Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 8, Número 1- Julho 2016

ISSN: 1647-0818

Volume 8, Número 1 – Julho 2016

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Con este número chegamos ao oitavo ano de Linguamática, un fito que logramos atinxir grazas á implicación e axuda de todo o equipo de colaboradores da revista. Autores, revisores e editores logramos cada semestre encher de contidos relevantes este espazo de divulgación científica concibido para a comunidade do PLN no ámbito lingüístico e cultural da Península Ibérica.

O labor non é doado, pois non é pouca a presión das autoridades académicas e políticas para favorecer as publicacións en lingua inglesa, en editoriais estranxeiras e en revistas catalogadas consonte estes principios subordinantes e ben pouco favorábeis a valorar axeitadamente a actividade científica orientada ao desenvolvemento das nosas propias comunidades lingüísticas, culturais e nacionais.

Precisamente por estas razóns, celebramos aínda con máis orgullo a saída do prelo de Linguamática e agradecemos con maior agarimo o traballo desinteresado de todas as persoas que semestre a semestre xuntan os seus esforzos entusiastas para conseguir unha publicación científica digna escrita nas nosas linguas e cada vez máis recoñecida, indexada e citada no ámbito das tecnoloxías das linguas ibéricas.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade de Santiago de Compostela

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Rui Pedro Marques,
Universidade de Lisboa

Salvador Climent Roca,
Universitat Oberta de Catalunya

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Artigos de Investigação

Proposta recebida em Abril 2016 e aceite para publicação em Junho 2016.

Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas?

Compiling Specialised Comparable Corpora. Should we always thrust (Semi-)automatic Compilation Tools?

Hernani Costa
Universidade de Málaga
hercos@uma.es

Isabel Dúran Muñoz
Universidade de Málaga
iduran@uma.es

Gloria Corpas Pastor
Universidade de Málaga
g.corpas@uma.es

Ruslan Mitkov
Universidade de Wolverhampton
r.mitkov@wlv.ac.uk

Resumo

Decisões tomadas anteriormente à compilação de um corpo comparável têm um grande impacto na forma em que este será posteriormente construído e analisado. Diversas variáveis e critérios externos são normalmente seguidos na construção de um corpo, mas pouco se tem investigado sobre a sua distribuição de similaridade textual interna ou nas suas vantagens qualitativas para a investigação. Numa tentativa de preencher esta lacuna, este artigo tem como objetivo apresentar uma metodologia simples, contudo eficiente, capaz de medir o grau de similaridade interno de um corpo. Para isso, a metodologia proposta usa diversas técnicas de processamento de linguagem natural e vários métodos estatísticos, numa tentativa bem sucedida de avaliar o grau de similaridade entre documentos. Os nossos resultados demonstram que a utilização de uma lista de entidades comuns e um conjunto de medidas de similaridade distribucional são suficientes, não só para descrever e avaliar o grau de similaridade entre os documentos num corpo comparável, mas também para os classificar de acordo com seu grau de semelhança e, consequentemente, melhorar a qualidade do corpos através da eliminação de documentos irrelevantes.

Palavras chave

corpos comparáveis, linguística computacional, medidas de similaridade distribucional, compilação manual e semi-automática.

Abstract

Decisions at the outset of compiling a comparable corpus are of crucial importance for how the corpus is to be built and analysed later on. Several variables and external criteria are usually followed when building a corpus but little is been said about textual distributional similarity in this context and the quality that it brings to research. In an attempt to fulfil this gap, this paper aims at presenting a simple but efficient methodology capable of measuring a corpus internal degree of relatedness. To do so, this methodology takes advantage of both available natural language processing technology and statistical methods in a successful attempt to access the relatedness degree between documents. Our findings prove that using a list of common entities and a set of distributional similarity measures is enough not only to describe and assess the degree of relatedness between the documents in a comparable corpus, but also to rank them according to their degree of relatedness within the corpus.

Keywords

comparable corpora, computational linguistics, distributional similarity measures, manual and semi-automatic compilation.

1 Introdução

O EAGLES — Expert Advisory Group on Language Engineering Standards Guidelines (EAGLES, 1996) define "corpos comparáveis" da seguinte forma: "Um corpo comparável é aquele que seleciona textos semelhantes em mais de um idioma ou variedade. Devido à escassez de exemplos de corpos comparáveis, ainda não existe um acordo sobre a sua similaridade."

Desde o momento em que esta definição foi criada em 1996, muitos corpos comparáveis foram compilados, analisados e utilizados em várias disciplinas.

A verdade é que este recurso acabou por se tornar essencial em várias áreas de investigação, tais como o Processamento de Linguagem Natural (PLN), terminologia, ensino de idiomas e tradução automática e assistida, entre outras. Neste momento podemos afirmar que não existe mais "escassez de exemplos de corpos comparáveis". Como Maia (2003) referiu: "os corpos comparáveis são vistos como uma resposta às necessidades de textos como exemplo de texto 'natural' original na cultura e idioma de origem" e, portanto, não é surpresa nenhuma que tenhamos assistido a um aumento no interesse por esses recursos e, um grande impulso na compilação de corpos comparáveis, especialmente no campo da investigação nas últimas décadas.

Contudo, de momento, "ainda não existe um acordo sobre a sua similaridade". A incerteza sobre os dados com que estamos a lidar ainda é um problema inerente para aqueles que lidam com corpos comparáveis. De facto, pouca investigação tem sido feita sobre a caraterização automática deste tipo de recurso linguístico, e tentar fazer uma descrição significativa do seu conteúdo é, muitas vezes, uma tarefa no mínimo arriscada (Corpas Pastor & Seghiri, 2009). Geralmente a um corpo é atribuído uma breve descrição do seu conteúdo, como por exemplo "transcrições de falas casuais" ou "corpo especializado comparável de turismo", juntamente com outras etiquetas que descrevem a sua autoria, data de criação, origem, número de documentos, número de palavras, etc. Na nossa opinião, estas especificações são de pouca valia para aqueles que procuram um corpo representativo de um domínio específico de elevada qualidade, ou até mesmo para aqueles que pretendem reutilizar um determinado corpo para outros fins. Desta forma, a maioria dos recursos à nossa disposição são construídos e partilhados sem que seja feita uma análise profunda ao seu conteúdo. Aqueles que os utilizam cegamente, confiam nas pessoas ou no grupo de investigação por detrás do seu processo de compilação, sem que conheçam a verdadeira qualidade interna do recurso, ou por outras palavras, sem conhecimento real sobre a quantidade de informação partilhada entre os seus documentos, ou quão semelhantes os documentos são entre si.

Assim, este trabalho tenta colmatar esta lacuna propondo uma nova metodologia que poderá ser utilizada em corpos comparáveis. Depois de selecionar o corpo que irá ser usado como cobaia em várias experiências, apresentamos a metodologia que explora várias técnicas de PLN juntamente com várias Medidas de Similaridade Distribucional (MSD). Para este efeito usámos uma lista de entidades comuns como parâmetro de entrada das MSD. Assumindo que os valores de saída das várias MSD podem ser usados como unidade de medida para identificar a quantidade de informação partilhada entre os documentos, a nossa hipótese é que estes valores possam ser posteriormente utilizados para descrever e caracterizar o corpo em questão.

O resto do artigo está estruturado da seguinte forma. A secção 2 descreve as vantagens e as desvantagens da compilação manual e automática de corpos e revela as atuais tendências de investigação usadas na compilação automática de corpos comparáveis. A secção 3 introduz alguns conceitos fundamentais relacionados com as MSD, ou seja, explica os fundamentos teóricos, trabalhos relacionados e as medidas utilizadas neste trabalho. A secção 4 apresenta o corpo utilizado nas nossas experiências, enquanto que a secção 5 descreve em detalhe a metodologia proposta, juntamente com todas as ferramentas, bibliotecas e *frameworks* utilizadas. E, finalmente, antes das conclusões finais (secção 7), a secção 6 descreve em detalhe os resultados obtidos.

2 Compilação Manual vs. Compilação Semi-automática

A compilação automática ou semi-automática de corpos comparáveis (ou seja, corpos compostos por textos originais semelhantes num ou mais idiomas usando os mesmos critérios de *design* (EAGLES, 1996; Corpas Pastor, 2001)) têm demonstrado muitas vantagens para a investigação atual, reduzindo particularmente o tempo necessário para construir um corpo e aumentando a quantidade de textos recuperados. Com ferramentas automáticas de compilação como o BootCaT (Baroni & Bernardini, 2004), WebBootCaT (Baroni et al., 2006) ou o Babouk (de Groc, 2011), hoje em dia é possível construir um corpo de grande tamanho num reduzido período de tempo, em contraste com o demorado protocolo de compilação e o número limitado de textos recuperados no mesmo intervalo de tempo quando a compilação é realizada manualmente. De facto, publicações recentes demonstram que a compilação automática está a superar a compilação manual, sendo cada vez maior o número de investigadores que tiram partido de ferramentas de compilação automática na construção dos seus corpos (Barbaresi, 2014; Jakubíček et al., 2014; Barbaresi, 2015; El-Khalili et al., 2015). A verdade é que neste momento é

um truísmo dizer que a compilação automática de corpos está a ganhar terreno sobre a compilação manual.

