Proposta recebida em 21 de Março 2018 e aceite para publicação em 16 de Julho 2018.

Estratégias Lexicométricas para Detetar Especificidades Textuais

Lexicometric strategies to detect textual specificities

Álvaro Iriarte
Universidade do Minho
Grupo Galabra-UMinho
alvaro@ilch.uminho.pt

Pablo Gamallo
Universidade de Santiago de Compostela
CiTIUSiTIUS-USC
pablo.gamallo@usc.es

Alberto Simões
2Ai — Polytechnic Institute of Cávado and Ave
Barcelos, Portugal
Grupo Galabra-UMinho
asimoes@ipca.pt

Resumo

Neste artigo propomo-nos a definir e desenvolver uma estratégia automática para procurar especificidades lexicais dentro de conjuntos de textos utilizando unidades lexicais simples e expressões com várias palavras, ou termos multipalavra (MWE, a sua sigla em inglês).

Propomos uma metodologia para o cálculo da divergência de distribuições de lemas e de MWE que permitirá encontrar, automaticamente, diferenças e semelhanças entre textos não anotados. Esta metodologia poderá ser utilizada para posteriormente identificar grupos de textos sobre os quais se procederá a análises quantitativas e qualitativas semiautomáticas e/ou com intervenção humana.

Num primeiro teste, utilizamos dois textos de especialidade (da área da pediatria) e um texto literário, presumindo que os textos de especialidade deveriam apresentar maiores divergências relativamente ao texto literário do que entre eles próprios. Como os testes feitos mostraram a tendência esperada, decidimos aplicar a mesma metodologia a um segundo grupo de textos (três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela).

Palavras chave

divergencia de Kullback-Leibler, divergência lexical, lexicometria

Abstract

In this article we propose to to define and develop an automatic strategy to search for lexical specificities within sets of texts using simple lexical units and multiword expressions (MWE).

We propose a methodology for calculating the divergence of lemma and MWE distributions that will automatically find differences and similarities between unlabeled texts. This methodology can be used to subsequently identify groups of texts to which quantitative and qualitative analyzes will be applied (semiautomatically and/or with human intervention).

In a first test, we used two specialized texts (from the area of Paediatrics) and a literary text, assuming that the texts of specialty should present greater divergences with respect to the literary text than among themselves. As the tests that were done showed the expected trend, we decided to apply the same methodology to a second set of texts (three sets of interviews done to visitors in the city of Santiago de Compostela).

Keywords

Kullback–Leibler divergence, lexical divergence, lexicometry

1 Introdução

Dentro das Ciências Humanas e Sociais e mais concretamente nas Humanidades Digitais, há uma necessidade cada vez maior de ter acesso a ferramentas computacionais e estatísticas que permitam detetar semelhanças e diferenças entre grupos de textos (Kilgarriff, 1996) ou medir a riqueza lexical dos mesmos (Tweedie & Baayen, 1998). As análises quantitativas baseadas na distribuição de traços linguísticos são essenciais para o desenvolvimento de trabalhos linguísticos e sociolinguísticos que procuram especificidades e diferenças em textos de natureza e origem diversas. Dentro dos traços linguísticos, têm especial relevância as características lexicais dos textos.

Propomo-nos a definir e desenvolver uma estratégia automática para procurar especificidades linguísticas, nomeadamente especificidades lexicais, dentro de conjuntos de textos. O léxico

DOI: 10.21814/lm.10.1.263

utilizado por diferentes indivíduos pode diferir substancialmente segundo as propriedades e características dos mesmos, incluindo, como veremos, género, profissão, estudos, etc. O nosso objetivo não é tanto identificar especificidades textuais em relação ao conteúdo específico do texto nem ao seu estilo (tamanho de frases e palavras, etc.), mas sim em relação ao uso de unidades lexicais simples (lemas/palavras) e termos multipalavra (MWE), entendidos aqui como combinações lexicais, n-grams ou cadeias de palavras (Maia et al., 2008; Stubbs & Barth, 2003) e não no sentido de combinações lexicais restritas, mais frequente dentro da linguistica e da lexicografia (Mel'čuk et al., 1995), combinações lexicais que deveriam funcionar melhor do que as palavras simples ou os lemas, para detetar divergências e convergências textuais, porque deveriam apresentar valores de divergência maiores.

