-se sobretudo em modelos de língua construídos a partir de corpora paralelos. Estes modelos são construídos a partir das co-ocorrências de palavras e, portanto, a distância entre línguas é resultado da similaridade interlinguística entre estas co-ocorrências (Liu & Cong, 2013; Gao et al., 2014; Asgari & Mofrad, 2016).

Também existem outras aproximações baseadas na entropia para investigar a mudança diacrónica no inglês científico, como em (Degaetano-Ortlieb et al., 2016) (Rama et al., 2015), utilizando a cross-entropy. Finalmente esta distância tem sido calculada utilizando a perplexidade em corpus sincrónicos Gamallo et al. (2017a) e diacrónicos Pichel et al. (2018).

3 Materiais e ferramentas

3.1 Corpora

Para a elaboração das nossas experiências, criámos um corpus diacrónico em OS para o português europeu, português do Brasil, espanhol europeu e espanhol da Argentina.

No que toca ao tamanho deste corpus, seguimos os critérios dos autores do Helsinki Corpus of Historical English (Rissanen et al., 1993), que indicam: "O primeiro problema a ser decidido na compilação de um corpus é o seu tamanho" e "O tamanho do corpus básico é de cerca de 1,5 milhões de palavras".

Em relação aos períodos, como só queremos estudar a distância entre as variantes diatópicas do português e do espanhol em períodos recentes, vamos dividir o nosso corpus exclusivamente em dois períodos históricos: segunda metade do século XX (XX-2) e século XXI até ao presente (XXI-1). Também para tornar este corpus representativo de todas as variantes diatópicas de português e espanhol, tendo em conta a representatividade definida por Biber (1993), incluímos 50% de ficção e 50% de não-ficção para cada período. Além disso, como queremos ver o papel que a ortografia desempenha na distância entre as variedades diatópicas, incluímos sempre textos em OS.

Tendo em conta todas estas características, alargámos o corpus histórico *Carvalho* em OS já desenvolvido em Pichel et al. (2019b) para o português europeu (Carvalho-PT-PT) e espanhol europeu (Carvalho-ES-ES), com o português do Brasil (Carvalho-PT-BR) e o espanhol da Argentina (Carvalho-ES-AR). Temos portanto o português europeu, português do Brasil, espanhol europeu e espanhol da Argentina para os períodos XX-2 e XXI-1. Além disso, os textos incluídos neste corpus estão na ortografia mais próxima possível do original, uma vez que as experiências que iremos realizar serão desenvolvidas tanto em OS como em TS automático. Criado para estas experiências, Carvalho[1] é um corpus histórico em OS disponível gratuitamente para inglês, português europeu, português do Brasil, espanhol europeu e espanhol da Argentina.

Finalmente, Carvalho-PT-PT, Carvalho-PT-BR, Carvalho-ES-ES, Carvalho-ES-AR foram di-

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

Carvalho	Train-pt	Test-pt	Train-br	Test-br	Train-es	Test-es	Train-arg	Test-arg
XX-2	1.688M	363K	1.261M	342K	1.231M	250K	1.280M	256K
XXI-1	1.389M	336K	1.222M	315K	1.270M	285K	1.202M	285K

Tabela 1: Tamanho dos corpora de treino e teste em dois períodos históricos de espanhol-Espanha (es), espanhol-Argentina (arg), português-Portugal (pt) e português-Brasil (br)

vididos em dois subcorpora (treino e teste) para calcular a distância entre variedades diatópicas baseadas na perplexidade. A tabela 1 mostra o tamanho dos corpora de treino e de teste nos dois períodos de cada variante diatópica de português e espanhol para os períodos XX-2 e XXI-1.

A próxima secção descreve as características do corpus diacrónico de Carvalho para cada uma das variedades diatópicas das línguas. Vamos concentrar-nos nos diferentes repositórios de onde foram extraídos todos os documentos e nas características significativas de cada língua.

3.1.1 Corpus do Português Europeu e do Brasil

Para a elaboração dos corpora Carvalho-PT-PT e Carvalho-PT-BR, seleccionámos textos com a ortografia o mais próxima possível do original (OS). Há que ter em conta que nessa OS estão incluídos textos com e sem o Acordo Ortográfico de 1990 (AO'90). As diferentes versões do português (português europeu, português do Brasil, português europeu AO'90 e português do Brasil AO'90) podem ser vistas na Tabela 2.

O português europeu e o português do Brasil têm variado especialmente no século XX do ponto de vista do padrão e da ortografia. Assim, desde o ano 1779 em Portugal, a Academia das Ciências de Lisboa tem promovido diferentes padrões e normas ortográficas (e.g.: 1885, 1911, 1945, 1973, 1990). Por sua vez, a Academia Brasileira de Letras tem convergido ou divergido com estas propostas (e.g.: 1907, 1915, 1919, 1924, 1929, 1931, 1943, 1971, 1986) até ao Acordo Ortográfico de 1990 (AO'90), que ainda hoje é objeto de grande controvérsia em ambos os países e não está totalmente espalhado.

Para criar os corpora de português Carvalho-PT-PT e Carvalho-PT-BR nos subperíodos XX-2 e XXI-1, identificámos e seleccionámos documentos dos seguintes repositórios: Wiki source[2], OpenLibrary[3], Linguateca[4], *Domínio Público*[5] *TesesUSP*[6]

3.1.2 Corpus do Espanhol Europeu e da Argentina

No caso do espanhol, as mudanças relevantes na ortografia ocorreram especialmente desde o aparecimento em 1713 da Real Academia Espanhola e mais tarde em 1741, com um padrão ortográfico diferente do resto das línguas românicas. Esta norma foi consolidada ao longo do tempo com pequenas variações na história, embora houvesse gramáticas na Argentina com orientações divergentes em relação ao espanhol europeu, como em Bello & Cuervo (1932) e Bello et al. (1951). Durante o século XX, a ortografia em espanhol europeu e argentino mudou muito pouco (1952, 1959 e 1999), mas houve contribuições para a gramática da Academia Argentina de las Letras fundada em 1931.

Na Tabela 3, mostram-se trechos do espanhol europeu e espanhol argentino. Para a realização dos corpora Carvalho-ES-ES e Carvalho-ES-AR, obtivemos documentos de ficção e não-ficção nos seguintes repositórios: OpenLibrary[7], Wiki source[8], *Repositorio Institucional CONICET Digital*[9], *TesesUniversidadBuenosAires*[10]

3.2 Perplexidade

Para medir a qualidade dos modelos de linguagem construídos com n-gramas extraídos a partir de corpora (Chen & Goodman, 1996; Sennrich, 2012; Dieguez-Tirado et al., 2005) utilizamos a perplexidade:

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

[2] https://en.wikisource.org/wiki/Category:Portuguese_authors
[3] https://openlibrary.org/
[4] https://www.linguateca.pt/
[5] http://www.dominiopublico.gov.br/
[6] https://www.teses.usp.br
[7] https://openlibrary.org/
[8] https://en.wikisource.org/wiki/Category:Spanish_authors
[9] https://ri.conicet.gov.ar/
[10] http://repositoriouba.sisbi.uba.ar/

Portugal (OS)	**Brasil(OS)**	**PT(AO'90) (OS)**	**BR(AO'90) (OS)**
o princípio da *acção* ou, também, a função essencial da vida animal. (...)	Ele existe – mas quase só por intermédio da *ação* das pessoas: de bons e maus. (...)	em primeiro lugar, porque tais deuses de *facto* não existem, (...)	Só o mau *fato* de se topar com eles, dava soloturno sombrio. (...)

Tabela 2: *Diferenças* entre variedades diatópicas do português europeu (OS), português do Brasil (OS), e ambos com Acordo Ortográfico (AO'90). Este extratos pertencem a documentos dos corpora Carvalho-PT-PT e Carvalho-PT-BR

Espanhol europeu (OS)	**Espanhol da Argentina (OS)**
-¿*Sabes* lo que te digo? -¡Qué! -Que si *tú* fueses el novio de mi hermana, te hubiera matado. (...)	Pero es que *vos* ya lo *sabés*, decía la Maga, resentida. (...)