Apesar de ser possível compilar mais rapidamente maiores corpos comparáveis num curto espaço de tempo – o que é sem dúvida a maior vantagem da compilação automática – é contudo necessário analisar todo o espectro de propriedades implícitas no processo. Em primeiro lugar, um dos inconvenientes mais importantes a considerar quando se lida com a compilação automática é o ruído, ou seja, a quantidade de informação irrelevante que acaba por ser adicionada ao corpo durante o processo. Ruído este que se tenta colmatar através de uma supervisão rigorosa nas primeiras fases, de modo a evitar possíveis repercussões nas fases seguintes. Deste modo, é quase desnecessário afirmar que a compilação automática também requer intervenção humana a fim de obter bons resultados durante o processo de compilação — daí a origem da palavra "semi-automática". Contudo, esta intervenção torna-se uma tarefa bastante tediosa e cansativa, dada a necessidade de filtrar determinados domínios na rede, eliminar pares de entidades ou páginas na rede irrelevantes oferecidas pela ferramenta de compilação (Gutiérrez Florido et al., 2013).

Outra caraterística interessante de analisar é o grau de semelhança entre documentos compilados manualmente e semi-automaticamente. Apesar de à primeira vista pensarmos que a compilação manual é a única que garante a qualidade em termos de forma e conteúdo num corpo, devido ao facto deste tipo de compilação ser mais minuciosa em termos de seleção dos textos a serem adicionados ao corpo, até ao momento ainda não existe um método formal que prove a sua veracidade. Sendo a forma e conteúdo de suma importância na construção de corpos comparáveis, e posteriormente na análise do mesmo, este trabalho tem como principal objetivo propor um método capaz de descrever, medir e classificar em termos de forma e conteúdo o grau de similaridade em corpos comparáveis. Noutras palavras, capaz de avaliar o grau de semelhança/ similaridade dentro de um corpo compilado manualmente ou semi-automaticamente. E assim permitir que o investigador responsável pela compilação tenha um conhecimento mais aprofundado sobre os documentos com que está a lidar para que possa posteriormente decidir quais devem ou não fazer parte do corpo.

Numa tentativa de estandardizar o nosso trabalho, e considerando as limitações de cada tipo de compilação, tivemos em conta vários fatores comuns que devem ser satisfeitos por ambos tipos de compilação. Estas variáveis devem ser estabelecidas de modo a garantir a fiabilidade do corpo, a sua coerência interna e a representatividade do domínio. Deste modo, Bowker & Pearson (2002) propõe vários critérios a serem seguidos, os quais estão relacionados com as línguas de trabalho e o nível de especialização. Em seguida enumeramos os vários critérios externos a serem considerados:

- Critério temporal: a data de publicação ou criação dos textos selecionados;
- Critério geográfico: origem geográfica dos textos;
- Critério formal: autenticidade dos textos completos ou fragmentados;
- Tipologia dos textos: o género textual a que os textos pertencem;
- Critério de autoria: a fonte dos textos (autor, instituição, etc.).

É importante referir que, de modo a garantir a homogeneidade do corpo usado neste trabalho, estes critérios foram seguidos durante o processo de compilação, como explicado na secção 4. Além disso, é também importante referir que neste trabalho ambas as abordagens (manual ou semi-automática) usam as mesmas ferramentas para recuperar documentos (ou seja, o mesmo motor de busca).

3 Medidas de Similaridade Distribucional (MSD)

Embora a tarefa de estruturar informação a partir de linguagem natural não estruturada não seja uma tarefa fácil, o Processamento de Linguagem Natural (PLN) em geral e, Recuperação de Informação (RI) (Singhal, 2001) e Extração de Informação (EI) (Grishman, 1997) em particular, têm melhorado o modo como a informação é acedida, extraída e representada. Em particular, RI e EI desempenham um papel crucial na tarefa de localizar e extrair informação específica de uma coleção de documentos ou outro tipo de recursos em linguagem natural, de acordo com um determinado critério de busca. Para isso, estas duas áreas do conhecimento tiram partido de vários métodos estatísticos para extrair informação sobre as palavras e suas coocorrências. Essencialmente, esses métodos visam encontrar as palavras mais frequentes num documento e usar essa informação como atributo quantitativo num determinado método estatístico. Partindo do teorema distribucional de Harris (1970), o qual assume que palavras semelhantes tendem a ocorrer em contextos semelhantes, esses métodos

estatísticos são adequados, por exemplo, para encontrar frases semelhantes com base nas palavras contidas nas mesmas (Costa et al., 2015a), ou, por exemplo, para extrair e validar automaticamente entidades semânticas extraídas de corpos (Costa et al., 2010; Costa, 2010; Costa et al., 2011). Para este efeito, assume-se que a quantidade de informação contida, por exemplo, num determinado documento poderá ser acedida através da soma da quantidade de informação contida nas palavras do mesmo. Além disso, a quantidade de informação transmitida por uma palavra pode ser representada pelo peso que lhe é atribuído (Salton & Buckley, 1988). Deste modo, o Spearman's Rank Correlation Coefficient (SCC) e o Chi-Square (χ^2), duas medidas frequentemente aplicadas na área de RI, podem ser utilizadas para calcular a similaridade entre dois documentos escritos no mesmo idioma (ver secção 3.1 e 3.2 para mais detalhes sobre estas medidas). Ambas as medidas são particularmente úteis para este trabalho, visto que ambas são: independentes do tamanho do texto (ambas usam uma lista das entidades comuns); e, independentes do idioma.

Devido a ser independente do tamanho dos textos e à sua simplicidade de implementação, a medida distribucional do SCC tem demonstrado a sua eficácia no cálculo da similaridade entre frases, documentos e até mesmo em corpos de tamanhos variados (Costa et al., 2015a; Costa, 2015; Kilgarriff, 2001).

A medida de similaridade do χ^2 também tem demonstrado a sua robustez e alto desempenho. A título de exemplo, o χ^2 tem vindo a ser utilizado para analisar o componente de conversação no Corpo Nacional Britânico (Rayson et al., 1997), para comparar corpos (Kilgarriff, 2001), e até mesmo para identificar grupos de tópicos relacionados em documentos transcritos (Ibrahimov et al., 2002). Embora seja uma medida estatística simples, o χ^2 permite avaliar se a relação entre duas variáveis numa amostra é devida ao acaso, ou, pelo contrário, a relação é sistemática.

Devido às razões mencionadas anteriormente, as Medidas de Similaridade Distribucional (MSD), em geral, e o SCC e χ^2 em particular, têm uma vasta gama de aplicabilidades (Kilgarriff, 2001; Costa, 2015; Costa et al., 2015b). Deste modo, este trabalho tem como objetivo provar que estas medidas simples, contudo robustas e de alto desempenho, permitem descrever o grau de similaridade entre documentos em corpos especializados. Em seguida descrevemos em detalhe como funcionam estas duas MSD.

3.1 Spearman's Rank Correlation Coefficient (SCC)

Neste trabalho o Spearman's Rank Correlation Coefficient (SCC) é utilizado e calculado do mesmo modo que no artigo do Kilgarriff (2001). Inicialmente é criada uma lista de entidades comuns[1] L entre dois documentos d_l e d_m, onde

$$L_{d_l,d_m} \subseteq (d_l \cap d_m).$$

É possível usar n entidades comuns ou todas as entidades comuns entre dois documentos, onde n corresponde ao total número de entidades comuns em $\mid L \mid$, ou seja,

$$\{n \mid n \in \mathbb{N}^0, n \leq \mid L \mid\}.$$

Neste trabalho são utilizadas todas as entidades comuns encontradas entre dois documentos, ou seja, $n = \mid L \mid$. Em seguida, por cada documento, as listas de entidades comuns (por exemplo, L_{d_l} and L_{d_m}) são ordenadas por ordem crescente de frequência ($R_{L_{d_l}}$ e $R_{L_{d_m}}$), ou seja, a entidade menos frequente recebe a posição *1* no ranking e a entidade mais frequente recebe a posição n. Em caso de empate, onde mais do que uma entidade aparece no documento o mesmo número de vezes, é atribuída a média das posições.

Por exemplo, se as entidades e_a, e_b e e_c ocorrerem o mesmo número de vezes e as suas posições forem *6*, *7* e *8*, a todas elas é atribuída a mesma posição no ranking, ou seja, a sua nova posição no ranking seria $\frac{6+7+8}{3} = 7$.