Propomos uma metodologia para o cálculo da divergência de distribuições de lemas e de MWE que permitirá encontrar, automaticamente, diferenças e semelhanças entre textos não anotados. Esta metodologia poderá ser utilizada para posteriormente identificar grupos de textos diferenciados sobre os quais se procederá a análises quantitativas e qualitativas semiautomáticas e/ou com intervenção humana. A identificação destes conjuntos de textos com maior grau de convergência poderá, assim, ser feita sem nenhum tipo de critério ou conhecimento prévio, como o que está a ser utilizado nos testes do presente artigo (*tradução literária vs. texto técnico*; *entrevistas a universitários vs. entrevistas a não universitários*, etc., permitindo assim abordagens, nas análises referidas, com menos riscos de vieses cognitivos ou até de preconceitos.

As nossas hipóteses de partida foram:

1. A divergência de Kullback-Leibler (divergência KL) permite comparar distribuições de palavras e MWE, o que poderá ser usado para comparar automaticamente textos não anotados previamente;
2. O uso de combinações lexicais para detetar divergências e convergências textuais deveria funcionar melhor do que o uso de palavras simples, porque apresentará valores de divergência maiores.

Para levar a cabo o objetivo acima sublinhado, desenvolveremos um método concreto de cálculo da divergência lexical entre textos com base, principalmente, na extração de MWE. Pensamos que esta abordagem poderá acrescentar valor às análises lexicométricas com base na unidade palavra. Uma vez que estas, as palavras, não funcionam como unidades isoladas (Saussure, 1999; Iriarte, 2001), consideramos uma mais-valia para os trabalhos de lexicometria e textometria o facto de ultrapassar a palavra como unidade de análise e descrição linguística. É no mínimo estranho que, com o surgimento de ferramentas informáticas que permitiram abordagens linguísticas mais empiristas, se continue a trabalhar com categorias gramaticais (PoS) já documentadas pelos gregos no ano 100 a.C. (Robins, 1997).

O artigo organiza-se da seguinte maneira: trabalho relacionado (secção 2), descrição do método (secção 3), experiências (secção 4) e conclusões.

2 Trabalho relacionado

O presente artigo presenta uma estratégia para a comparação textual. Existem numerosos métodos cujo objetivo é também a comparação quantitativa entre textos, alguns deles centrados no estilo formal e outros no conteúdo textual. Entre os métodos estilísticos e formais, um dos mais comuns é o que calcula a diversidade lexical a partir de uma família de medidas baseadas na relação entre o número de unidades lexicais e o tamanho total do texto, sendo a mais básica a ratio entre tipos e tokens, chamada *TTR* (McCarthy & Jarvis, 2010; Fergadiotis et al., 2013). Este método permite medir a riqueza lexical dos textos, mas não estabelece a comparação entre eles com base no tipo de unidades lexicais utilizadas. Outros métodos estilísticos centram-se na legibilidade e complexidade dos textos com base no cômputo do tamanho das frases (percentagem de palavras por oração) e das próprias palavras (percentagem de sílabas por palavras) (Loughran & McDonald, 2013), como o que deu lugar ao teste de legibilidade chamado *Flesch-Kincaid*, que mede a dificuldade de um texto para ser compreendido no processo de leitura. Em comparação com os estilísticos e formais, os métodos focados no conteúdo semântico dos textos calculam a similaridade textual com base em modelos distribucionais e *embeddings*, que por sua vez foram inspirados pelos métodos que calculam a similaridade semântica entre orações. As palavras são representadas como vetores de contextos e os documentos (ou pequenos extratos de textos) são modelados como a soma desses vetores. A comparação entre os extratos textuais é levada a cabo mediante medidas de similaridade entre vetores, sendo a mais comum a do *coseno* (Agirre et al., 2016; Mikolov et al., 2013). A diversidade lexical e a legibilidade são estratégias quantita-

tivas muito genéricas centradas no estilo e na forma, enquanto a similaridade distribucional é um método muito mais específico ao focar-se no conteúdo. O método proposto no presente artigo, centrado na divergência lexical mediante lemas e MWE, situa-se a um nível intermédio entre o formalismo estilístico e o conteúdo textual.