Tabela 3: *Diferenças* entre variedades diatópicas do espanhol europeu (OS) e o espanhol da Argentina (OS) em documentos do corpus Carvalho-ES-ES e Carvalho-ES-AR

onde as probabilidades de n-grama $P(\cdot)$ são definidas desta forma:

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

Esta métrica está orientada para conferir se um modelo de língua é bom a prever uma amostra de texto. Assim, se a perplexidade é baixa, o modelo de língua é bom a prever a amostra. Pelo contrário, uma perplexidade alta mostra que o modelo de linguagem não é bom a prever a amostra em questão.

A perplexidade tem sido usada também em tarefas muito específicas, tais como medir a dificuldade das tarefas de reconhecimento da fala (Jelinek et al., 1977), para classificar tweets formais e coloquiais (González, 2015), ou para identificar automaticamente línguas estreitamente relacionadas e até variedades diatópicas de línguas (Gamallo et al., 2016).

Tendo em conta isto, definimos recentemente em Pichel et al. (2019b) uma distância baseada na perplexidade chamada Perplexity Language Distance (PLD), para medir a distância diacrónica intralinguística em línguas como o inglês, português e espanhol. A PLD também foi aplicada para medir a *CrossDiaDist* entre duas línguas (Pichel et al., 2019a).

No nosso caso a *CrossDiaDist* será entre duas variedades diatópicas da mesma língua. Esta é definida comparando os n-gramas de um texto numa variedade da língua (português europeu) com o modelo de n-gramas treinado para a outra variedade de língua (português do Brasil). Esta comparação deve ser feita nas duas direcções, dado que PP é uma divergência com valores assimétricos. Além disso esta comparação ao ser diacrónica é por cada período histórico.

Finalmente, para tornar a medida simétrica, a perplexidade do texto do teste CH na variedade diatópica $VL1.2$, dado o modelo da linguagem LM da variedade diatópica $VL1.1$, bem como a perplexidade do texto do teste em $VL1.1$, dado o modelo da linguagem LM de $VL1.2$, são utilizadas para definir *CrossDiaDist* baseada na perplexidade, PLD, entre $VL1.1$ e $VL1.2$, da seguinte forma:

$$PLD(VL1.1, VL1.2) = (PP(A) + PP(B))/2 \quad (3)$$

$$PP(A) = PP(CH_{VL1.2}, LM_{VL1.1}) \quad (4)$$

$$PP(B) = PP(CH_{VL1.1}, LM_{VL1.2}) \quad (5)$$

No trabalho actual, o nosso objectivo é aplicar a PLD para medir a *CrossDiaDist* entre variedades diatópicas de línguas nos mesmos períodos históricos. Com este fim, utilizámos modelos de linguagem baseados em 7-gramas de caracteres, que incorporam uma técnica de alisamento baseada em interpolação linear. Os corpora de treino/teste contêm aproximadamente 1,25M/250K palavras, respectivamente, para que os nossos resultados possam ser comparados e comentados mais tarde na Secção 5.

Finalmente, para que se possa medir a PLD entre períodos de qualquer outro idioma, outros

pares de idiomas ou outros pares de variedades diatópicas de idiomas, desenvolvemos uma arquitetura de pipeline em Perl, disponível em GitHub[11].

4 Métodos e procedimento

O nosso método para calcular a *CrossDiaDist* entre variantes diatópicas de línguas está dividido nas seguintes tarefas sequenciais:

1. Definir períodos históricos comuns para todas as línguas ou variedades diatópicas das línguas. No nosso caso teremos dois períodos (XX-2 e XXI-1) para as seguintes línguas: português europeu, português do Brasil, espanhol europeu e espanhol do Brasil.

2. Obter textos suficientes para todas as variedades diatópicas dos idiomas nos períodos históricos previamente definidos. Antes de incorporá-los no corpus é importante verificar se estão em OS. Para isso, temos de olhar para a história das mudanças ortográficas de cada variedade diatópica. Os excertos em qualquer outra língua são eliminados.

3. Dividir o corpus anterior em treino e teste para cada um dos períodos históricos. A tipologia dos textos deve estar equilibrada em 50% aproximadamente entre ficção e não-ficção. O treino contém pelo menos 1,25M palavras por período, enquanto o teste tem pelo menos 20% do tamanho da partição do treino, ou seja, entre 250K e 350K palavras.

4. Realização da *CrossDiaDist* em OS, que será calculada entre cada variedade diatópica de idioma (PLD(VL1.1, VL1.2), PLD(VL2.1, VL2.2)), e para cada período.

5. Realização da *CrossDiaDist* em TS. A TS é o resultado da aplicação de uma normalização ortográfica em todos os textos com a finalidade de unificar ortograficamente os textos das variedades do português europeu e do português do Brasil, e também da variedade do espanhol europeu e do espanhol da Argentina. Uma vez unificados ortograficamente, é calculada a *CrossDiaDist*, mas em TS. Para isso, foi implementado um transcritor cujo alfabeto consiste em 34 símbolos, representando 10 vogais (incluindo acentos) e 24 consoantes, destinados a cobrir a maioria dos sons mais comuns, incluindo várias palatizações. A codificação é, portanto, próxima da fonológica e, assim, permite simplificar e homogeneizar os casos em que sons semelhantes (geralmente palatalizações) são transcritos de forma diferente em diferentes idiomas. Como as grafias do português europeu e do português do Brasil são muito próximas, a normalização da TS só afecta especialmente a diferenças nas acentuações gráficas. Por exemplo, "acadêmico" no português do Brasil e "académico" no português europeu, ou "assembléia" no português do Brasil e "assembleia" no português europeu são unificados em TS como "academico" e "assembleia". O mesmo acontece com o espanhol europeu e espanhol da Argentina, embora sem diferenças ortográficas salientáveis.

6. Finalmente, avaliação dos resultados finais da *CrossDiaDist* em OS e TS.

5 Avaliação

Após aplicar a metodologia para o cálculo da *CrossDiaDist* baseado em *PLD* em OS e TS, sobre os corpora Carvalho-PT-PT (português europeu), Carvalho-PT-BR (português do Brasil), Carvalho-ES-ES (espanhol europeu) e Carvalho-ES-AR (espanhol da Argentina), e sobre os dois períodos históricos XX-2 e XXI-1, obtemos os resultados que serão explicados a seguir.

5.1 Resultados

A Tabela 4 mostra os resultados da aplicação da metodologia para os corpora de português europeu e português do Brasil nos dois períodos XX-2 e XXI-1 tanto em OS como em TS. Nela vemos que a distância aumenta ligeiramente desde o período XX-2 até a actualidade, entre o português europeu e o português do Brasil, tanto em OS com em TS. Em OS aumenta de PLD: 4,12 para PLD: 4,36 e em TS aumenta de PLD: 3,65 para 3,83.

A Tabela 5 mostra os resultados para o espanhol espanhol europeu e o espanhol da Argentina em OS e TS. Para as variedades diatópicas do espanhol, vemos que a distância diminui ligeiramente entre espanhol de Espanha e espanhol da Argentina entre os períodos XX-2 e XXI-1 em OS e também em TS. Assim, em OS diminui a PLD: 4,27 para PLD: 4,04 e em TS diminui de PLD: 3,60 para 3,45.

Finalmente a figura 1 e a figura 2 retratam a informação da distância entre as variedades diatópicas do português europeu e do português

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

do Brasil, e do espanhol europeu e o espanhol da Argentina.

PLD(PT/BR)	PLD (OS)	PLD (TS)
XX-2	4.12	3.65
XXI-1	4.36	3.83

Tabela 4: Distância diacrónica (PLD) entre o português europeu e o português do Brasil nos períodos XX-2 e XXI-1 em OS e TS.

PLD(ES/AR)	PLD (OS)	PLD (TS)
XX-2	4.27	3.60
XXI-1	4.04	3.45

Tabela 5: Distância diacrónica (PLD) entre o espanhol europeu e o espanhol da Argentina nos períodos XX-2 e XXI-1 em OS e TS.

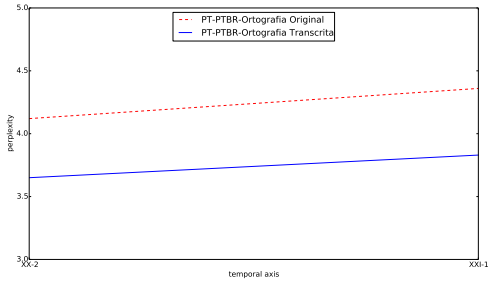


Figura 1: *CrossDiaDist* entre o português europeu e o português da Brasil através do eixo temporal em OS e TS.