Finalmente, para cada entidade comum $\{e_1, ..., e_n\} \in L$ em cada um dos documentos é calculada a diferença entre as suas posições e posteriormente normalizada através da soma dos quadros das suas diferenças

$$\left(\sum_{i=1}^{n} s_i^2\right).$$

A equação completa do SCC é apresentada na Equação 1, onde

$$\{SCC \mid SCC \in \mathbb{R}, -1 \leq SCC \leq 1\}.$$

Como exemplo, imagine-se que e_x é uma entidade comum (ou seja, $\{e_x\} \in L$) e,

$$R_{L_{d_l}} = \{1\#e_{n_{d_l}}, 2\#e_{n-1_{d_l}}, \ldots, n\#e_{1_{d_l}}\}, \quad e$$

$$R_{L_{d_m}} = \{1\#e_{n_{d_m}}, 2\#e_{n-1_{d_m}}, \ldots, n\#e_{1_{d_m}}\}$$

[1]Neste trabalho, o termo "entidade" refere-se a "palavras simples", as quais podem ser um *token*, um lema ou um stem.

são as listas ordenadas de entidades comuns de d_l e d_m, respetivamente. Assumindo que e_x é o $3\#e_{n-2_{d_l}}$ e $1\#e_{n_{d_m}}$, ou seja, e_x está na posição 3 do ranking em $R_{L_{d_l}}$ e na posição 1 em $R_{L_{d_m}}$, s seria calculado da seguinte forma: $s^2_{e_x} = (3-1)^2$ e, o resultado seria 4. Em seguida este processo seria repetido para as restantes $n-1$ entidades e o resultado do SCC corresponderia ao valor de similaridade entre d_l e d_m.

$$SCC(d_i, d_j) = 1 - \frac{6 \times \sum_{i=1}^{n} s_i^2}{n^3 - n} \quad (1)$$

3.2 Chi-Square (χ^2)

A medida do Chi-square (χ^2) também usa uma lista de entidades comuns (L). E à semelhança do SCC, também é possível usar n entidades comuns ou todas as entidades comuns entre dois documentos. Também neste caso optámos por usar a lista completa, ou seja, todas as entidades comuns encontradas entre dois documentos ($n = \mid L \mid$). O número de ocorrências de uma determinada entidade em L, que seria expectável em cada um dos documentos, é calculado usando a lista de frequências. Se o tamanho do documento d_l e d_m forem N_l e N_m e a entidade e_i tiver as seguintes frequências observadas $O(e_i, d_l)$ e $O(e_i, d_m)$, então os valores esperados seriam

$$e_{i_{d_l}} = \frac{N_l * (O(e_i, d_l) + O(e_i, d_m))}{N_l + N_m}, \quad e$$

$$e_{i_{d_m}} = \frac{N_m * (O(e_i, d_l) + O(e_i, d_m))}{N_l + N_m}.$$

Na Equação 2 é apresentada a fórmula completa do χ^2, onde O corresponde ao valor da frequência observada e E a frequência esperada. Assim, o valor resultante do χ^2 deverá ser interpretado como a distância interna entre dois documentos. Também é importante referir que

$$\{\chi^2 \mid \chi^2 \in \mathbb{R}, 1 \leq \chi^2 < +\infty\},$$

o que significa que quanto menos relacionadas as entidades forem em L, menor será o valor do χ^2.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

A Tabela 1 apresenta um exemplo de uma tabela de contingências. Assumindo que existem duas entidades comuns e_i e e_j entre dois documentos d_l e d_m (ou seja, $L = \{e_i, e_j\}$), esta tabela apresenta: i) as frequências observadas (O); ii) os totais nas margens; iii) as frequências esperadas (E), que foram obtidas através da seguinte fórmula:

$$\frac{column\ total}{N} \times row\_total,$$

por exemplo, $E(e_i, d_l) = \frac{14}{26} \times 15 = 8.08$. Assim que calculadas as frequências esperadas, o próximo passo seria calcular o χ^2 (veja-se a Equação 3).

$$\frac{(11-8.08)^2}{8.08} + \frac{(3-5.92)^2}{5.92} + \frac{(4-6.92)^2}{6.92} + \frac{(8-5.08)^2}{5.08} = 5.41 \quad (3)$$

	d_l	d_m	*Total*
e_i	O=11 E=8.08	O=4 E=6.92	15
e_j	O=3 E=5.92	O=8 E=5.08	11
Total	14	12	26

Tabela 1: Exemplo de uma tabela de contingência.

4 O Corpo INTELITERM

O corpo INTELITERM[2] é um corpo comparável especializado composto por documentos recuperados da Internet. Inicialmente foi compilado manualmente, por investigadores, com o objetivo de construir um corpo em inglês, espanhol, alemão e italiano livre de ruído e representativo na área do Turismo e Beleza. No entanto, numa fase posterior, a fim de aumentar o tamanho do mesmo, mais documentos foram recuperados automaticamente usando a ferramenta de compilação BootCaT[3] (Baroni & Bernardini, 2004). De modo a manter a homogeneidade e a qualidade do corpo, em ambos os processos de compilação foram seguidas as mesmas variáveis e critérios externos (ver Tabela 2).

Em detalhe, o corpo comparável INTELITERM pode ser dividido em quatro subcorpos de acordo com o idioma, ou seja, inglês, espanhol, alemão e italiano. Estes subcorpos, por sua vez podem ser subdivididos por tipo de documento, isto é, textos originais compilados manualmente, textos traduzidos compilados manualmente e textos originais compilados

[2] http://www.lexytrad.es/

[3] http://bootcat.sslmit.unibo.it

Critério	**Descrição**
Temporal	A data de publicação ou criação dos textos selecionados deve ser tão recente quanto possível.
Geográfico	De modo a evitar uma possível variação terminológica diatópica, como o espanhol falado no México ou Venezuela, todos os textos selecionados são geograficamente limitados, ou seja, todos os textos utilizados, por exemplo, em espanhol são provenientes de Espanha, e todos os textos italianos são da Itália.
Formal	Os textos selecionados referem-se a um contexto de comunicação especializado, ou seja, a um contexto de nível médio-alto de especialização, são originalmente escritos nas línguas do estudo e estão no seu formato eletrónico original.
Género ou tipologia textual	Todos os textos selecionados pertencem ao mesmo género, ou seja, são textos promocionais recuperados da Internet contendo informação sobre produtos e serviços de bem-estar e beleza na área do turismo.
Autor	Todos os textos são documentos autênticos criados por autores relevantes, instituições ou empresas.

Tabela 2: Variáveis e critérios externos utilizados durante o processo de compilação.

automaticamente. Dado o reduzido tamanho do corpo (veja-se Tabela 3), decidimos usar todos os seus documentos, ou seja, todos os documentos originais e traduzidos compilados manualmente para o inglês (*i_en_od* e *i_en_td*), espanhol (*i_es_od* e *i_es_td*), alemão (*i_de_od* e *i_de_td*) e italiano (*i_it_od* — os investigadores não encontraram textos traduzidos para o italiano), assim como todos os documentos compilados automaticamente usando a ferramenta de compilação automática bootcaT para o inglês, espanhol, alemão e italiano (*bc_en*, *bc_es*, *bc_de* and *bc_it*, respetivamente). Toda a informação relativa aos subcorpos referidos anteriormente é apresentada na Tabela 3. Esta tabela apresenta o número de documentos (nD), o número de palavras únicas (*types*), o número total de palavras (*tokens*), a relação entre palavras únicas e o número total de palavras (*types*/*tokens*) por subcorpos e o tipo de fonte (sT), a qual pode ser original, tradução ou *crawled*/recuperado automaticamente (ori., trans. e craw., respetivamente). Os valores apresentados na Tabela 3 foram obtidos através da ferramenta de análise de concordância Antconc 3.4.3 (Anthony, 2014).

	nD	types	tokens	$\frac{types}{tokens}$	sT
i_en_od	151	11.6k	496.2k	0,023	ori.
i_en_td	60	6.9k	83.1k	0,083	trans.
i_es_od	224	13.0k	207.3k	0,063	ori.
i_es_td	27	3.4k	16.4k	0,207	trans.
i_de_od	138	21.4k	199.8k	0,049	ori.
i_de_td	109	5.5k	26.8k	0,205	trans.
i_it_od	150	19.9k	386.2k	0,051	ori.
bc_en	111	41.1k	563.5k	0,073	craw.
bc_es	246	32.8k	735.4k	0,045	craw.
bc_de	253	58.3k	482.4k	0,121	craw.
bc_it	122	11.9k	81.5k	0,147	craw.

Tabela 3: Informação estatística dos vários subcorpos do INTELITERM.

5 Medindo o Grau de Similaridade entre Documentos

Esta secção tem como objetivo apresentar uma metodologia simples, contudo eficiente capaz de descrever e extrair informação sobre o grau interno de similaridade de um determinado corpo. De facto, em última instância, esta metodologia permitir-nos-á não só descrever os documentos num corpo, mas também medir e classificar documentos com base nos seus valores de similaridade. Em seguida descrevemos a metodologia usada para este fim, juntamente com todas as ferramentas, bibliotecas e *frameworks* utilizadas no processo.

i) **Pré-processamento dos dados**: em primeira instância processámos o corpo com o OpenNLP[4] de modo a delimitar as frases e as palavras. Relativamente ao processo de anotação, utilizámos o TT4J[5], uma biblioteca em Java que permite invocar a ferramenta TreeTagger (Schmid, 1995) – uma ferramenta criada especificamente para identificar a categoria gramatical e o lema das palavras. Em relação ao *stemming*, usámos o algoritmo Porter stemmer fornecido pela biblioteca Snowball[6]. Também foi implementado manualmente um módulo para remover sinais de pontuação e caracteres especiais dentro das palavras. Além disso, de modo a eliminarmos o ruído, foi criada uma lista de stopwords[7] para identificar as palavras mais frequentes no corpo, ou seja, palavras vazias sem informação semântica. Uma vez processado um determinado documento, ou seja, depois de delimitar as frases, identificar

[4] https://opennlp.apache.org
[5] http://reckart.github.io/tt4j/
[6] http://snowball.tartarus.org
[7] Disponíveis através do seguinte endereço na rede: https://github.com/hpcosta/stopwords.