Outros trabalhos mais próximos do nosso exploram traços linguísticos concretos —por exemplo uso de pronomes pessoais, de palavras com polaridade, de modificadores nominais, etc.—, com o intuito de detetar características textuais próprias de grupos sociais: homens/mulheres, jovens, etc. (Argamon et al., 2003)

Por outro lado, e já fora do âmbito do PLN, o trabalho da Rede Galabra[1] centra-se, de maneira especial, em projetos de investigação relacionados com os discursos e práticas culturais na comunidade e os seus impactos, nas suas dimensões económicas, ambientais, socioculturais ou simbólicas (Torres Feijó, 2015; Pazos-Justo et al., 2018)

As nossas responsabilidades, dentro dos referidos projetos, estão relacionadas com o tratamento linguístico (nomeadamente a extração terminológica e a análise lexicométrica) dos *corpora* constituídos: inquéritos, entrevistas, gravações de grupos de discussão e *corpus* documental já catalogado (*vd. infra*).

Utilizando as mesmas orientações metodológicas e o mesmo *corpus*, tentar-se-á replicar alguns resultados dos projetos da Rede Galabra (finalizados e em curso)[2]. Reivindicar a replicação dos resultados na área das Ciências Humanas e Sociais (CHS) é de suma importância, uma vez que é uma prática pouco frequente, o que impede consolidar e validar muitas das nossas pesquisas em CHS como investigação dita científica.

As nossas tarefas consistem em explorar, de maneira automática ou semiautomática, todo o potencial dos inquéritos e entrevistas feitos aos visitantes, bem como a importante base de dados composta pelos produtos literários e culturais já catalogados, mediante o tratamento estatístico e linguístico do *corpus*, neste caso concreto, com a extração terminológica (da base de dados documental já construída e das entrevistas já realizadas), focando, de maneira especial, o que podemos chamar, de maneira genérica, *termos multipalavra (MWE)*, que corresponderão, no nosso trabalho, ao que conhecemos como expressões idiomáticas (*deitar foguetes antes da festa*; *to ask for the moon*), colocações (*ódio mortal*; *bitter hatred*), quase-frasemas (*cartão vermelho*; *black belt*) ou entidades nomeadas (*Santiago de Compostela*, *Cavaleiros da Ordem de Santiago*, *25 de Julho*, etc.) mas também outras combinações lexicais frequentes, não necessariamente restritas (Mel'čuk et al., 1995).

À partida, o uso de MWE deveria ser mais eficaz do que o uso de palavras simples (mesmo que previamente selecionadas) permitindo poder trabalhar com *corpora* não anotados. Por exemplo, em análises posteriores de outros trabalhos do projeto, a palavra *água* será contabilizada na categoria *Gastronomia* quando ocorre em combinações como *beber água*, *água de mesa*, *água mineral*, etc., mas não em *cair na água* ou *água do rio*, por exemplo. Esta última ocorrência, porém, seria contabilizada (erradamente) ao utilizarmos a unidade palavra.

Outro exemplo: formas do adjetivo *caro* que ocorrem nas combinações *preços caros*, *uma cidade cara*, etc. serão contabilizadas na categoria *Economia* ao usarmos MWE com anotação morfossintáctica, evitando a contagem de formas como *cara a cara*, *dar a cara*, *caro colega*, *fazer-se caro*, etc.

3 Descrição do método

Para além da comparação direta de frequências da mesma palavra em dois *corpora*, é possível comparar toda a distribuição das frequências, como um todo. Para isso foi usada a Divergência de Kullback-Leibler (Kullback & Leibler, 1951)

[1] https://redegalabra.org/

[2] Entre outros:

- Discursos, imágenes y prácticas culturales sobre Santiago de Compostela como meta de los Caminos de Santiago. Projeto de 3 anos de duração financiado pela Subdirección General de Proyectos de Investigación. Dirección General de Investigación Científica y Técnica. Ministerio de Economía y Competitividad. Gobierno de España [Código: FFI2012-35521] (2012-2015); https://redegalabra.org/discursos-imagens-e-praticas-culturais-sobre-santiago-de-compostela-como-meta-dos-caminhos

- Bienestar de la comunidad local através de narrativas y usos culturales: Santiago y el Camino actual [em curso];

- Discursos sobre Santiago de Compostela y el/los Camino(s) de Santiago en la novela española actual (2010) a través de técnicas analíticas digitales: Posibilidades y valor del conocimiento generado [Tese de doutoramento de María Luisa Fernández, orientada por Elias J. Feijó e Roberto Samartim (Grupo Galabra da Univ. de Santiago de Compostela e Grupo Galabra-UMinho)];

