Proposta recebida em Janeiro 2017 e aceite para publicação em Maio 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 & Trask, 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, 1997), 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.

DOI: 10.21814/lm.12.1.319

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 de Cavnar & Trenkle (1994) e Dunning (1994), que são os primeiros 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., 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 caracteres 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., 2009), 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 (1997) 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 & Borin, 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 divididos 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

[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)

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] e *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 (1984) 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] e *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:

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

onde as probabilidades de n-grama $P(\cdot)$ são definidas desta forma:

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

[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

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:

$$PLD(VL1.1, VL1.2) = \frac{PP(A) + PP(B)}{2} \quad (3)$$

$$PP(A) = PP(CH_{VL1.2}, LM_{VL1.1}) \quad (4)$$

$$PP(B) = PP(CH_{VL1.1}, LM_{VL1.2}) \quad (5)$$

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:

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

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 nos textos com a finalidade de unificar ortograficamente os textos das variedades do português europeu e do Brasil, e também da variedade do espanhol europeu e do 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 as Figuras 1 e 2 retratam a informação da distância entre as variedades diatópicas do português europeu e do português do Brasil, e do espanhol europeu e o espanhol da Argentina.

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

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.

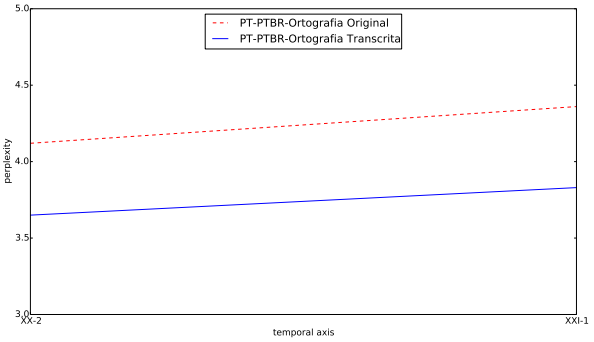

Figura 1: *CrossDiaDist* entre o português europeu e o português da Brasil através do eixo temporal em OS e TS.

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 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.

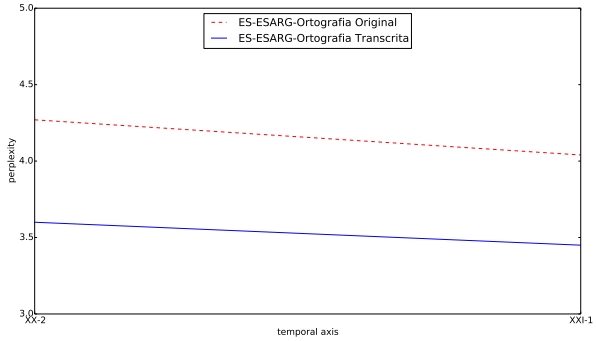

Figura 2: *CrossDiaDist* entre o espanhol europeu e o espanhol da Argentina através do eixo temporal em OS e TS.

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 *Workshop on Multilingual and Cross-lingual Methods in NLP*, 65–74. 10.18653/v1/W16-1208.

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. 10.1515/LITY.2009.009.

Barbançon, François, Steven N. Evans, Luay Nakhleh, Don Ringe & Tandy Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30(2). 143–170. 10.1075/dia.30.2.01bar.

Bello, Andrés. 1984. *Gramática de la lengua castellana*. EDAF.

Bello, Andrés et al. 1951. *Gramática: gramática de la lengua castellana destinada al uso de los americanos*. Caracas: Ministerio de Educación.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and linguistic Computing* 8(4). 243–257. 10.1093/llc/8.4.243.

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupilla. 2009. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals* 61(4). 285–308. 10.1524/stuf.2008.0026.

Cavnar, William B & John M Trenkle. 1994. N-gram-based text categorization. Em *3rd annual symposium on document analysis and information retrieval*, 161–175.

Chen, Stanley F. & Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. Em *34th Annual Meeting on Association for Computational Linguistics*, 310–318. 10.3115/981863.981904.

Chiswick, Barry R. & Paul W. Miller. 2004. *Linguistic distance: A quantitative measure of the distance between english and other languages*. Bonn: IZA Discussion Papers.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis & Elke Teich. 2016. An information-theoretic approach to modeling diachronic change in scientific english. Em *From Data to Evidence in English Language Research*, 258–281. Brill. 10.1163/9789004390652_012.

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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, I/833–I/836. 10.1109/ICASSP.2005.1415243.

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 *3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 170–177.

Gamallo, Pablo, Marcos Garcia, 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*, 12–16.

Gamallo, Pablo, José Ramom Pichel & Iñaki Alegria. 2017a. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications* 484. 152–162. 10.1016/j.physa.2017.05.011.

Gamallo, Pablo, Jose Ramom Pichel, Santiago de Compostela & Inaki Alegria. 2017b. A perplexity-based method for similar languages discrimination. *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* 109–114. 10.18653/v1/W17-1213.

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. 579–589. 10.1016/j.physa.2013.08.075.

González, Meritxell. 2015. An analysis of Twitter corpora and the differences between formal and colloquial tweets. Em *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 *15th 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 *8th Workshop on Statistical Machine Translation*, 365–372.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Muller & Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42(3–4). 331–354. 10.1515/FLIN.2008.331.

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. 10.1121/1.2016299.

Kondrak, Grzegorz. 2005. N-gram similarity and distance. Em *International Symposium on String Processing and Information Retrieval (SPIRE)*, 115–126. 10.1007/11575832_13.

Kroon, Martin, Masha Medvedeva & Barbara Plank. 2018. When simple n-gram models outperform syntactic approaches: Discriminating between Dutch and Flemish. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 244–253.

Liu, HaiTao & Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin* 58. 1139–1144. 10.1007/s11434-013-5711-8.

Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez & Pedro Moreno. 2014. Automatic language identification using deep neural networks. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5337–5341. 10.1109/ICASSP.2014.6854622.

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 *3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, 1–14.

Millar, Robert McColl & Larry Trask. 2015. *Trask's historical linguistics*. Abington, UK: 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. 1997. Measuring dialect distance phonetically. Em *3rd Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, 11–18.

Petroni, Filippo & Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389(11). 2280–2283. 10.1016/j.physa.2010.02.004.

Pichel, José Ramom, Pablo Gamallo & Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to european portuguese. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 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. 10.1017/S1351324919000378.

Rama, Taraka & Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. Em *Sequences in Language and Text*, 171–200. De Gruyter Mouton. 10.1515/9783110362879-012.

Rissanen, Matti, Merja Kytö & Minna Palander-Collin. 1993. *Early english in the computer age: Explorations through the helsinki corpus*. Berlin: De Gruyter Mouton.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. Em *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 539–549.

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 (PROPOR)*, 121–127. 10.1007/978-3-642-28885-2_14.

Singh, Anil Kumar & Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? Em *9th meeting of the ACL special interest group in computational morphology and phonology*, 40–47.

Specia, Lucia, Carolina Scarton & Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies* 11(1). 1–162. 10.2200/S00854ED1V01Y201805HLT039.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. *American Philosophical Society* 96(4). 452–463.

Tiedemann, Jörg & Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. Em *International Conference on Computational Linguistics (COLING)*, 2619–2634.

Yujian, Li & Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29(6). 1091–1095. 10.1109/TPAMI.2007.1078.

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 *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 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 *6th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 1–16. 10.18653/v1/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. 729–766. 10.1007/s10579-015-9317-4.