as palavras, a sua categoria gramatical, o seu lema e o seu stem, o sistema cria um novo ficheiro onde é guardada toda esta nova informação. Além disso, também é adicionado ao ficheiro um vetor booleano que descreve se uma entidade é uma palavra irrelevante (ou seja, stopword) ou não. Desta forma, o sistema irá ser capaz de utilizar somente as palavras, lemas e stems que não sejam stopwords.

ii) Identificação da lista de entidades comuns entre documentos: de modo a identificar a lista de entidades comuns (para futura referência, EC), foi criada uma matriz de coocorrências por cada par de documentos. Neste trabalho, somente pares de documentos com pelo menos uma entidade em comum são processados. Como exigido pelas MSD (ver secção 3), a frequência das EC em ambos os documentos são guardadas numa matriz de coocorrências

$$\begin{aligned} L_{d_l,d_m} \quad = \quad & \{e_i, (f(e_i,d_l), f(e_i,d_m)); \\ & e_j, (f(e_j,d_l), f(e_j,d_m)); \ldots \\ & e_n, (f(e_n,d_l), f(e_n,d_m))\} \end{aligned}$$

onde f representa a frequência de uma entidade num determinado documento d). Com o objetivo de analisar e comparar o desempenho das várias MSD foram criadas três listas para serem utilizadas como parâmetros de entrada: a primeira usando o número de tokens em comum (NTC), a segunda usando o número de lemas em comum (NLC) e a terceira usando o número de stems em comum (NSC).

iii) Calcular a similaridade entre documentos: a similaridade entre documentos foi calculada aplicando as várias MSD ($MSD = \{MSD_{EC}, MSD_{SCC}, MSD_{\chi^2}\}$, onde os índices EC, SCC e χ^2 correspondem ao número de entidades comuns ao Spearman's Rank Correlation Coefficient e ao Chi-Square, respetivamente), usando os três parâmetros de entrada (NTC, NLC e NSC).

iv) Calcular a pontuação final do documento: a pontuação final do documento $MSD(d_l)$ resulta da média das similaridades entre o documento d_l com todos os demais documentos na coleção de documentos, ou seja,

$$MSD(d_l) = \frac{\sum_{i=1}^{n-1} MSD_i(d_l, d_i)}{n-1},$$

onde n representa o número total de documentos na coleção e $MSD_i(d_l, d_i)$ o valor de similaridade entre o documento d_l com o documento d_i.

v) Classificar os documentos: por fim, os documentos são classificados por ordem descendente de acordo com o valor resultante final das várias MSD (ou seja, MSD_{EC}, MSD_{SCC} ou MSD_{χ^2}).

6 Avaliando o Corpo usando MSD

Depois de apresentado o problema que pretendemos explorar, a metodologia que iremos aplicar e os dados com os quais iremos trabalhar, é hora de juntar todas as peças num cenário de teste e explicar as nossas descobertas. Para este efeito, as Medidas de Similaridade Distribucional (MSD), apresentados na secção 3, serão aplicadas para explorar e classificar os documentos do corpo INTELITERM. Esta experiência divide-se em duas partes distintas. Na primeira parte, usaremos os vários subcorpos compilados manualmente para explorar e comparar o conteúdo dos documentos originais com os traduzidos, de modo a compreender como eles diferem entre si de um ponto de vista estatístico (secção 6.1). Depois, na segunda parte, faremos uma análise comparativa entre os documentos compilados manualmente com os semi-automaticamente compilados (secção 6.2). Por fim, esta secção termina com uma discussão geral sobre os resultados obtidos (secção 6.3).

A fim de descrever os dados em mãos é aplicada a metodologia apresentada na secção 5, juntamente com as três diferentes MSD, ou seja: o número de entidades comuns (EC); o Spearman's Rank Correlation Coefficient (SCC); e o Chi-Square (χ^2). Como parâmetro de entrada para as diferentes MSD, usaremos três diferentes listas de entidades (isto é, tokens, lemas e stems). As Figuras 1, 2 e 3 apresentam o número médio (av) do número de tokens comuns (NTC) entre documentos, os valores resultantes do SCC e do χ^2, juntamente com os seus desvios padrão correspondentes (σ — linhas verticais que se estendem a partir das barras) por medida e subcorpos (ou seja, documentos *o*riginais, *t*raduzidos e compilados automaticamente com o *b*oot*c*aT. Usaremos os seus acrónimos, a partir deste momento: *i_od*, *i_td* and *bc*, respetivamente).

É importante referir que neste trabalho usamos todos os documentos do corpo INTELITERM e, portanto, todos os resultados observados resultam de toda a população, e não de uma amostra. Ou seja, são utilizados todos os documentos em: inglês (*i_en_od*, *i_en_td* e

bc_en); espanhol (*i_es_od*, *i_es_td* e *bc_es*); alemão (*i_de_od*, *i_de_td* e *bc_de*); e italiano (*i_it_od* e *bc_it*) — importante referir novamente que para o italiano não existe um o subcorpo de documentos traduzidos (ver secção 4).

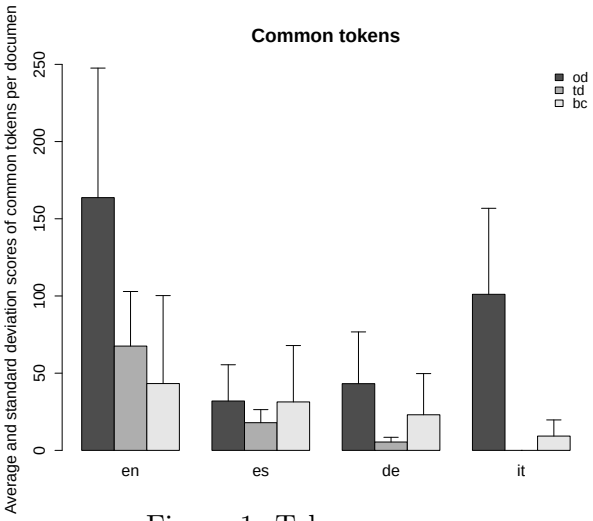

Figura 1: Tokens comuns.

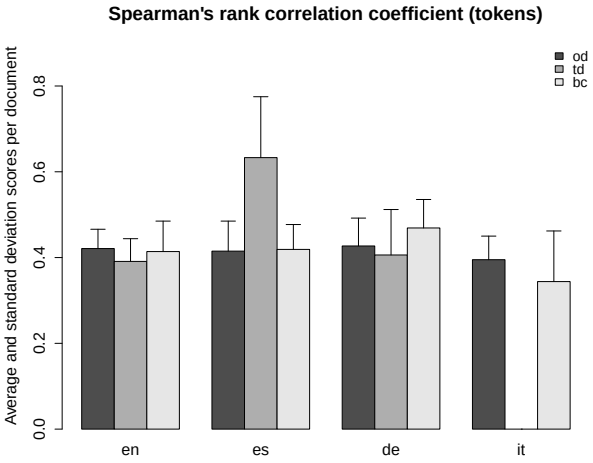

Figura 2: SCC.

6.1 Documentos Originais vs. Traduzidos

As Figuras 4 a 12 apresentam os valores médios por documento num formato de *box plot* para todas as combinações MSD *vs.* subcorpo. Em cada uma das *box plot* é apresentada a gama de variação (mínimo e máximo), o intervalo de variação (variação interquartil), a mediana e os valores mínimos e máximos extremos (também conhecidos como *outliers*).

A primeira observação que podemos fazer a partir das Figuras 4, 7 e 10 é que as distribuições entre os distintos parâmetros de entrada são bastante semelhantes. Embora não seja possível generalizar estes resultados para outros tipos de corpos ou domínios, todas as MSD sugerem a

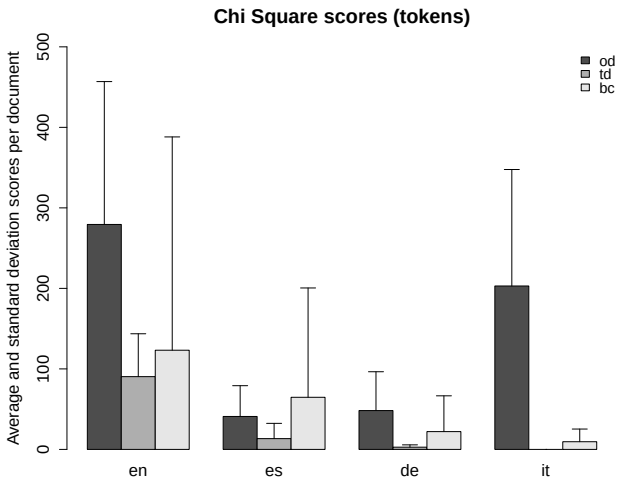

Figura 3: χ^2.

mesma conclusão: é possível alcançar resultados aceitáveis apenas usando tokens, ou seja, palavras na sua forma original. Como os stems e os lemas exigem mais poder computacional e tempo para serem processados — especialmente os lemas, devido à sua dependência à categoria gramatical e ao tempo de processamento subjacente – a possibilidade de usar apenas tokens é uma mais valia não só para as MSD, mas principalmente para o método proposto neste trabalho.