- "Narrativas, usos e consumos de visitantes como aliados ou amenaças para o bem-estar da comunidade local: o caso de Santiago de Compostela" (Ref: FFI2017-88196-R), parcialmente subsidiado pelo Ministerio de Industria, Economía y Competitividad do Governo da Espanha no quadro do Programa Estatal de I+D+I Orientada a los Retos de la Sociedad (2018-2021)

que, dada a distribuição de dois corpos diferentes (P_{c_i} e P_{c_j}) pode ser definida por:

$$D_{\mathrm{KL}}(P_{c_i}||P_{c_j}) = \sum_k F_{c_i}(k) \log \frac{F_{c_i}(k)}{F_{c_j}(k)} \quad (1)$$

onde $F_{c_i}(k)$ é a probabilidade (frequência relativa) de palavra k no *corpus* c_i.

A equação 1 permite obter uma medida de quanto a distribuição P_{c_j} se distancia da distribuição P_{c_i}, tomando em conta as probabilidades (ou frequências relativas) das palavras de cada corpus. Para as análises aqui apresentadas foi usada uma implementação em Perl `Math::KullbackLeibler::Discrete` de um dos autores.[3]

4 Testes

4.1 Objetivos

O nosso objetivo é utilizar a divergência KL para calcular graus de especificidade ou de convergência lexical entre grupos de textos, sem intervenção humana e sem necessidade de trabalhar com *corpora* anotados.

Numa primeira experiência comparamos três textos: dois textos de especialidade (da área da pediatria) e um texto literário (a tradução para o espanhol do romance *Ensaio sobre a Cegueira*, de José Saramago), sabendo à partida que as divergências deverão ser maiores entre qualquer um dos textos de especialidade e o texto literário.

Na segunda experiência, aplicaremos a divergência KL para calcular graus de especificidade ou de convergência lexical entre vários reagrupamentos das entrevistas realizadas a 24 visitantes da cidade de Santiago de Compostela.

Como foi referido, as nossas hipóteses de partida foram:

1. A divergência de Kullback-Leibler (divergência KL) permite comparar distribuições de palavras e MWE, o que poderá ser usado para comparar automaticamente textos não anotados previamente;
2. O uso de combinações lexicais para detetar divergências e convergências textuais deveria funcionar melhor do que o uso de palavras simples, porque deveria apresentar valores de divergência maiores.

Deixamos para futuros trabalhos, com colegas de outras áreas da rede Galabra, as análises relacionais e contrastivas (do ponto de vista qualitativo e quantitativo) destes mesmos lemas e combinações lexicais extraídos das transcrições das entrevistas.

4.2 *Corpus*

A base de dados utilizada pela Rede Galabra disponibiliza o acesso a informação retirada de um *corpus* documental e a um conjunto de inquéritos e entrevistas[4].

Uma vez que, neste momento falta ainda por finalizar o processo de transcrição das entrevistas a portugueses e brasileiros, o trabalho aqui apresentado é realizado apenas sobre 24 entrevistas em castelhano (para além dos três textos utilizados na primeira experiência: dois da área da pediatria e um texto literário).

Com base nos dados disponíveis nos inquéritos, relativos às mesmas pessoas entrevistadas (idade, género, nível de estudos, e autoidentificação como peregrinos ou como turistas), subdividimos as entrevistas nos seguintes subgrupos[5]:

1. Autoidentificação
   - Peregrinos (11 entrevistas)
   - Não peregrinos (13 entrevistas)
2. Nível de estudos
   - Universitários (16 entrevistas)
   - Não universitários (8 entrevistas)
3. Género
   - Mulheres (11 entrevistas)
   - Homens (13 entrevistas)

[3] https://github.com/ambs/Math-KullbackLeibler-Discrete

[4] O *corpus* documental disponibiliza aos investigadores do grupo uma base de dados avançada com 560 livros catalogados (Samartim, 2015), procedente de produtos culturais e literários publicados entre 2008 e 2012. O *corpus* foi limitado às produções culturais efetivamente consumidas pelos turistas procedentes da Galiza, Espanha, Portugal e Brasil desde 2008 (Portugal e Brasil são os países de procedência do maior número de visitantes não espanhóis e não comunitários respetivamente).