5.2 Discussão

Em primeiro lugar, observamos que a *CrossDiaDist* entre as variedades diatópicas do português e do espanhol são muito semelhantes em OS e TS sendo a PLD inferior a 5. Assim a distância mais pequena é de 3.45, entre espanhol de Espanha e espanhol da Argentina em TS, e a máxima é de 4.36 entre o português europeu e português do Brasil em OS. Segundo os resultados reportados em Gamallo et al. (2016), línguas muito próximas como bósnio e croata têm em TS uma distância muito superior, com PLD: 5,90.

Para o caso do português europeu e do português do Brasil, observamos um ligeiro distanciamento no século XXI. Por um lado, talvez este distanciamento se fique a dever a Portugal e o Brasil funcionarem como sistemas culturais diferenciados. O AO'90 foi apresentado como factor

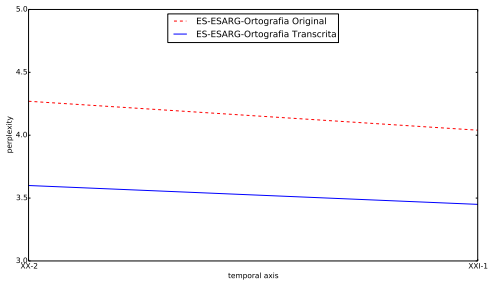


Figura 2: *CrossDiaDist* entre o espanhol europeu e o espanhol da Argentina através do eixo temporal em OS e TS.

de aproximação mas, no entanto, tem tido uma implementação lenta e com muitas resistências, o que talvez seja sintoma das barreiras culturais entre os dois países. De qualquer forma, os valores que apresentamos em TS mostram que a ortografia é um fator pouco relevante no que toca à distância entre o português de Portugal e o português do Brasil. Por outro lado, os valores relativos ao espanhol mostram que é possível registar uma aproximação entre variantes nacionais da mesma língua.

Pelo contrário, no caso do espanhol europeu e do espanhol argentino, vemos que existe uma ligeira aproximação no mesmo período (XXI-1), talvez devido aos esforços de coordenação entre as diferentes academias de língua espanhola e à existência de mais troca de materiais entre os sistemas culturais de Espanha e Argentina.

Finalmente, observamos que a ortografia entre as duas variantes diatópicas de português e espanhol não desempenha um papel importante nesta distância, pois quando calculamos a PLD em TS, ela diminui ligeiramente, mantendo a mesma tendência que em OS.

6 Conclusões e trabalhos futuros

Compilaremos agora as principais conclusões das nossas experiências a partir da aplicação da metodologia de cálculo da distância diacrónica *CrossDiaDist* a variantes diatópicas do português e do espanhol. Também detalharemos na Secção 6.2 próximas investigações em relação à distância automática entre idiomas.

6.1 Conclusões

O cálculo da distância entre idiomas ou variantes diatópicas baseado na perplexidade (PLD) identifica automaticamente idiomas e variantes

diatópicas de idiomas (Gamallo et al., 2017b), mede a distância síncrona entre idiomas (Gamallo et al., 2017a), a distância diacrónica intralinguística em varias línguas Pichel et al. (2018), a *CrossDiaDist* entre línguas (Pichel et al., 2019a) e agora a *CrossDiaDist* entre variantes diatópicas.

Observamos que esta distância entre as variedades diatópicas de português e espanhol é inferior à distância entre línguas muito próximas. Além disso, vemos que o português europeu e o português do Brasil estão a distanciar-se ligeiramente no século XXI. Pelo contrário, o espanhol europeu e o espanhol da Argentina estão a aproximar-se.

Finalmente, a ortografia nestas variantes diatópicas do português e do espanhol não desempenha um papel relevante, pois estas variantes são escritas com ortografias muito próximas ou indistinguíveis.

6.2 Trabalhos futuros

Queremos alargar esta metodologia ao cálculo de distância entre três línguas. Aplicaremos esta metodologia a três línguas muito próximas, como é o caso do galego em relação ao português e ao castelhano.

Outro objectivo é construir um corpus de redes sociais (p.e.: twitter) e comentários em plataformas digitais (p.e: Tripadvisor, AirBnB, Booking, etc.), para variedades diatópicas de português e espanhol, e observar a distância linguística com um corpus de textos mais afastados da gramática padrão e mais próximo das falas populares.

Finalmente, gostaríamos de investigar a relação entre a distância do idioma usando *PLD* e a estimativa da qualidade da tradução automática (Specia et al., 2018; Han et al., 2013).

Agradecimentos

Estamos muito gratos aos professores Dr. Carlos Quiroga e Dr. José António Souto Cabo da Universidade de Santiago de Compostela, Dr. Fernando Venâncio da Universidade de Amesterdão pelas suas observações sobre a história do português europeu e do Brasil, para além da ajuda na escolha de textos de Portugal e do Brasil. Também à professora Maria Isabel Fernández Domínguez pelo seu conhecimento sobre a história do espanhol europeu e ao Dr. Ernesto Vázquez Souza no que toca à história do espanhol da Argentina. Também a ambos, pela ajuda na escolha de textos de referência de ambas as variedades diatópicas. Finalmente, ao Dr. Marcos Garcia da Universidade da Corunha pelos seus conselhos durante as experiências.

Referências

Asgari, Ehsaneddin & 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. Em *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, 65–74. San Diego, California. http://arxiv.org/abs/1604.08561.

Bakker, Dik, Andre Muller, Viveka Velupillai, Soren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant & 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 & T. Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30. 143–170.

Bello, Andrés & Rufino José Cuervo. 1932. Gramática castellana .

Bello, Andrés et al. 1951. Gramática: gramática de la lengua castellana destinada al uso de los americanos .

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

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

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

Chen, Stanley F. & Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. Em *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics* ACL '96, 310–318. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981863.981904. http://dx.doi.org/10.3115/981863.981904.

Chiswick, B.R. & P.W. Miller. 2004. *Linguistic distance: A quantitative measure of the dis-*

tance between english and other languages Discussion papers. IZA. https://books.google.es/books?id=nebHnQEACAAJ.

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

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

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

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

Gamallo, Pablo, José Ramom Pichel & 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 & Inaki Alegria. 2017b. A perplexity-based method for similar languages discrimination. *VarDial 2017* 109.

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

Gao, Yuyang, Wei Liang, Yuming Shi & 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. http://EconPapers.repec.org/RePEc:eee:phsmap:v:393:y:2014:i:c:p:579-589.

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

Gonzalez-Dominguez, Javier, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez & Pedro J Moreno. 2014. Automatic language identification using long short-term memory recurrent neural networks. Em *Fifteenth Annual Conference of the International Speech Communication Association*, .

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

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

Jelinek, Fred, Robert L Mercer, Lalit R Bahl & James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62(S1). S63–S63.

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

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

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

Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez & Pedro Moreno. 2014. Automatic language identification using deep neural networks. Em *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5337–5341. IEEE.

Malmasi, Shervin, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali & Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL Shared Task. Em *Proceedings of the 3rd Workshop on Language Technology for Closely Related Langua-*

ges, Varieties and Dialects (VarDial), 1–14. Osaka, Japan.

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

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

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

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

Petroni, Filippo & Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389(11). 2280–2283. http://EconPapers.repec.org/RePEc:eee:phsmap:v:389:y:2010:i:11:p:2280-2283.

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

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

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

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

Rissanen, Matti, Merja Kytö & Minna Palander-Collin. 1993. *Early english in the computer age: Explorations through the helsinki corpus* 11. Walter de Gruyter.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. Em *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* EACL '12, 539–549. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2380816.2380881.

Simões, Alberto, Álvaro Iriarte Sanromán & José João Almeida. 2012. Dicionário-aberto: A source of resources for the portuguese language processing. Em *International Conference on Computational Processing of the Portuguese Language*, 121–127. Springer.

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

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

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

Tiedemann, Jörg & Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. Em *Proceedings of COLING 2012*, 2619–2634.

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

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. Em *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 1–17.

Zampieri, Marcos, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru & Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. Em *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–16. Ann Arbor, Michigan: Association for Computational Linguistics. doi:10.

18653/v1/W19-1401. https://www.aclweb.org/anthology/W19-1401.