Deste modo vamo-nos focar nas Figuras 4, 5 e 6. Com base nos resultados apresentados nas mesmas, podemos afirmar que os valores obtidos por cada subcorpo é simétrico (distribuição simétrica com a mediana no centro do retângulo), o que significa que os dados seguem uma distribuição normal. Contudo, há algumas exceções, como por exemplo nos valores médios para o SCC e para o χ^2, mais precisamente para o subcorpo *i_es_td* e para o *i_de_td*, os quais serão mais tarde analisados em detalhe nesta secção. Outra observação interessante está relacionada com o elevado número de entidades comuns (EC) — veja-se Figuras 1, 4, 7 e 10 — nos documentos originais (*i_en_od*, *i_es_od* e *i_de_od*) quando comparado com os documentos traduzidos (*i_en_td*, *i_es_td* e *i_de_td*, respetivamente). Por exemplo, o subcorpo *i_en_od* (o subcorpo em inglês que contém documentos originais) contém 163,70 tokens em comum por documento em média (av) com um desvio padrão (σ) de 83,89, enquanto que o subcorpo *i_en_td* (o qual contém textos traduzidos em inglês) tem somente 67,54 tokens comuns por documento em média com um σ=35,35 (ver Figura 1).

A mesma observação pode ser feita para os subcorpos originais em espanhol e alemão (*i_es_od*={av=31,97; σ=23,48} e

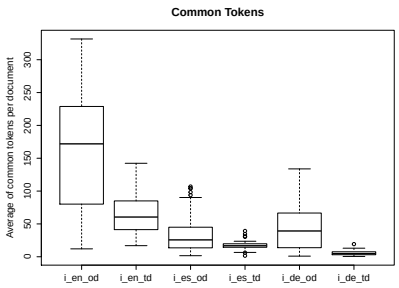

Figura 4: NTC.

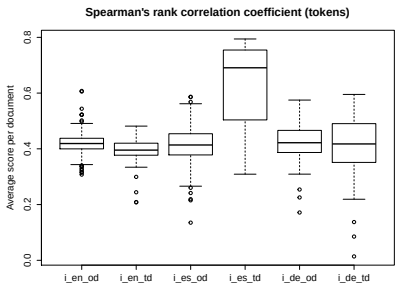

Figura 5: SCC.

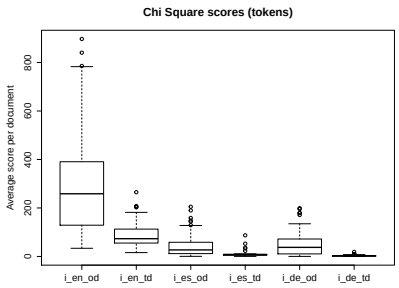

Figura 6: χ^2.

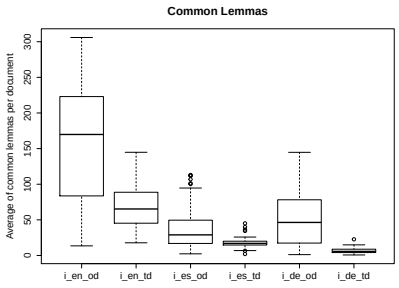

Figura 7: Lemas.

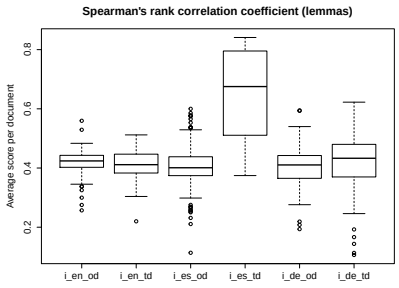

Figura 8: SCC.

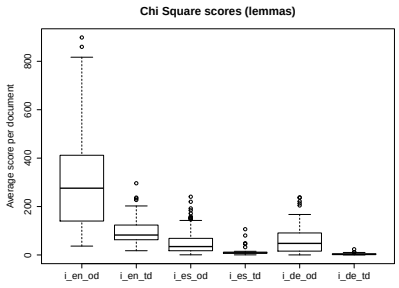

Figura 9: χ^2.

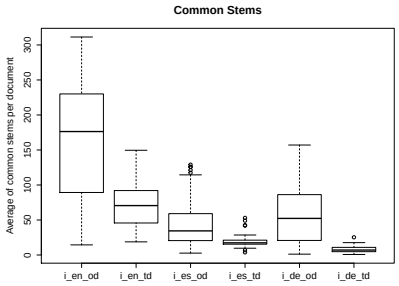

Figura 10: Stems.

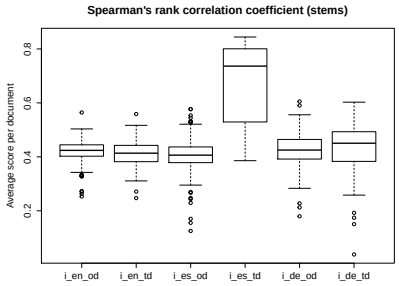

Figura 11: SCC.

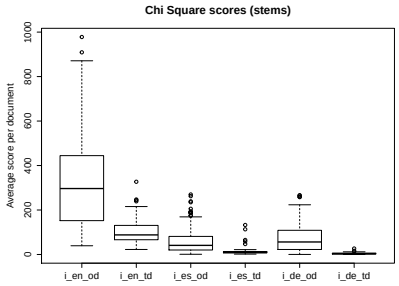

Figura 12: χ^2.

i_de_od={av=43,21; σ=33,52}) com os seus subcorpos traduzidos (*i_es_td*={av=17,93; σ=8,46} e *i_de_td*={av=5,42; σ=3,05}), ver Figuras 1 e 4 — repare-se que a Figura 4 mostra como os dados estão distribuídos acima e abaixo da mediana e a Figura 1 apresenta as distintas médias e seus desvios padrão correspondentes.

Uma possível explicação para estes valores baseia-se no fato destes documentos, recuperados da Internet, serem documentos traduzidos (ou seja, traduzidos de diferentes línguas e por diferentes tradutores) e, consequentemente, devido à variabilidade das várias caraterísticas linguísticas, tais como vocabulário, estilo, repetição, fontes, etc., em cada um dos documentos, pode muito bem explicar o porquê de haver um menor número de EC entre os documentos traduzidos quando comparado com os documentos originais.

Embora a média do número de tokens comuns por documento (NTC) seja maior para o corpo *i_en_od*, a amplitude inter-quartis (IQR) é maior que nos demais subcorpos (ver Figuras 1 e 4), o que significa que em média, 50% dos dados estão mais distribuídos e, consequentemente, a média de NTC por documento é mais variável. Além disso, na Figura 4 podemos verificar que os *whiskers* são longos (ou seja, as linhas que se estendem verticalmente a partir do retângulo), o que poderá indicar uma certa variabilidade fora dos quartis superiores e inferiores (ou seja entre o máximo e o Q3 e entre o Q1 e o mínimo). Portanto, podemos dizer que o subcorpo *i_en_od* contêm uma grande variedade de tipos de documentos e, consequentemente, alguns deles estão minimamente correlacionados com os demais documentos do subcorpo. No entanto, os dados são positivamente assimétricos, o que significa que a maioria está fortemente correlacionada, isto é, os documentos partilham um elevado NTC entre si. Esta ideia pode ser sustentada pelos valores médios do SCC e o elevado número de *outliers* positivos que se observam na Figura 5. Além disso, a média de 0,42 para o SCC e σ=0,045 também corroboram a existência de uma forte correlação entre os documentos no subcorpo *i_en_od*. Em relação aos valores do χ^2, o longo *whisker* que sai do Q1, na Figura 6, também deve

ser interpretado como indício de um elevado grau de similaridade entre os documentos.

Em relação ao subcorpo *i_en_td*, os valores do NTC, do SCC e do χ^2 (Figuras 4, 5 e 6) e, a média de 67,54 tokens comuns por documento e o σ=35,35 (Figura 1) sugerem que os dados estão normalmente distribuídos (Figura 5) e os documentos — não tanto como no subcorpo *i_en_od*, contudo — também estão fortemente relacionados entre si.

De todos os subcorpos, o *i_es_od* é o maior, contendo 224 documentos (Tabela 3). No entanto, as Figuras 1 e 4 revelam que o NTC é mais baixo em comparação com os dois subcorpos em inglês. Embora uma análise linguística mais aprofundada nos daria uma explicação mais precisa, uma possível teoria passa pelo facto de que o espanhol tem uma morfologia mais rica em relação ao inglês. E, portanto, devido a um maior número de formas flexionadas por lema, existe um maior número de tokens e, consequentemente, menos tokens em comum entre os documentos em espanhol. Ao analisarmos as Figuras 4 e 6, ambas as *box plots* do subcorpo *i_es_od* resultam bastante similar às do *i_en_td* caso haja um valor médio de tokens maior por documento. Com a exceção do *whisker* mais longo na Figura 5, os valores do SCC também apresentam distribuições, médias e desvios padrão bastante similares quando comparados com o subcorpo *i_en_td* (veja-se Figura 1).

Apesar do subcorpo alemão *i_de_od* ter mais *tokens* e menos *types* (21,4k e 199,8k, respetivamente) quando comparado com o *i_es_od* (13k types e 207,3k tokens), o seu rácio $\frac{types}{tokens}$ não varia muito entre eles (0,049 contra 0.063, para mais detalhes veja-se Tabela 3). O mesmo ocorre com os valores do NTC, do SCC e do χ^2 (Figuras 1, 2 e 3). Por exemplo, o NTC entre os documentos, em média, para o subcorpo *i_es_od* é de 31,97 com um σ=23,48, contra uma av=43,21 e um σ=33,52 para o subcorpo *i_de_od*. Além disso, a média e o desvio padrão do seu SCC e χ^2 são ainda mais expressivos:

- SCC={av=0,415 e σ=0,07} para o *i_es_od*;
- SCC={av=0,427} e σ=0,065 para o *i_de_od*

e também

- χ^2={av=40,922; σ=38,212} para o *i_es_od*;
- χ^2={av=48,235; σ=45,301} para o *i_de_od*.