O *corpus vivo* é constituído por inquéritos e entrevistas a visitantes, comerciantes e comunidade local. Os inquéritos, num total de 2157, foram realizados entre 27/03/2013 e 26/03/2014 a turistas galegos e espanhóis (1323), portugueses (428) e brasileiros (406). Foram gravadas 41 entrevistas a turistas galegos e espanhóis, 59 entrevistas a turistas portugueses e 56 entrevistas a turistas brasileiros.

[5] A divisão por grupos etários será excluída, para o presente trabalho, devido ao reduzido tamanho dos 4 grupos etários estabelecidos no projeto "Discursos, imagens e práticas culturais sobre Santiago de Compostela como meta dos Caminhos": Idade < 30; $30 \leq$ Idade < 45; $45 \leq$ Idade < 69; $70 \leq$ idade.

Antes de calcular o grau de divergência lexical entre os seis conjuntos de entrevistas, foi feito um teste prévio com as frequências dos lemas e das MWE extraídos dos dois textos[6] de especialidade (da área da pediatria). e um texto literário (a tradução para o espanhol do romance *Ensaio sobre a Cegueira*, de José Saramago), sabendo à partida, como já referimos, que os dois textos de pediatria deveriam apresentar maior convergência entre si.

Atendendo à Lei de Zifp, as distribuições de palavras baseadas em frequências relativas produzem diferentes escalas de valores com textos de diferentes tamanhos. Para podermos trabalhar com textos de tamanho semelhante e assim comparar os valores usando frequências absolutas, reduzimos os tamanhos dos conjuntos dos textos (as entrevistas e os dois textos utilizados no primeiro teste) ao tamanho do documento mais pequeno de cada grupo. Assim, reduzimos todos os grupos de entrevistas ao tamanho do conjunto de entrevistas feitas a não universitários (64 751 palavras) e o tamanho dos três textos utilizados no primeiro teste, ao tamanho de um dos textos da especialidade de pediatria (25 998 palavras).

O tamanho original dos grupos de entrevistas e dos textos utilizados no primeiro teste são descritos na tabela 1.

texto	*# palavras*
11 entrevistas a peregrinos	71 652
13 entrevistas a não peregrinos	116 285
16 entrevistas a universitários	123 186
8 entrevistas a não universitários	64 751
11 entrevistas a mulheres	92 650
13 entrevistas a homens	102 497
Texto de *Pediatría 1*	35 713
Texto de *Pediatría 2*	25 998
Texto do romance *Ensayo sobre la Ceguera*	107 296

Tabela 1: Tamanhos originais dos textos analisados.

Para os propósitos dos testes aqui realizados, pensamos ser irrelevante o facto de termos cortado os textos de maneira aleatória.

A partir destes conjuntos de textos, foram extraídos os lemas e as MWE com as respetivas frequências e construídas as correspondentes matrizes usadas no cálculo das divergências. A extração de lemas e MWE foi feita com os módulos correspondentes da ferramenta LinguaKit (Gamallo & Garcia, 2017)[7]. O anotador morfossintáctico (que inclui o lematizador) integrado no LinguaKit foi avaliado para três línguas, nomeadamente inglês, português e espanhol, com resultados próximos do estado da arte: ≈ 96% para português e espanhol, e ligeiramente mais baixos (≈ 94%) para inglês (Gamallo et al., 2015; Garcia & Gamallo, 2015). Quanto ao extrator de MWE, foi descrito e avaliado qualitativamente em (Gamallo & Garcia, 2017).

Na Tabela 2 apresentamos alguns dados quantitativos relativos aos resultados da extração de lemas e de MWE dos dois textos de especialidade e do texto literário referidos na secção 4.2

	lemas	*lemas (total >1)*	*MWE*	*MWE (total >1)*
Total	4754	2495	6725	798
Ceguera	2209	1371	1226	72
Pediatría 1	2089	1456	2698	419
Pediatría 2	2273	1485	2905	411

Tabela 2: Nº de lemas e MWE extraídos (primeiro teste).