Zubiaga, Arkaitz, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza & Víctor Fresno. 2016. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation* 50(4). 729–766.

Zubiaga, Arkaitz, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza & Víctor Fresno. 2015. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation* 1–38. doi:10.1007/s10579-015-9317-4. http://dx.doi.org/10.1007/s10579-015-9317-4.

V Descrição dos corpora

- Corpus de inglês (Carvalho-EN-UK)
- Corpora de português europeu (Carvalho-PT-PT) e português do Brasil (Carvalho-PT-BR)
- Corpora de espanhol europeu (Carvalho-ES-ES) e espanhol da Argentina (Carvalho-ES-AR)
- Corpus de galego (Carvalho-GL)

Os corpora necessários para as nossas experiências por cada língua ou variante diatópica de línguas foram concebidos tendo em conta que devem ser representativos, de tamanho suficiente, divididos em diferentes períodos históricos relevantes (no caso de cálculos de distância diacrónicos) e escritos com uma ortografia o mais próxima possível dos textos originais, a fim de medir a importância da ortografia como parâmetro de distância entre línguas ou variantes linguísticas.

Para a sua concepção, tivemos em conta a Biber [1993], que assegura que, para que um corpus seja representativo, deve incluir *"a range of text types in a language"*. Por outro lado, no que respeita à dimensão de cada período histórico do corpus, seguimos as recomendações dos autores do corpus histórico do inglês Helsinki [Rissanen et al., 1993b], que afirmam o seguinte: *"The first problem to be decided upon in compiling a corpus is its size"* e *"The size of the basic corpus is c. 1.5 million words"*.

Como trabalhámos com distâncias diacrónicas nas nossas experiências, seguimos a recomendação de Rissanen et al. [1993b], que definiu que um corpus histórico deve ser dividido em pelo menos três períodos: Medieval (séculos XII a XV), Idade Moderna (XVI-XVIII), e Idade Contemporânea (séculos XIX e XX). No entanto, não devemos perder de vista os problemas que comenta Klarer [2013]: *"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"*.

Tendo em conta todas estas questões, criámos um corpus histórico chamado *Carvalho* contendo textos equilibrados de ficção e não-ficção, com um tamanho total de pelo menos 1,5 milhões de palavras para cada período histórico de cada língua ou variante diatópica da língua (se tiver produção escrita suficiente). Este corpus contém textos com todas estas características para todas estas línguas ou variantes diatópicas: galego, português europeu, português do Brasil, espanhol europeu, espanhol da Argentina e inglês.

No entanto, *Carvalho* é um corpus que foi concebido não só para poder verificar hipóteses históricas de cada uma das línguas ou entre elas, mas também para poder ter resultados comparáveis entre as quatro línguas pesquisadas (inglês, português, galego e espanhol) ou entre os dois pares de variedades diatópicas (português europeu – português do

Brasil, espanhol europeu – espanhol da Argentina), e ser capaz de gerar novas observações sobre elas e ainda para ver o papel da ortografia na distância entre as línguas.

Por estas razões, *Carvalho* não inclui textos editados ou transcritos para uma ortografia moderna, mas textos com uma ortografia original ou o mais próxima da original (no caso das medievais), uma vez que as experiências de distância entre línguas ou variedades diatópicas foram realizadas tanto na ortografia original (OS) como numa ortografia transcrita automaticamente (TS) para ver o papel que a ortografia desempenha como parâmetro de distância entre línguas.

Também devido às mudanças ortográficas nas línguas ou a uma maior produção de textos escritos desde o século XIX, *Carvalho* foi dividido nos séculos XIX e XX em dois sub-períodos de 50 anos cada um, para melhor observar as mudanças produzidas nas línguas.

Assim, o inglês sofreu alterações profundas entre a Idade Média e outros períodos históricos, enquanto o português e o espanhol sofreram alterações ortográficas significativas nos séculos XIX e XX.

Por outro lado, o galego não teve produção escrita suficiente para as nossas experiências desde o século XVI até à segunda metade do século XIX, para além da falta de um padrão estável até ao final do século XX. Devemos também salientar, como caso especial, que nos estendemos até ao presente (XXI) para medir a distância entre as variantes diatópicas de português (europeu e brasileiro) e espanhol (europeu e argentino).

Portanto, o corpus histórico *Carvalho* contém seis períodos diacrónicos para três línguas: Carvalho-EN-UK (inglês britânico), Carvalho-PT-PT (português Europeu) e Carvalho-ES-ES (espanhol europeu). Os seis períodos são: medieval (XII-XV), era moderna (XVI-XVIII), primeira metade do século XIX (XIX-1), segunda metade do século XIX (XIX-2), primeira metade do século XX (XX-1), e segunda metade do século XX (XX-2). Também *Carvalho* contém quatro períodos diacrónicos para Carvalho-GL (galego), e dois períodos diacrónicos para Carvalho-PT-BR (português do Brasil) e Carvalho-ES-AR (espanhol da Argentina). *Carvalho* está livremente disponível, excepto para o galego, devido a questões de direitos de autor.[1]

Finalmente, *Carvalho* é dividido num corpus do modelo de língua (MDL) e diferentes corpora de teste, divididos nos seis períodos, a

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

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

Tabela VI.1: Tamanho do corpus Carvalho-EN-UK dividido em modelo de língua (MDL) e Teste por períodos históricos.

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

Tabela VI.2: Tamanho do corpus Carvalho-PT-PT dividido em modelo de língua (MDL) e Teste por períodos históricos.

fim de calcular a PLD. O corpus MDL contém um mínimo de 1.25 milhões de palavras e o segundo, o corpus de teste, contém um mínimo de 250.000 palavras. Isto aplica-se a todas as línguas excepto o galego, o português brasileiro e o espanhol argentino, que não têm todos os períodos.

Na Tabela VI.1, Tabela VI.2 e Tabela VI.3 podemos ver o tamanho do corpus MDL/teste para as línguas em que temos todos os períodos históricos (inglês, português europeu e espanhol europeu). Na Tabela VI.4 vemos o tamanho do corpus no caso do galego, uma língua especial porque não temos corpus suficiente para dois períodos históricos (XV-XVIII e XIX-1). Finalmente, no caso do português do Brasil e do espanhol da Argentina, como já dissemos, prolongámos os períodos históricos de XX-2 até ao presente. Mostramos na Tabela VI.5 e Tabela VI.6 o tamanho destes corpus.

Nas subsecções seguintes detalharemos o corpus diacrónico de *Carvalho* para cada uma das línguas. Centrar-nos-emos nos diferentes repositórios dos quais todos os documentos foram extraídos, em exemplos de textos, obras de referência e nas características significativas de cada língua.

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

Tabela VI.3: Tamanho do corpus Carvalho-ES-ES dividido em modelo de língua (MDL) e Teste por períodos históricos.

Carvalho	**MDL-gl**	**Teste-gl**
XII-XV	1.515M	308K
XIX-2	1.390M	385K
XX-1	1.404M	319K
XX-2	1.504M	398K

Tabela VI.4: Tamanho dos corpora do modelo de língua (MDL)/teste de quatro períodos históricos do galego

Corpus de inglês

O inglês tem sido escrito continuamente desde o período medieval (séculos XII-XV) até aos dias de hoje. Durante estes períodos, há textos suficientes para as nossas experiências.

A Tabela VI.7 mostra informação relevante que utilizámos para construir o corpus Carvalho-EN-UK: os estudos históricos que utilizámos para preparar o material, os recursos do corpus a partir dos quais foram seleccionados os documentos na ortografia original, e algumas amostras de documentos de ficção e não-ficção presentes no corpus final.

No que toca à ortografia, há dois períodos diferentes: o primeiro correspondente aos textos da Idade Média e outra da Idade Moderna até à actualidade. Na Tabela VI.8 podemos ver amostras diferentes dos textos históricos em ortografia original.