Como podemos observar nas Figuras 4, 5 e 6, a média de valores por documento para ambos os subcorpos *i_es_td* e *i_de_td* são ligeiramente diferentes dos valores apresentados nas *box plots* do subcorpo *i_en_td*. Além do reduzido NTC por documento, os desvios padrão do χ^2 resultarem maiores que as suas médias (*i_es_td*={av=13,40; σ=18,95} e *i_de_td*={av=2,771; σ=2,883}), e a expressiva variabilidade dentro e fora do IQR do SCC no subcorpo *i_es_td* indiciam uma certa inconsistência nos dados. Esta instabilidade poderá ser explicada pelo reduzido número de types (*i_es_td*=3,4k e *i_ed_td*=5,5k) e tokens (*i_es_td*=16,4k e *i_de_td*=26,8k) e pelo seu rácio $\frac{types}{tokens}$ de 0,207 e 0,205, respetivamente (Tabela 3).

Como referido por Baker (2006), a análise do rácio $\frac{types}{tokens}$ torna-se útil quando estamos perante subcorpos de tamanho reduzido. Assim, é bastante interessante observar que estes dois subcorpos só têm em média 607 e 246 tokens

$$i\_es\_td = \frac{16400}{27} \approx 607, e$$

$$i\_de\_td = \frac{26800}{109} \approx 246,$$

e, 126 e 50 types por documento

$$i\_es\_td = \frac{3400}{27} \approx 126, e$$

$$i\_de\_td = \frac{5500}{109} \approx 50,$$

o que os converte numa excelente prova de conceito. Quando comparados com os baixos rácios dos demais subcorpos (ver Tabela 3), — mesmo para este tipo de corpos — estes valores podem muito bem serem considerados elevados. Deste modo, podemos concluir que o elevado rácio sugere que estamos perante uma forma mais diversificada do uso da linguagem, o que consequentemente também pode explicar os baixos valores no NTC e do χ^2 para estes dois subcorpos. Por outro lado, um rácio baixo também pode indicar um grande número de repetições (uma mesma palavra ocorrendo uma e outra vez), o que pode implicar que estamos perante um domínio bastante especializado. Apesar do elevado valor do SCC, os dados são assimétricos e variáveis (veja-se a grande amplitude interquartis na Figura 5). Isso acontece porque a maioria das entidades comuns ocorrem poucas vezes nos documentos e, consequentemente, estas posicionam-se próximas umas das outras nas listas de ranking, o que depois resulta em elevados valores no SCC, principalmente por causa da sua influência no numerador da fórmula (ver Equação 1).

Depois de analisados os vários subcorpos, o próximo passo passou por entender como os documentos traduzidos afetariam a similaridade interna quando adicionados aos subcorpos originais

correspondentes. Para esse fim, realizamos várias experiências adicionando diferentes percentagens de documentos traduzidos, selecionados aleatoriamente, aos subcorpos originais. Mais precisamente, começamos por adicionar 10%, 20%, 30% e por fim 100%[8] dos documentos aos subcorpos originais. As Figuras 13, 14 e 15 apresentam os valores médios por documento para cada uma das diferentes percentagens. Como esperado, quanto mais documentos são adicionados menor é o NTC (veja-se Figura 13). No entanto, é necessária uma análise mais profunda dos resultados obtidos.

Embora o NTC para o espanhol seja menor quando 100% dos documentos traduzidos são adicionados ao subcorpo original, resultando em ≈9.3% menos tokens comuns por documentos, a queda em si não é muito significativa. Na verdade, o valor médio de tokens por documento aumenta ≈1.19% e ≈1.22% quando adicionados 20% e 30% dos documentos traduzidos, respetivamente. A reduzida variação nos valores do SCC e χ^2 também corrobora este facto (veja-se Figuras 14 e 15, respetivamente). O mesmo fenómeno pode-se observar para o inglês quando são adicionados os documentos traduzidos. O subcorpo original tem uma av=163,70 tokens e quando 10%, 20%, 30% e 100% dos documentos traduzidos são adicionados o NTC somente diminuiu ≈3.2%, ≈3.4%, ≈6.1% e ≈23.6%, respetivamente.

Deste modo, podemos inferir com base nos resultados estatísticos obtidos, que caso um subcorpo com mais documentos seja necessário para uma determinada tarefa em particular, os respetivos documentos originais e traduzidos em espanhol e inglês podem ser adicionados sem que a sua similaridade interna seja gravemente comprometida. Mesmo que esta junção signifique que hajam alguns documentos ruidosos dentro dos novos subcorpos, particularmente para o espanhol esta união representa um aumento no número de documentos de ≈12% e, a uma perda de somente ≈9.3% no seu grau de similaridade interno. Apesar de uma diminuição de ≈23,6% no NTC para o inglês, o aumento no número de documentos é mais significativa que para o espanhol, mais precisamente de ≈39.7%.

Relativamente ao alemão, a união dos seus subcorpos resulta numa diminuição abrupta de ≈53.4% no grau interno de similaridade. Este facto é bem visível nas Figuras 13 e 15, o que nos leva a ser ainda mais cautelosos em relação à junção dos seus dois subcorpos.

[8] O número de documentos correspondentes a estas percentagens podem ser inferidas a partir da Tabela 3.

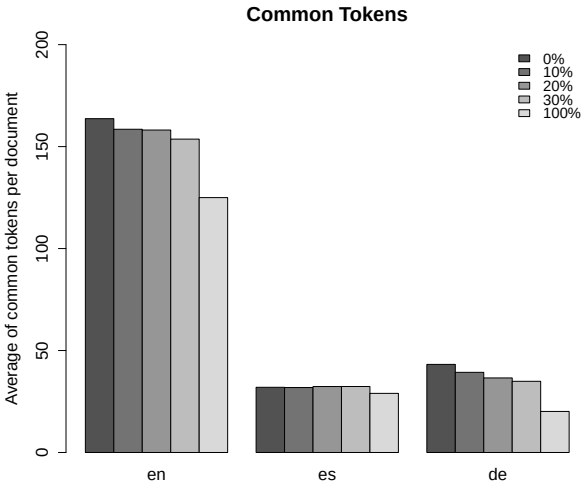

Figura 13: NTC.

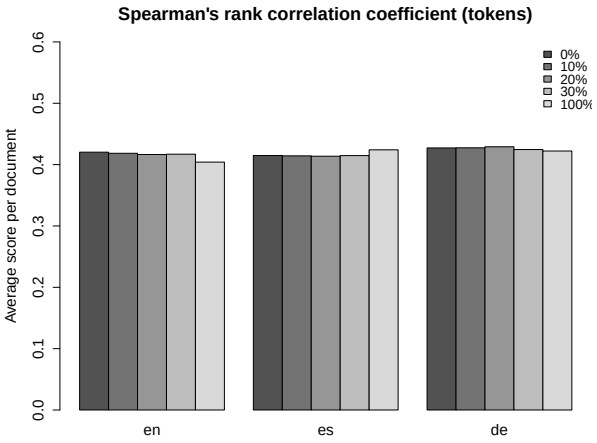

Figura 14: SCC.

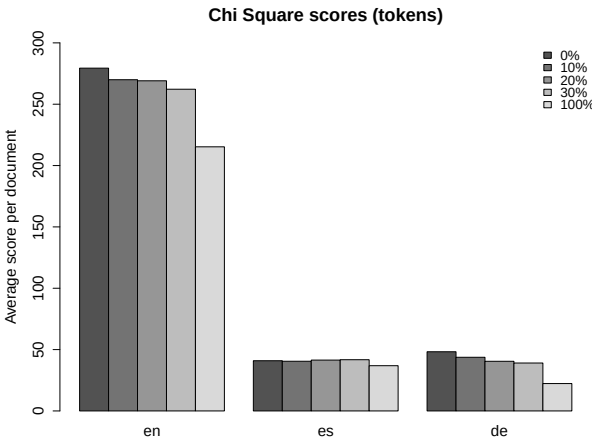

Figura 15: χ^2.

Dado os resultados analisados até ao momento podemos afirmar, de um ponto de vista teórico e estatístico, que os subcorpos *i_en_od*, *i_en_td* e *i_de_od* agregam documentos com um elevado grau de similaridade. E, pelo contrário, o mesmo não se pode afirmar para os subcorpos *i_es_od*, *i_es_td* and *i_de_td*. A segunda conclusão a retirar

dos dados analisados é que se fosse necessário um subcorpo especializado maior para o espanhol e/ou inglês, as evidências estatísticas mostram que ambos os seus subcorpos, originais e traduzidos, poderiam ser agregados sem que diminuísse drasticamente o seu grau de similaridade interno — especialmente para o espanhol em que a queda seria de apenas ≈9.3%. Contudo, é aconselhável que qualquer tipo de trabalho de investigação seja feito no subcorpo original e, somente em casos que este não seja suficientemente grande para a tarefa em questão é que se deve prosseguir com a fusão com o respetivo subcorpo traduzido.