Na Tabela 3 apresentamos dados quantitativos relativos aos resultados da extração de lemas e de MWE da transcrição das 24 entrevistas referidas *supra*.

	lemas	*lemas (total >1)*	*MWE*	*MWE (total >1)*
Total	3907	3370	4271	3145
Mulheres	2203	2062	1526	1280
Homens	2319	2293	1669	1661
Universitários	2211	2165	1679	1553
Não universitários	2204	2124	1658	1441
Peregrinos	2261	2017	1712	1183
Não peregrinos	2281	2281	1687	1687

Tabela 3: Nº de lemas e MWE extraídos das entrevistas.

4.3 Primeiro teste: texto científico *vs.* texto literário

Nesta primeira experiência, a partir das matrizes com as frequências dos lemas e dos MWE extraídos dos dois textos de especialidade e do texto literário referidos na secção 4.2, calculamos o grau de divergência lexical entre os mesmos, presumindo, como dissemos, que os dois textos da área de especialidade deveriam apresentar maior convergência entre si e que as divergências deveriam ser maiores entre estes e o texto literário.

Dado tratar-se de uma divergência, para um par de documentos (d_1, d_2) foi calculada a média das divergências das suas distribuições $D_{\mathrm{KL}}(P_{d_1}||P_{d_2})$ e $D_{\mathrm{KL}}(P_{d_2}||P_{d_2})$.

[6] Cifuentes, Javier & Ventura-Juncá, Patricio (2001). *Manual de Pediatría*. Retrieved January 16, 2018, from http://botica.com.ve/PDF/6mlped.pdf;
Cerrolaza, Javier, Mercé, Luis. & Emilio Jardón, Emílio (2008). *1. Consideraciones generales 5 Consideraciones clínicas previas 6*. Retrieved January 16, 2018, from http://www.espanito.com/1-consideraciones-generales-5-consideraciones-clnicas-previas.html

[7] https://github.com/citiususc/Linguakit

Os resultados são apresentados em duas colunas, sendo que a segunda corresponde aos resultados de frequências > 1 .

Como veremos, todas as configurações devolvem resultados consistentes entre elas.

4.3.1 Cálculo de divergências usando lemas

Na Tabela 4 apresentamos os resultados da comparação dos dados relativos às listas de frequências dos lemas extraídos de cada um dos textos, comparados dois a dois.

Considerando que a divergência de Kullback-Leibler é nula para duas distribuições idênticas, pode-se concluir, como esperado, que as maiores divergências aparecem entre o texto literário e cada um dos textos de especialidade.

	Lemas	*Lemas(freq > 1)*
Ceguera - Pediatría 1	3,1445	2,8761
Ceguera - Pediatría 2	3,3874	3,0859
Pediatría 1 - Pediatría 2	1,9542	1,6351

Tabela 4: Cálculo de divergências usando lemas (primeiro teste).

4.3.2 Cálculo de divergências usando MWE

Na Tabela 5 apresentamos os resultados da comparação dos dados relativos às listas de frequências das MWE extraídas de cada um dos textos, comparados dois a dois.

Também aqui, como esperado, as maiores divergências aparecem, novamente, entre o texto literário e cada um dos textos de especialidade.

	MWE	*MWE(freq > 1)*
Ceguera - Pediatría 1	13,3517	15,6336
Ceguera - Pediatría 2	13,3193	15,6978
Pediatría 1 - Pediatría 2	12,1861	12,3519

Tabela 5: Cálculo de divergências usando MWE (primeiro teste).

4.4 Segundo teste: entrevistas

Na segunda experiência, a partir das matrizes com as frequências dos lemas e das MWE extraídos da transcrição de 24 entrevistas realizadas, entre 27/03/2013 e 26/03/2014, a 24 pessoas que visitaram a cidade de Santiago de Compostela, calculamos o grau de divergência lexical entre os seis conjuntos de entrevistas já referidos (peregrinos *vs.* não peregrinos; universitários *vs.* não universitários; mulheres *vs.* homens).

Como no caso anterior, dado tratar-se de uma divergência, para um par de documentos (d_1, d_2) foi calculada a média das divergências das suas distribuições $D_{\mathrm{KL}}(P_{d_1}||P_{d_2})$ e $D_{\mathrm{KL}}(P_{d_2}||P_{d_2})$.

Com os conjuntos de entrevistas estudados, o esperado é que as maiores divergências apareçam entre os grupos que se opõem diretamente: entrevistas a mulheres *vs.* entrevistas a homens; entrevistas a peregrinos *vs.* entrevistas a não peregrinos; entrevistas a universitários *vs.* entrevistas a não universitários. Como veremos, todas as configurações devolvem resultados consistentes entre elas.