[2] `http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html`
[3] `http://www.helsinki.fi/varieng/CoRD/corpora/ZEN/`
[4] `http://www.gutenberg.org/catalog/`
[5] `https://openlibrary.org/`
[6] `https://en.wikisource.org/wiki/Main_Page`

Carvalho	MDL-pt	Teste-pt	MDL-br	Teste-br
XX-2	1.688M	363K	1.261M	342K
XXI-1	1.389M	336K	1.222M	315K

Tabela VI.5: Tamanho dos corpora do modelo de língua (MDL)/teste de dois períodos históricos de português europeu (pt) e português do Brasil (br)

Carvalho	MDL-es	Teste-es	MDL-ar	Teste-arg
XX-2	1.231M	250K	1.280M	256K
XXI-1	1.270M	285K	1.202M	285K

Tabela VI.6: Tamanho dos corpora do modelo de língua (MDL)/teste de dois períodos históricos de espanhol europeu (es) e espanhol da Argentina (arg)

Corpora de português europeu e português do Brasil

O português tem sido escrito continuamente desde o Período Medieval, uma fase conhecida como galego-português, até os dias de hoje. Por esta razão, existe um corpus de tamanho suficiente em português para as nossas experiências em todos os períodos históricos.

A Tabela VI.9 mostra a informação relevante necessária para construir o corpus Carvalho-PT-PT: os estudos históricos que utilizámos para preparar o material, as fontes das quais extraímos os textos na ortografia original, e algumas amostras de documentos de ficção e não-ficção incluídos no corpus final.

Quanto aos diferentes padrões ortográficos em português, houve um

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

Estudos	"A history of the English language" [Baugh and Cable, 1993], "The Short Oxford History of English Literature" [Sanders, 1994], "The Story of English: How the English Language conquered the World" [Gooden, 2009], "The history of English" [Mastin, 2011], "The historical development of the English spelling system" [Jurić, 2013], "An Historical Study of English Function, form and change" [Smith, 2003]
Fontes	The Helsinki Corpus[2], Zurich English newspaper corpus (ZEN)[3], Project Gutenberg[4], OpenLibrary[5], Wikisource[6],
Ficção	"Canterbury Tales" by George Chaucer, "The Complete works" by Shakespeare, "Dracula" by Bram Stoker, "The fifth child" by Doris Lessing
Não-ficção	"Þe Story of Englande als Robert Mannyng", "Theological Tracts" by Bacon, "The blind watchmaker" by Richard Dawkins

Tabela VI.7: Metadados do corpus Carvalho-EN-UK: estudos de referência para o desenho do corpus, fontes de corpus, obras de ficção e não-ficção presentes no corpus

primeiro padrão ortográfico em 1779 promovido pela Academia das Ciências de Lisboa, que foi posteriormente reformado em 1885, 1911, 1945, 1973 e 1990.

Em relação ao português do Brasil, a Academia Brasileira de Letras convergiu ou divergiu com estas propostas (por exemplo: 1907, 1915, 1919, 1924, 1929, 1931, 1943, 1971, 1986) até ao Acordo Ortográfico de 1990 (AO'90). É importante salientar que o AO'90 gerou tanta controvérsia que no período de XX-2 e, especialmente no século XXI, podemos encontrar em Portugal e no Brasil textos que seguem o AO'90 e outros que não o seguem.

Na Tabela VI.10 podemos ver diferentes amostras destes textos históricos de português europeu em ortografia original em todos os períodos históricos. Por último, na Tabela VI.11 veremos o mesmo mas para o português do Brasil, com a excepção de que apenas cobre os períodos históricos de XX-2 e XXI-1.

Corpora de espanhol europeu e da Argentina

Quanto ao espanhol, existe, como no caso do português ou do inglês, um corpus de tamanho suficiente em todos os períodos históricos, o que

Texto XII-XV	"sodenly withdrawe ageyne al be thei greuous synnes I gesse that thei ben nought dedly Now myghte men axe wherof that pride sourdeth" (The Canterbury Tales, Chaucer)
Texto XVI-XVIII	"Tut, man, one fire burns out another's burning, One pain is lessen'd by another's anguish; Turn giddy, and be holp by backward turning; One desperate grief cures with another's languish: Take thou some new infection to thy eye, And the rank poison of the old will die." (Romeo and Juliet, Shakespeare)
Texto XIX-1	"Oh cold, cold, rigid, dreadful Death, set up thine altar here, and dress it with such terrors as thou hast at thy command: for this is thy dominion!" (A Christmas Carol, Charles Dickens)
Texto XIX-2	"Rudimentary organs are eminently variable; and this is partly intelligible, as they are useless or nearly useless, and consequently are no longer subjected to natural selection. They often become wholly suppressed." (The Descent of Man, Charles Darwin)
Texto XX-1	"Mrs. Dalloway said she would buy the flowers herself. For Lucy had her work cut out for her. The doors would be taken off their hinges; Rumpelmayer's men were coming. And then, thought Clarissa Dalloway, what a morning–fresh as if issued to children on a beach."(Mrs Dalloway, Virginia Woolf)
Texto XX-2	"A towel, it says, is about the most massively useful thing an interstellar hitch hiker can have. Partly it has great practical value - you can wrap it around you for warmth as you bound across the cold moons of Jaglan Beta" (The Hitchhiker's guide to Galaxy, Douglas Adams)

Tabela VI.8: Amostra de textos históricos em inglês

nos permitiu realizar as nossas experiências de distância entre línguas.

A Tabela VI.12 mostra alguma informação relevante necessária para construir o corpus Carvalho-ES-ES, como também comentámos no caso de Carvalho-EN-UK ou Carvalho-PT-PT. Nele podemos encontrar os estudos históricos que utilizámos para preparar o material, os recursos do corpus a partir dos quais foram seleccionados os documentos em ortografia original, e algumas amostras de documentos de ficção e não-ficção que participaram no corpus final.

Em relação à ortografia, desde a época de Alfonso X, em Castela, havia um desejo de harmonizar a ortografia e criar um padrão único.

[17] `https://www.gutenberg.org/browse/languages/es`
[18] `https://en.wikisource.org/wiki/Category:Spanish_authors`
[19] `https://openlibrary.org/`
[20] `http://repositoriouba.sisbi.uba.ar/gsdl/cgi-bin/library.cgi`

Estudos	History of Portuguese Language [Teyssier, 1982], Historical Phonology and Morphology of the Portuguese Language [Williams, 1962], *História da Literatura Portuguesa* (History of Portuguese Literature) [Saraiva, 2001], *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]
Fontes	Tycho Brahe corpus[7][Galves and Faria, 2010], Colonia[8] [Zampieri, 2017], *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) [Xavier et al., 1994], Project Gutenberg, especificamente para o século XIX[9], Wiki source[10], OpenLibrary[11], Arquivo Pessoa[12], Linguateca[13], *Corpus de Textos antigos* [14], *Domínio Público*[15], *TesesUSP*[16]
Ficção	Cantigas de Dom Dinis, "Cancioneiro Geral de Resende", "Elegia" por Barbosa du Bocage, "A relíquia" por Eça de Queiroz, "Elegias" por Teixeira de Pascoaes, "Caim" por José Saramago
Não-Ficção	"Chronica de Dom João I", "Documentos Notariais", "Opúsculos" por Alexandre Herculano, "Descobrimento de Philipinas", "Páginas Archeologicas" por Felix Alves, "Este mundo da injustiça globalizada" por Saramago

Tabela VI.9: Metadados do corpus Carvalho-PT-PT e Carvalho-PT-BR: estudos de referência para o desenho do corpus, fontes de corpus e algumas obras de ficção e não-ficção presentes no corpus

Contudo, só depois da criação da Real Academia Espanhola em 1713 e do padrão ortográfico em 1741 [Lapesa and Pidal, 1942] uma ortografia padronizada começou a espalhar-se. O padrão ortográfico espanhol não incluía soluções que ainda são utilizadas nas outras línguas românicas, tais como "ss", "ç" e latinismos [Alatorre, 2002].

Esta norma foi consolidada ao longo do tempo com pequenas variações na história, embora houvesse gramáticas na Argentina com orientações divergentes em relação ao espanhol europeu, como em Bello and Cuervo [1932] e Bello et al. [1951]. Durante o século XX, a ortografia do espanhol europeu e argentino mudaram muito pouco (1952, 1959 e 1999), mas houve contribuições para a gramática pela Academia Argentina de Letras fundada em 1931.

Na Tabela VI.13 podemos ver diferentes amostras destes textos históricos para o espanhol europeu em ortografia original.

Finalmente, na Tabela VI.14 podemos ver amostras para o espanhol da Argentina referentes aos dois períodos históricos de investigação, também em ortografia original.