6.2 Compilação Manual vs. Semi-automática

Esta secção tem como objetivo comparar os subcorpos compilados manualmente com os corpos compilados semi-automaticamente pelo BootCaT (ver secção 4 para mais informação sobre os diversos subcorpos). Como não existem documentos traduzidos em italiano, decidiu-se realizar as seguintes experiências apenas usando os subcorpos originais (ou seja, usando os subcorpos *i_en_od*, *i_es_od*, *i_de_od* e *i_it_od* — ver Tabela 3). Em primeiro lugar foi feita uma comparação estatística entre os dois tipos de subcorpos de modo a compreender como a sua similaridade interna difere entre si. Em seguida, analisámos se a junção dos documentos compilado semi-automaticamente com o documentos originais comprometem o grau de similaridade interno dos mesmos.

De um modo semelhante ao que foi feito na secção anterior, as Figuras 16, 17 e 18 colocam lado a lado os valores médios por documento para as várias línguas (inglês, espanhol, alemão e italiano). A primeira observação que podemos fazer sobre a Figura 16 é a surpreendente diferença no NTC entre os documentos originais e os compilados semi-automaticamente. Por exemplo veja-se o NTC médio para o subcorpo *i_en_od* de 163,70 com um σ=83,89 quando comparado com o *bc_en* que apenas tem uma av=43.28 com um σ=56.97, ou seja, ≈74% menos tokens em comum por documento em média. De facto a diferença para o italiano é ainda maior, ≈91% menos tokens em comum por documento em média para sermos mais precisos (*i_it_od*={av=101,08; σ=55,71} e *bc_it*={av=9,26; σ=10,46}). Estes resultados podem ser corroborados pela variação dos valores do SCC e pelos baixos valores do χ^2 resultantes para o *bc_en* e para o *bc_it* quando comparados com os subcorpos *i_en_od* e *i_it_od*, respetivamente (Figuras 17 e 18). Contudo, note-se que o subcorpo *bc_en* tem vários *outliers* por cima do máximo, o que significa que estes documentos têm um elevado grau de similaridade com os do subcorpo *i_en_od* e, portanto, devem ser cuidadosamente analisados pela pessoa responsável pela manutenção do corpo.

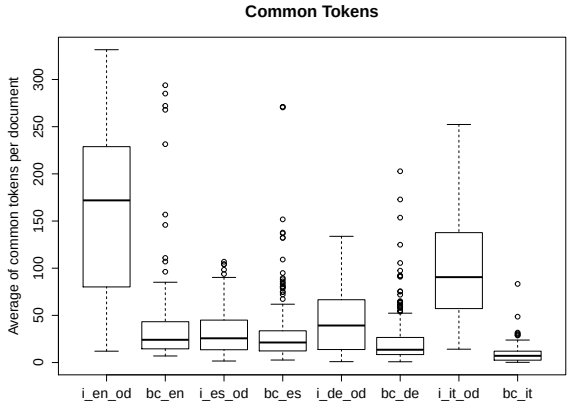

Figura 16: NTC.

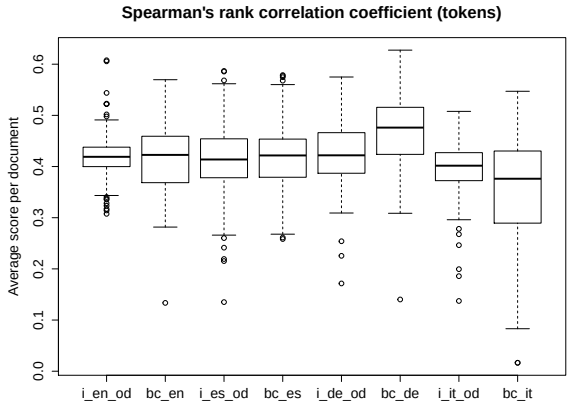

Figura 17: SCC.

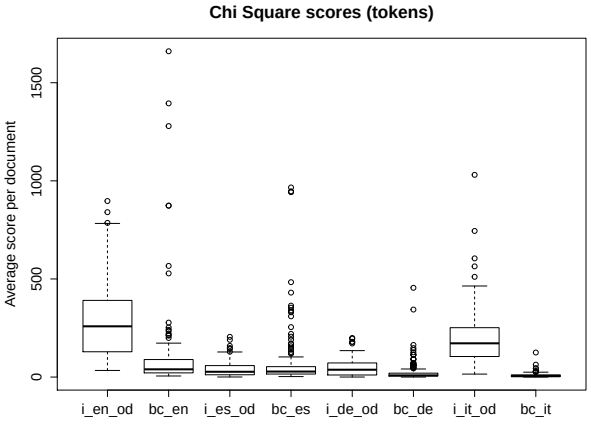

Figura 18: χ^2.

Relativamente ao subcorpo *bc_de*, este tem ≈22% menos tokens comuns por documento em média quando comparado com o subcorpo *i_de_od* (*i_de_od*={av=43,21; σ=33,52} e *bc_de*={av=23,06; σ=26,68}). Apesar desta

diferença de 22% entre os dois subcorpos em alemão, não devemos rejeitar a hipótese de que estes dois subcorpos não podem ser unidos sem diminuir drasticamente o grau de similaridade interno — no entanto, é necessária uma análise mais profunda, como veremos mais tarde nesta secção. Em relação aos subcorpos em espanhol, estes, à primeira vista, parecem conter documentos com um grau de similaridade idêntico, pois as suas médias e desvios padrão não diferem muito entre eles (i_es_od={av=31,97; σ=23,48} e bc_es={av=31,38; σ=36,51}). Além do mais, os valores do SCC e χ^2 também parecem confirmar esta hipótese (veja-se as Figuras 17 e 18).

Em suma, por um lado, os valores médios das MSD apresentados nas Figuras 16, 17 e 18 oferecem fortes evidências de que os subcorpos compilados manualmente e os compilados semi-automaticamente para o inglês e italiano não têm muito em comum. Por outro lado, as MSD sugerem que os subcorpos alemão e, principalmente os subcorpos espanhóis, partilham um elevado grau de similaridade entre os seus subcorpos e, portanto, a sua união pode ser considerada caso necessário. Para pôr à prova estes indícios, aleatoriamente selecionámos e adicionámos diferentes percentagens de documentos compilados semi-automaticamente aos subcorpos originais. A nossa hipótese é que os valores médios das MSD diminuam quanto mais documentos semi-automaticamente compilados são adicionados. Com base nos resultados anteriores, é esperada uma queda drástica para o inglês e italiano e uma queda mais suave para o alemão e, particularmente, para o espanhol.

As Figuras 19, 20 e 21 apresentam os valores médios por documento quando adicionadas diferentes percentagens de documentos semi-automaticamente compilados aos subcorpos originais. De modo a entendermos como o grau interno de similaridade varia, foram aleatoriamente selecionados e incrementalmente adicionados conjuntos de 10% aos subcorpos originais. Acima de tudo o que é importante analisar nas Figuras 19, 20 e 21 é o seguinte: i) os valores médios iniciais, ou seja os valores dos subcorpos compilados manualmente (0%); ii) como estes valores variam quando mais documentos são adicionados (de 10% a 100%); iii) e comparar o valor inicial com o valor final, ou seja quando a totalidade dos documentos semi-automáticos é adicionada ao subcorpo original (0% e 100%). Já anteriormente, quando colocámos as Figuras 16, 17 e 18 lado a lado, deu para ter uma ideia sobre o que aconteceria quando fosse feita esta união dos dois tipos de subcorpos e, de facto as Figuras 19 e 21 vêm corroborar a nossa tese inicial. Como podemos ver na Figura 19, quanto mais conjuntos de documentos são adicionados, menor é o NTC para as quatro línguas de trabalho.

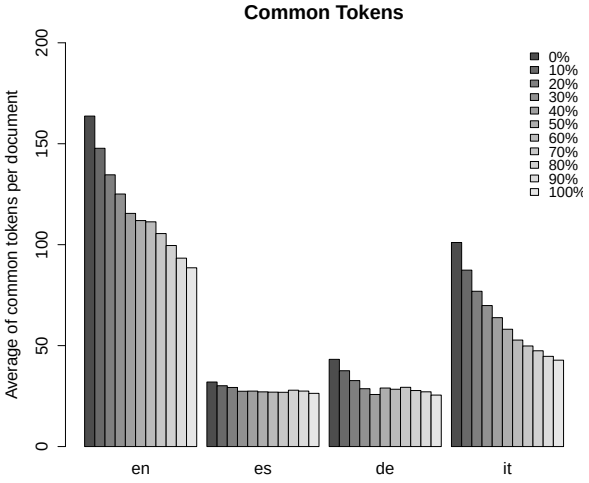

Figura 19: NTC.

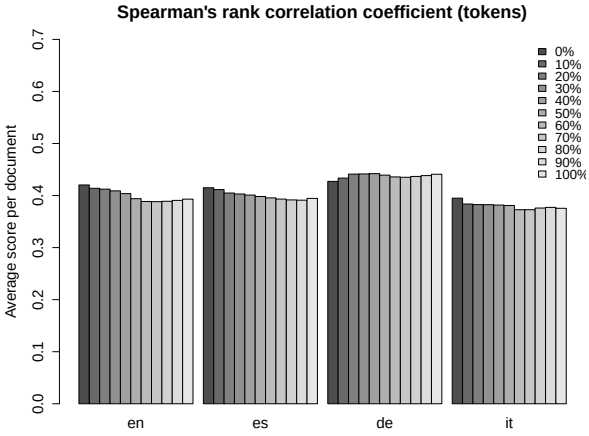

Figura 20: SCC.