4.4.1 Cálculo de divergências usando lemas

Na Tabela 6 apresentamos os resultados da comparação dos dados relativos às listas de frequências dos lemas extraídos de cada um dos seis grupos de entrevistas comparados dois a dois.

Pode-se concluir que, ao compararmos os grupos de entrevistas dois a dois, as maiores divergências aparecem entre os grupos que se opõem diretamente: homem–mulher; peregrino–não peregrino; universitário–não universitário.

	Lemas	*Lemas(freq > 1)*
mulheres – homens	**0,5059**	**0,48865**
mulheres – Univer.	0,389	0,3696
mulheres – NãoUniver.	0,2818	0,25865
mulheres – Peregr.	0,3057	0,26615
mulheres – NãoPeregr.	0,41205	0,39725
homens – Univer.	0,16135	0,1539
homens – NãoUniv	0,37225	0,36145
homens – Peregr.	0,4111	0,3841
homens – NãoPeregr.	0,10655	0,10375
Univers. – NãoUniver.	**0,50035**	**0,48785**
Univer. – Peregr.	0,3924	0,36345
Univer. – NãoPeregr	0,13005	0,12535
NãoUniver. – Peregr.	0,30215	0,26945
NãoUniver. – NãoPeregr.	0,36475	0,3567
Peregr. - NãoPeregr.	**0,4699**	**0,44575**

Tabela 6: Cálculo divergências usando lemas.

4.4.2 Cálculo de divergências usando MWE

Na Tabela 7 apresentamos os resultados da comparação dos dados relativos às listas de frequências das MWE extraídas de cada um dos seis grupos de entrevistas, comparados dois a dois.

Neste caso, também se pode concluir que, ao compararmos os grupos de entrevistas dois a dois, as maiores divergências aparecem entre os grupos que se opõem diretamente: homem–mulher;

peregrino–não peregrino; universitário–não universitário.

	MWE	*MWE(freq > 1)*
mulheres – homens	**11,7211**	**11,6629**
mulheres – Univer.	9,4818	9,1155
mulheres – NãoUniver.	6,77665	5,88005
mulheres – Peregr.	7,82915	6,5231
mulheres – NãoPeregr.	9,6372	9,38765
homens – Univer.	3,2259	2,86875
homens – NãoUniver.	8,57	8,3035
homens – Peregr.	9,70665	9,1854
homens – NãoPeregr.	2,12425	2,10125
Univers – NãoUniver.	**11,6121**	**11,53955**
Univer. – Peregr	9,62975	8,99675
Univer. – NãoPeregr.	3,1857	2,8462
NãoUniver. – Peregr.	7,84815	6,634
NãoUniver. – NãoPeregr.	8,5075	8,2438
Peregr. – NãoPeregr.	**11,54975**	**11,3894**

Tabela 7: Cálculo de divergências usando MWE.

5 Conclusões

As duas experiências apresentadas (a comparação de dois textos de especialidade e de um texto literário —primeiro teste— e a comparação dos três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela) permitem confirmar que a divergência de Kullback-Leibler (divergência KL) é uma medida robusta porque extrai, em ambos os casos, os valores esperados.

A configuração que apresenta divergências maiores é a que só toma em conta frequências > 1 e usa lemas.

Portanto, das duas hipóteses de partida:

1. A divergencia de Kullback-Leibler (divergência KL) permite comparar automaticamente textos não anotados;
2. O uso de MWE será mais adequado do que o uso dos lemas porque apresentará valores de divergência maiores;

só se confirmou a primeira, embora se possa afirmar que o uso das MWE também é válido para comparar textos com a divergência KL pois se conseguem resultados igualmente robustos. É preciso também sublinhar que o número de MWE utilizados para calcular as divergências é menor que o de lemas, o que nos leva a inferir que uma extração automática com mais cobertura e exaustividade deveria melhorar os resultados.

No primeiro teste, utilizamos dois textos de especialidade (da área da pediatria) e um texto literário, presumindo que os textos de especialidade deveriam apresentar maiores divergências relativamente ao texto literário do que entre eles próprios. Como as experiências feitas mostraram a tendência esperada, decidimos aplicar a metodologia a um segundo grupo de textos (três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela).