Corpus de galego

Em relação ao galego, o período medieval (séculos XII-XV) é conhecido como período galego-português: *"From the late twelfth century to the early fourteenth, Galician-Portuguese, a convenient term limited to the period when the two languages had not yet become clearly differentiated"* [Azevedo, 2005]. Neste período há textos suficientes pertencentes ao período medieval para as nossas experiências, que duraram entre o século XII e o XV. No entanto, durante os séculos XVI-XVIII e a primeira metade do século XIX (XIX-1) a escrita em galego foi abandonada de forma relevante e, portanto, não existem textos suficientes escritos para as nossas experiências. Pelo contrário, a partir da segunda metade do século XIX (XIX-2), o período chamado *Rexurdimento*, até hoje, temos documentos suficientes para podermos medir a PLD.

O corpus Carvalho-GL que recolhemos para o período medieval (XII-XV) faz parte do corpus TMILG (Tesouro Medieval Informatizado da Lingua Galega) [Varela Barreiro, 2004, Moura et al., 2008]. Para os períodos XIX-2, XX-1 e XX-2, foram utilizados textos do corpus TILG (Tesouro Informatizado da Lingua Galega) [Santamarina, 2003]. Estes corpora apenas permitem direitos de acesso com consultas previamente

definidas, embora os seus autores possam ser contactados para investigação sobre os mesmos.

A Tabela VI.15 contém meta-informação relevante que utilizámos para conceber e construir o corpus Carvalho-GL: os estudos históricos que utilizámos para preparar o material, as fontes dos documentos em ortografia original, e algumas amostras de obras relevantes de ficção e não-ficção incluídas no corpus final.

Em termos de ortografia, desde a Idade Média até à actualidade, as ortografias galegas oscilaram entre a aproximação às ortografias portuguesas (período medieval) e à ortografia castelhana (período moderno e contemporâneo).

Finalmente, na Tabela VI.16 podemos ver diferentes exemplos destes textos históricos na ortografia original.

[21] `https://ilg.usc.es/tmilg/`
[22] `https://ilg.usc.es/TILG/`

Texto XII-XV	"A quantos esta carta uiren faço saber que Domingos perez filho de Maria. martjz dicta Daynha mj mostrou hũa mha carta que de mjn ten pola qual eu enprazei a el. e aa primeira molher con que fosse casado dous casaaes e hũu Moynho que eu ei na quintaa de Maceeira." (Chancelaria de Dom Afonso. Volume I)
Texto XVI-XVIII	"Hum Sabio disse, que naõ havia neste mundo homem, que se conhecesse; porque todos para comsigo saõ como os olhos, que vendo tudo, naõ se vem a si mesmos; e daqui vem naõ darem muita fé em si de suas perfeiçoens" (Arte de furtar, Padre Manuel da Costa)
Texto XIX-1	"Foi um anjo? Foi um demonio? Foi algum feiticeiro? Mysterio. Não ha, nem haverá, talvez, nunca, philosopho que o explique; salvo se tal phenomeno é uma das maravilhas do magnetismo animal." (Lendas e Narrativas. Tomo I, Alexandre Herculano)
Texto XIX-2	"O galego, ao servir-lhe o nabo e grão, rosnou com estima: Ora, seja bem aparecidinho o Senhor Lino! Ao cozido este cavalheiro, abandonando a Nação onde percorrera miudamente os anúncios, pousou em mim os olhos amarelentos de bílis e baços, e observou que estávamos gozando desde os Reis um tempinho de apetite." (A Relíquia, Eça de Queiroz)
Texto XX-1	"Mas eu fico triste como um pôr de sol Para a nossa imaginação, Quando esfria no fundo da planície E se sente a noite entrada Como uma borboleta pela janela. "(O guardador de rebanhos, Alberto Caeiro)
Texto XX-2	"Uns com os outros, não mostram qualquer relutância em reconhecer que a vida no céu é a coisa mais aborrecida que alguma vez se inventou, sempre o coro dos anjos a proclamar aos quatro ventos a grandeza do senhor, a generosidade do senhor, inclusive, a beleza do senhor " (Caim, José Saramago)
Texto XXI-1	"No entanto, tranquilizava-o o facto de que o homem, embora não parecesse ter o físico adequado, deveria pertencer, outra possibilidade não cabia, pelo menos, ao grupo daqueles que haviam sido contratados para ajudar a empurrar " (A viagem do Elefante, José Saramago)

Tabela VI.10: Amostra de textos históricos em português

Texto XX-2	"e o próprio doutor Teodoro, cujo nome não deve ser objeto de injusto esquecimento pelo simples fato de o termos como preclaro herói desta despretensiosa crônica de costumes." (Dona Flor. Jorge Amado)
Texto XXI-1	"A frota Targaryen fora esmagada enquanto estava ancorada e enormes blocos de pedra foram arrancados dos parapeitos e desabaram sobre as águas encapeladas do mar estreito. A mãe morrera ao dá-la à luz, e por esse fato Viserys nunca a perdoara." (Crónicas de gelo, Martin)

Tabela VI.11: Amostra de textos históricos em português do Brasil

Estudos	"Historia de la lengua española" (History of Spanish Language) by Rafael Lapesa [Lapesa and Pidal, 1942], "Los 1001 años de la Lengua española" (1001 years of Spanish Language) by Antonio Alatorre [Alatorre, 2002]
Fontes	Project Gutenberg[17], Wikisource[18], OpenLibrary[19], Repositorio Digital Institucional Universidad Buenos Aires[20]
Ficção	"Libro Buen Amor" por el Arcipreste of Hita, "Don Quixote de la Mancha" por Cervantes, "La Gaviota" por Fernán Caballero, "La Regenta" por Leopoldo Alas Clarín, "Platero y Yo" por Juan Ramón Jiménez, "Pascual Duarte" por Camilo José Cela
Não-Ficção	"General estoria" by Alfonso X, "Naufragios" por Cabeça de Vaca, "Historia de Castilla", "Historia del Derecho español" por Eduardo Hinojosa, "Historia de la decadencia de España" por Cánovas del Castillo, "Análisis del Protágoras de Platón" por Gustavo Bueno

Tabela VI.12: Metadados do corpus Carvalho-ES-ES e Carvalho-ES-AR: estudos de referência para o desenho do corpus, fontes de corpus e algumas obras de ficção e não-ficção presentes no corpus

Texto XII-XV	"et este nombre munene quiere dezir en arauigo tanto como enel nuestro lenguage de castiella lo que desseamos. et cuenta aquel sabio que esta duenna era de buen seso et de grand conseio." (General Estoria. Alfonso X)
Texto XVI-XVIII	"Con esto dexaron la hermita y picaron hazia la venta, y a poco trecho toparon vn mancebito que delante dellos yua caminando no con mucha priesa, y assi le alcaçaron" (Don Quixote, Miguel de Cervantes)
Texto XIX-1	"El alemán le hizo entonces un fiel relato de su vida. Era el sexto hijo de un profesor de una ciudad pequeña de Sajonia, el cual había gastado cuanto tenía en la educación de sus hijos." (La gaviota, Fernán Caballero)
Texto XIX-2	"¿Cuánto tiempo dura la belleza del hombre crapuloso, de la mujer liviana, del malvado, en cuyo rostro contraído no tardan en reflejarse sus pensamientos siniestros?" (La igualdad social, Concepción Arenal)
Texto XX-1	"La puerta se abrió al poco rato, asomando á ella Sebastiana, sorprendida por este llamamiento cuando iba á acostarse. "(La Tierra de todos, Blasco Ibáñez)
Texto XX-2	"Tenía la mirada en un punto muerto. Lucio alzaba los ojos al amarillo cielo, raso, que se vencía por el centro, como una gran barriga." (El Jarama, Sánchez Ferlosio)

Tabela VI.13: Amostra de textos históricos em espanhol

Texto XX-2	"Las recovas de la plaza Independencia, vos también las conocés, Horacio, esa plaza tan triste con las parrilladas, seguro que por la tarde hubo algún asesinato y los canillitas están voceando el diario en las recovas." (Rayuela. Julio Cortázar)
Texto XXI-1	"— ¡Vos no sabés nada! Te quedaste con las rimas de Becker. ¡Antigua! Julia Prilutzky Farny es lo mejor." (Las muertes de Juana, Susana Irene Astellanos)