No segundo teste, baseado em entrevistas, os resultados não só ajudam a confirmar a eficácia da medida de divergência, mas também permitem confirmar a pertinência das categorias socioculturais utilizadas para desenhar as entrevistas, bem como a sua pertinência nas análises qualitativas futuras que pretendemos desenvolver no projeto. Neste sentido, conjeturamos que uma categoria social ou cultural tem traços distintivos diferenciadores se o seu discurso é divergente do de indivíduos doutras categorias.

Deixámos para futuros trabalhos:

1. Procurar estratégias que permitam combinar o uso de formas, lemas e MWE no cálculo de divergências/convergências em textos não anotados.
2. As análises relacionais e contrastivas (do ponto de vista qualitativo e quantitativo) dos lemas e MWE mais relevantes extraídos das transcrições das entrevistas referidas na nota 4.

Agradecimentos

Este trabalho é apoiado pelo projeto *Narrativas, usos e consumos de visitantes como aliados ou amenaças para o bem-estar da comunidade local: o caso de Santiago de Compostela.* Ref: FFI2017-88196-R, parcialmente subsidiado pelo *Ministerio de Industria, Economía y Competitividad* espanhol no quadro do *Programa Estatal de I+D+i Orientada a los Retos de la Sociedad (2018-2021).*

Referências

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau & Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. Em *International Workshop on Semantic Evaluation (SemEval)*, 497–511.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk* 23(3). 321–346.

Fergadiotis, Gerasimos, Heather H. Wright & Thomas M. West. 2013. Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology* 22(2). 397–408. doi:10.1044/1058-0360.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28. doi:10.21814/lm.9.1.243.

Gamallo, Pablo, Juan Carlos Pichel, Marcos Garcia, José Manuel Abuín & Tomás Fernández-Pena. 2015. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. *Procesamiento del Lenguaje Natural* 53. 17–24.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies* CCIS, 65–75. Springer.

Iriarte, Álvaro. 2001. *A unidade lexicográfica. palavras, colocações, frasemas, pragmatemas*: Universidade do Minho. Tese de Doutoramento.

Kilgarriff, Adam. 1996. Why chi-square doesn't work, and an improved LOB-Brown comparison. Em *Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 169–172.

Kullback, S. & R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86. doi:10.1214/aoms/1177729694.

Loughran, Tim & Bill McDonald. 2013. Measuring readability in financial disclosures. *Journal of Finance* doi:10.2139/ssrn.1920411.

Maia, Belinda, Rui Sousa Silva, Anabela Barreiro & Cecília Fróis. 2008. N-grams in search of theories. Em Barbara Lewandowska-Tomaszczyk (ed.), *Corpus Linguistics, Computer Tools, and Applications: State-of-the Art*, 71–84. Peter Lang.

McCarthy, PM & J Jarvis. 2010. Mtld, voc-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 41(2). 381–392.

Mel'čuk, Igor, André Clas & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in Neural Information Processing Systems*, 3111–3119.

Pazos-Justo, Carlos, María Luísa del Río Araujo & Roberto Samartim. 2018. Políticas culturais e comunidade local: contributos para a análise do caso de santiago de compostela como meta dos caminhos de santiago. Em *Atas do III Congresso Internacional sobre Culturas: Interfaces da Lusofonia*, vol. 7, Instituto de Ciências Sociais da Universidade do Minho. No prelo.

Robins, Robert Henry. 1997. *A short history of linguistics* Longman linguistics library. Longman.

Samartim, Roberto. 2015. Bases de dados para o estudo da cultura: apresentação do catalogador e possibilidades de abordagem sobre o corpus documental do projeto caminho de santiago. Em *Estudos da AIL sobre teoria e metodologia*, vol. 2, 115–125. AIL Editora.

Saussure, Ferdinand. 1999. *Curso de linguística geral*. Lisboa: Dom Quixote.

Stubbs, Michael & Isabel Barth. 2003. Using recurrent phrases as text type discriminators: a quantitative method and some findings. *Functions of Language* 10(1). 61–104.

Torres Feijó, Elias. 2015. Identity sustainability, identity affectivity, and the ithaca traveler: Conceptual tools for measuring and modeling tourism as an opportunity. Em Gabriel R. Ricci (ed.), *Travel, Tourism and Identity, Culture & Civilization*, vol. 7, 143–162. Transaction Publishers.

Tweedie, Fiona J. & Harald R. Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.