Tabela VI.14: Amostra de textos históricos em espanhol da Argentina

Estudos	"Historia da Literatura galega contemporánea" [Carvalho, 1981], "Galician and Castilian in contact: historical, social and linguistic aspects" [Monteagudo and Santamarina, 1993], "A construção da língua portuguesa frente ao castelhano: o galego como exemplo a contrario." [Corredoira, 1998], "Historia social da lingua galega: idioma, sociedade e cultura a través do tempo" [Monteagudo and Romero, 1999], "Historia da Literatura galega" [Vilavedra and Fdez, 1999], "Gramática da lingua galega II. Morfosintaxe " [Freixeiro Mato, 2000], "O estudo do mundo lusófono no sistema literário galego. Bases metodológicas para o estudo dos sistemas emergentes e as suas relaçons intersistémicas." [Torres Feijó, 2002] "A fouce, o hórreo eo prelo: Ánxel Casal ou o libro galego moderno" [Vázquez Souza, 2003] "Historia de Galicia"[Villares, 2004] "Historia da lingua galega" [Paz, 2008], "O galego (im)possível" [Rodrigues Fagim, 2001]
Fontes	TMILG (Tesouro Medieval Informatizado da Lingua Galega) [21], TILG (Tesouro Informatizado da Lingua Galega) [22],
Ficção	"Cantigas de Santa Maria" por Alfonso X, "Follas Novas" por Rosalia de Castro, "Queixumes dos Pinos" por Eduardo Pondal, "Da Terra asoballada" por Ramón Cabanillas, "Crónica de nós" por Xosé Luís Méndez Ferrín
Não-Ficção	"Crónica Geral de Castela", "O Tío Marcos da Portela" by Valentín Lamas Carvajal, "A nosa terra" uma revista galega, "Para un axeitado dereito foral galego" por Carlos Abraira López

Tabela VI.15: Metadados do corpus Carvalho-GL: estudos de referência para o desenho do corpus, fontes de corpus, obras de ficção e de não-ficção presentes no corpus

Texto XII-XV	"Esta é como Santa Maria juigou a alma do romeu que ya a Santiago, que sse matou na carreira por engano do diabo, que tornass ' ao corpo e fezesse pẽedença." (Cantigas de Santa Maria, Alfonso X)
Texto XIX-2	"Ós meus compañeiros de monteira. Eiquí me tedes tan prantado como Dios me deu, co'ista cara de home de ben, pois é necesario que sepiades que eu son un gallego enxebre; tan enxebre cal os do sigro dazasete, que gastaban monteira e calzós de rizo." (O Tío Marcos da Portela)
Texto XX-1	"¡Ilusos! Tiven que facerlle ver que a dictadura primorriverista, que causou o derrubamento da monarquía en Hespaña, era equivalente á de Xohan Franco, que causara o derrubamento da monarquía portuguesa. I engadínlle: "A dictadura de Oliveira Salazar e a que vai vir a Hespaña se vostedes non saben evitá-la."(Sempre en Galiza, Castelao)
Texto XX-2	"D. MARCIAL Recuando, diante do sinxelo razonamento de Alberte, e voltando ao procedimento afirmativo, mandón, propio deste tipo de cregos en certos países. Berrando." (Teatro pra a xente, Blanco Amor)

Tabela VI.16: Amostra de textos históricos em galego

Bibliografia

Al-Onaizan, Y. and Knight, K. (2002). Machine transliteration of names in arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–13. Association for Computational Linguistics.

Alatorre, A. (2002). *Los 1001 años de la lengua española*, volume 3. Fondo de Cultura Económica.

Alecha, E. V. and González, M. G. (2016). Variación e distancia lingüística na romania antiqua: unha contribución dialectométrica ao debate sobre o grao de individuación da lingua galega. *Estudos de Lingüística Galega*, 8:229–246.

Álvarez, R. and Monteagudo, H. (2005). *Norma lingüística e variación: unha perspectiva desde o idioma galego*. Inst. da lingua Galega.

Anna, A. A. S. and Weller, L. (2020). The threat of communism during the cold war: A constraint to income inequality? *Comparative Politics*, 52(3):359–393.

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

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

Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., Egorov, D., Mailhammer, R., Grant, A., and Holman,

E. W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181.

Baldwin, T. and Lui, M. (2010). Language identification: The long and the short of the matter. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237.

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

Barrault, L., Bojar, O., Costa-Jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., et al. (2019). Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Baugh, A. C. and Cable, T. (1993). *A history of the English language*. Routledge.

Bello, A. and Cuervo, R. J. (1932). Gramática castellana.

Bello, A. et al. (1951). Gramática: gramática de la lengua castellana destinada al uso de los americanos.

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

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

Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036*.

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

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

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

Calero, R. C. (1981). *Problemas da língua galega*, volume 2. Sá da Costa Editora.

Capelo, R. G., Monteiro, A., Nunes, J., Rodrigues, A., Torgal, L., and Vitorino, F. (1994). *História de Portugal em datas*. Círculo de Leitores, Lisboa.

Carling, G., Larsson, F., Cathcart, C., Johansson, N., Holmer, A., Round, E., and Verhoeven, R. (2018). Diachronic atlas of comparative linguistics (diacl)—a database for ancient language typology. *PLoS ONE*, 13(10).

Carrera, A. (2014). O occitano do val d'aran: uma aproximação sociolinguística. *Quem fala a minha língua*, 2:63–93.

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

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

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

Chavula, C. and Suleman, H. (2020). Intercomprehension in retrieval: User perspectives on six related scarce resource languages. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 263–272.

Chen, S. F. and Goodman, J. (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 Miller, P. (2004). *Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages*. Discussion papers. IZA.

Collazo, S. D. (2014). O estándar galego: reintegracionismo vs. autonomismo. *Romanica Olomucensia*, (1):1–13.

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

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

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

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

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

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

Diez, X. C. L. (2008). Sobre a noção de galego-português. *Cadernos de Letras da UFF–Dossiê: Patrimônio cultural e latinidade*, 35:61–82.

Donoso, G. and Sánchez, D. (2017). Dialectometric analysis of language variation in twitter. *arXiv preprint arXiv:1702.06777*.

Dubert, F. and Sousa, X. (2016). On quantitative geolinguistics: an illustration from galician dialectology. *Dialectologia: revista electrònica*, pages 191–221.

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

Eden, S. E. (2018). *Measuring phonological distance between languages*. PhD thesis, UCL (University College London).

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

Freixeiro Mato, X. R. (2000). Gramática da lingua galega ii. morfosintaxe. *Vigo: A Nosa Terra*.

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

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

Gamallo, P., Pichel, J. R., and Alegria, I. (2020). Measuring language distance of isolated european languages. *Information*, 11(4):181.

Gamallo, P., Pichel, J. R., de Compostela, S., and Alegria, I. (2017b). A perplexity-based method for similar languages discrimination. *VarDial 2017*, page 109.

Gamallo, P., Sotelo, S., and Pichel, J. R. (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, Y., Liang, W., Shi, Y., and Huang, Q. (2014). Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393(C):579–589.

Goebl, H. (1982a). Ansätze zu einer computativen dialektometrie. *Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Handbücher zur Sprach-und Kommunikationswissenschaft*, 1:778–92.

Goebl, H. (1982b). Dialektometrie; prinzipien und methoden des einsatzes der numerischen taxonomie im bereich der dialektgeographie, volume 157 of philosophisch-historische klasse denkschriften. verlag der osterreichischen akademie der wissenschaften, vienna. with assistance of w. *D. Rase and H. Pudlatz*.

Goebl, H. (2006). Recent advances in salzburg dialectometry. *Literary and linguistic computing*, 21(4):411–435.

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.

Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H., Gonzalez-Rodriguez, J., and Moreno, P. J. (2014). Automatic language identification using long short-term memory recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*.

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

Gooskens, C., Nerbonne, J., Vaillette, N., et al. (2007). Conditional entropy measures intelligibility among related languages. *LOT Occasional Series*, 7:51–66.

Goutte, C., Léger, S., Malmasi, S., and Zampieri, M. (2016). Discriminating similar languages: Evaluations and explorations. *arXiv preprint arXiv:1610.00031*.

Haizhou, L., Min, Z., and Jian, S. (2004). A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting on association for Computational Linguistics*, page 159. Association for Computational Linguistics.

Han, A. L.-F., Lu, Y., Wong, D. F., Chao, L. S., He, L., and Xing, J. (2013). Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 365–372.

Heeringa, W. and Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, 13(3):375–400.

Heeringa, W. and Nerbonne, J. (2013). Dialectometry.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, Citeseer.

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

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

Jauhiainen, T. S., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Jones, M. C. and Mooney, D. (2017). *Creating orthographies for endangered languages*. Cambridge University Press.

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

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

Klarer, M. (2013). *An introduction to literary studies*. Routledge.

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

Knight, K. and Graehl, J. (1998). Machine transliteration. *Computational linguistics*, 24(4):599–612.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., and Verkerk, A. (2018). A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504.

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

Kroon, M., Medvedeva, M., and Plank, B. (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.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Lai, M., Patti, V., Ruffo, G., and Rosso, P. (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.

Lapa, M. R. (1973). A recuperação literária do galego. *Grial*, 11(41):278–287.

Lapesa, R. and Pidal, R. M. (1942). Historia de la lengua española.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., and Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144.

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

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., and Moreno, P. (2014). Automatic language

identification using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5337–5341. IEEE.

Lui, M. and Cook, P. (2013). Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 5–15.

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

Malvar, P., Pichel, J. R., Senra, Ó., Gamallo, P., and García, A. (2010). Vencendo a escassez de recursos computacionais. carvalho: Tradutor automático estatístico inglês-galego a partir do corpus paralelo europarl inglês-português. *Linguamática*, 2(2):31–38.

Mastin, L. (2011). The history of english.

Mattoso, J. and Ramos, R. (1994). *História de portugal*. Editorial Estampa.

Michael, L. D. (2015). A bayesian phylogenetic classification of tupí-guaraní. *LIAMES*, 15.

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

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2019). Overview for the second shared task on language identification in code-switched data. *arXiv preprint arXiv:1909.13016*.

Monteagudo, H. and Romero, H. M. (1999). *Historia social da lingua galega: idioma, sociedade e cultura a través do tempo*, volume 1. Editorial Galaxia.

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

Moura, A. d. C., López, A., and Pichel, J. R. (2008). Tmilg (tesouro medieval informatizado da lingua galega). *Procesamiento del lenguaje Natural*, (41):303–304.

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

Müller, A., Wichmann, S., Velupillai, V., Brown, C. H., Brown, P., Sauppe, S., Holman, E. W., Bakker, D., List, J.-M., Egorov, D., et al. (2010). Asjp world language tree of lexical similarity: Version 3 (july 2010). *Retrieved*, 10(19):2015.

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

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

Nerbonne, J. and Heeringa, W. (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.

Nerbonne, J., Heeringa, W., Van den Hout, E., Van der Kooi, P., Otten, S., Van de Vis, W., et al. (1996). Phonetic distance between dutch dialects. In *CLIN VI: Proceedings of the sixth CLIN meeting*, pages 185–202.

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

Nerbonne, J. and Kretzschmar, W. (2006). Progress in dialectometry: Toward explanation. *Literary and Linguistic Computing*, 21(4):387–397.

Nerbonne, J. and Kretzschmar Jr, W. A. (2013). Dialectometry++. *Literary and linguistic computing*, 28(1):2–12.

Nichols, J. and Warnow, T. J. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

Paz, R. M. (2008). *Historia de la lengua gallega*. Lincom Europa.

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

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

Pichel, J. R., Fernández, P. M., Gómez, O. S., Otero, P. G., and García, A. (2009). Carvalho: English-galician smt system from europarl english-portuguese parallel corpus. *Procesamiento del lenguaje natural*, (43):379–381.

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

Pichel, J. R., Gamallo, P., and Alegria, I. (2019a). Cross-lingual diachronic distance: Application to portuguese and spanish. *Procesamiento del Lenguaje Natural*, 63:77–84.

Pichel, J. R., Gamallo, P., and Alegria, I. (2019b). Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering*, pages 1–22.

Pichel, J. R., Gamallo, P., Alegria, I., and Neves, M. (2020a). A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, pages 1–31.

Pichel, J. R., Gamallo, P., Neves, M., and Alegria, I. (2020b). Distância diacrónica automática entre variantes diatópicas do português e do espanhol. *Linguamática*, 12(1):117–126.

Pichel Campos, J. R. and Fagim, V. (2012). O galego é uma oportunidade. *El gallego es una oportunidad*.

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

Porta, J., Sancho, J.-L., and Gómez, J. (2013). Edit transducers for spelling variation in old spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, number 087, pages 70–79. Linköping University Electronic Press.

Purver, M. (2014). A simple baseline for discriminating similar languages. Association for Computational Linguistics.

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

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

Reynaert, M., Hendrickx, I., and Marquilhas, R. (2012). Historical spelling normalization. a comparison of two statistical methods: Ticcl and vard2. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pages 87–98.

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

Rissanen, M., Kytö, M., and Palander-Collin, M. (1993b). *Early English in the computer age: Explorations through the Helsinki Corpus*. Number 11. Walter de Gruyter.

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

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

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

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

Saraiva, A. J. (2001). *História da literatura portuguesa*. Porto: Porto Editora, 2001.

Saraiva, J. H. (1978). *História concisa de Portugal*. Publ. Europa-América.

Satapathy, R., Guerreiro, C., Chaturvedi, I., and Cambria, E. (2017). Phonetic-based microtext normalization for twitter sentiment analysis. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 407–413. IEEE.

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.

Schneider, P. (2002). Computer assisted spelling normalization of 18th century english. In *New Frontiers of Corpus Research*, pages 199–211. Brill Rodopi.

Séguy, J. (1971). *La relation entre la distance spatiale et la distance lexicale*. G. Straka, Palais de l'université.

Sennrich, R. (2012a). 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.

Sennrich, R. (2012b). 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*, pages 539–549. Association for Computational Linguistics.

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

Singh, A. K. and Surana, H. (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, J. (2003). *An historical study of English: Function, form and change*. Routledge.

Specia, L., Scarton, C., and Paetzold, G. H. (2018). Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

SuanzeS-Carpegna, J. V. (2013). La constitución de 1931 y la organización terrirorial del estado. *Iura vasconiae: revista de derecho histórico y autonómico de Vasconia*, (10):323–354.

Suzuki, I., Mikami, Y., Ohsato, A., and Chubachi, Y. (2002). A language and character set determination method based on n-gram statistics. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(3):269–278.

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

Szmrecsanyi, B. (2008). Corpus-based dialectometry: Aggregate morphosyntactic variability in british english dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):279–296.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: a methodological sketch. *Corpora*, 6(1):45–76.

Tang, G., Cap, F., Pettersson, E., and Nivre, J. (2018). An evaluation of neural machine translation models on historical spelling normalization. *arXiv preprint arXiv:1806.05210*.

Teyssier, P. (1982). História da língua portuguesa.

Tiedemann, J. and Ljubešić, N. (2012). Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634.

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

Trask, R. L. (1995). Origins and relatives of the basque language: Review of the evidence. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 65–100.

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

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

Vilavedra, D. and Fdez, V. F. V. (1999). *Historia da literatura galega*, volume 2. Editorial Galaxia.

Villares, R. (2004). *Historia de Galicia*, volume 6. Editorial Galaxia.

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

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

Wieling, M. and Nerbonne, J. (2015). Advances in dialectometry.

Williams, E. B. (1962). *From Latin to Portuguese: Historical Phonology and Morphology of the Portugese Language*. Univ. Pennsylvania Press.

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

Yujian, L. and Bo, L. (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*.

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

Zampieri, M., Gebre, B. G., Costa, H., and Van Genabith, J. (2015). Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 66–72.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN*, volume 2, pages 580–587.

Zampieri, M., Malmasi, S., and Dras, M. (2016). Modeling language change in historical corpora: the case of portuguese. *arXiv preprint arXiv:1610.00030*.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., Scherrer, Y., Samardžić, T., Ljubešić, N., Tiedemann, J., 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.

Zampieri, M., Malmasi, S., Scherrer, Y., Samardzic, T., Tyers, F., Silfverberg, M., Klyueva, N., Pan, T.-L., Huang, C.-R., Ionescu, R. T., et al. (2019). A report on the third vardial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16.

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.

Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V. (2016). Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4):729–766.

Zubiaga, A., Vicente, I. S., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V. (2015). Tweetlid: a benchmark

for tweet language identification. *Language Resources and Evaluation*, pages 1–38.