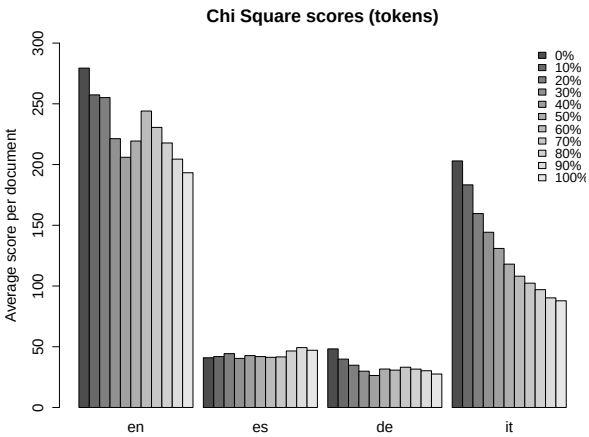

Figura 21: χ^2.

Como mencionado anteriormente, o NTC por documento para o subcorpo *i_en_od* é, em média, de 163,70. Contudo, quando o *bc_en* é adicionado — o que significa um aumento de ≈73.5% em termos de tamanho — o NTC diminui para quase metade (ou seja, há uma diminuição de ≈46%: {*i_en_od* + *bc_en*}={av=88.55}). Para o italiano a redução do NTC é ainda mais acentuada, mais precisamente de ≈58% ({*i_it_od* + *bc_it*}={av=42.79}, enquanto que o aumento no número de documentos é de ≈81.3%). E, o alemão segue a mesma tendência com uma redução no NTC de ≈41%, contudo é necessário ter em conta que esta união representa um aumento no número de documentos de ≈183.3%.

Os valores do χ^2 também apontam na mesma direção, ou seja, os valores do χ^2 diminuem em ≈31%, ≈57% e ≈43% para os subcorpos {*i_en_od* + *bc_en*}, {*i_it_od* + *bc_it*} e {*i_de_od* + *bc_de*}, respetivamente. Um fenómeno semelhante ocorre com o espanhol, observe-se a Figura 16. Contudo, e apesar da diminuição do NTC em ≈17% para o espanhol quando este sofre um aumento de ≈103.8% no número de documentos, o grau de similaridade interno parece estabilizar assim que o primeiro conjunto de documentos é adicionado, o que poderá significar que o subcorpo *bc_es* segue uma distribuição normal em termos de conteúdo, neste caso no NTC por documento. Em relação aos valores do χ^2, este sofre um aumento de ≈15%, o que mostra indícios de um aumento da similaridade interna.

De forma semelhante à conclusão retirada na secção 6.1 (quando comparámos os subcorpos originais com os traduzidos), os valores do NTC e os valores χ^2 das Figuras 19 e 21, assim como os resultados observados nas Figuras 16, 17 e 18, leva-nos a concluir que caso seja necessário um maior subcorpo especializado para o espanhol a união entre os textos originais e os compilados semi-automaticamente pode ser realizada sem que o grau interno de similaridade seja drasticamente comprometido. Ou, pelo menos, é mais aconselhável sugerir esta união do que a união dos subcorpos do italiano, do alemão ou mesmo do inglês. Embora, em geral, os valores do SCC diminuam para três das quatro línguas, estes, no entanto, não são suficientemente explícitos para nos permitir tirar uma conclusão sólida sobre os mesmos (veja-se Figura 20).

6.3 Discussão

Depois de apresentados todos os resultados estatísticos é hora de seguir em frente e analisar o problema de uma perspetiva diferente e centrarmo-nos sobre a seguinte questão: "*Devemos sempre confiar nas ferramentas semi-automáticas para compilar corpos comparáveis especializados?*". A questão em si é simples, mas como foi demonstrado nas secções anteriores, a resposta não é trivial. Por um lado, podemos assumir que as ferramentas de compilação semi-automáticas têm uma abrangência maior quando comparadas com a compilação manual, pois estas são capazes de compilar mais documentos do que um humano no mesmo espaço de tempo. Contudo, a sua precisão não é tão elevada como a de um humano — embora esta ideia seja discutível, o humano é quem tem a última palavra a dizer e, consequentemente, aquele que julga se os documentos devem pertencer ao corpo ou não. Porém, também podemos afirmar que a compilação manual nem sempre é viável, uma vez que é muito demorada e exige um grande esforço intelectual. Na verdade é que derivado à enorme quantidade de variáveis envolvidas no processo de compilação, tais como o domínio, as línguas de trabalho, os motores de busca utilizados, entre outros, que não se pode afirmar que exista uma resposta simples para a questão anterior. Por exemplo, cada motor de busca utiliza um método de indexação diferente para armazenar e encontrar páginas na rede, o que significa que diferentes motores de busca devolvem diferentes resultados. De volta à questão, e com base nos nossos resultados, o que podemos afirmar é que as ferramentas de compilação semi-automáticas podem-nos ajudar a impulsionar o processo de compilação. E, embora algumas fases do processo possam ser semi-automatizadas, estas ferramentas não funcionam corretamente sem a intervenção humana. Contudo, devemos ter sempre muito cuidado ao compilar corpos comparáveis em geral e corpos comparáveis especializados em particular, não só durante o processo inicial de *design*, mas também na última instância do processo de compilação, ou seja, ao analisar e filtrar os documentos compilados que devem fazer parte do corpo. E, é precisamente nesta etapa do processo onde a metodologia proposta neste trabalho se encaixa, podendo não só ser usada para ter uma ideia sobre os documentos em mãos, mas também para comparar diferentes conjuntos de documentos, e classificar os mesmos de acordo com o seu grau de similaridade. Deste modo, a pessoa em cargo da compilação poderá usar esta metodologia como uma ferramenta extra para a ajudar a descrever um corpo e até mesmo para decidir se um determinado documento ou conjunto de documentos devem fazer parte do mesmo ou não.

7 Conclusão

Neste artigo descrevemos uma metodologia simples, contudo eficiente, capaz de medir o grau de similaridade no contexto de corpos comparáveis. A metodologia apresentada reúne vários métodos de diferentes áreas do conhecimento com a finalidade de descrever, medir e classificar documentos com base no conteúdo partilhado entre eles. De modo a provar a sua eficácia foram realizadas várias experiências com três diferentes Medidas de Similaridade Distribucional (MSD).

Resumidamente, a primeira parte deste trabalho focou-se na análise dos diversos subcorpos compilados manualmente e as principais conclusões foram as seguintes: i) foram obtidos resultados semelhantes utilizando diferentes parâmetros de entrada para as várias MSD; ii) os documentos originais contêm um maior número de entidades comuns quando comparados com os traduzidos; e iii) as MSD sugerem que os subcorpos em inglês e italiano originais são compostos por documentos com um maior grau de similaridade em comparação com os restantes subcorpos analisados neste trabalho. O passo seguinte passou por demonstrar como os documentos traduzidos afetariam o grau de similaridade interno nos vários subcorpos originais quando unidos. Embora o grau de similaridade tenha reduzido drasticamente, ≈53,4% para o alemão após a fusão, o subcorpo espanhol e inglês diminuiu apenas ≈23,6% e ≈9,3%, respetivamente. Deste modo, demos por concluída a primeira parte deste trabalho afirmando que, caso fosse necessário um subcorpo especializado maior para o espanhol ou inglês, as MSD demonstraram que a união entre o subcorpo original e o subcorpo traduzido poderia ser realizada sem que se reduza drasticamente o seu grau interno de similaridade.

A segunda parte deste trabalho focou-se na comparação entre os documentos compilados manualmente e os documentos compilados semi-automaticamente. Mais uma vez começámos por realizar uma análise estatístico-descritiva entre os dois tipos de documentos de modo a obter uma ideia geral de como a similaridade média interna diferia entre eles. Como resultado, observou-se que os subcorpos compilados manualmente continham documentos com um maior grau de similaridade quando comparados com os correspondentes subcorpos compilados semi-automaticamente. Especialmente para o inglês e italiano, observamos que a diferença entre a média no número de entidades comuns era muito elevada, para sermos mais precisos, ≈74% e ≈91% menos entidades comuns, respetivamente. Estes valores já nos dão uma ideia sobre o que ocorreria quando uníssemos os subcorpos compilados manualmente com os semi-automáticos. De modo a demonstrar a sua veracidade, juntámos os vários subcorpos e as MSD demonstraram uma queda drástica em termos de similaridade interna. Mais precisamente, foi observada uma queda muito acentuada, na ordem dos 41%, 46% e 58% para o alemão, inglês e italiano, respetivamente, e uma queda não tão abrupta de ≈17% para o espanhol. Com estes resultados, concluímos que caso fosse necessário um subcorpo especializado maior para o espanhol, esta união deveria ser ponderada. Pois, se por um lado a similaridade interna caíra 17%, por outro, esta união aumentaria o número de documentos em ≈109.8%.

Como observação final, concluímos que as várias MSD podem ser consideradas uma ferramenta muito útil e versátil para descrever corpos comparáveis, o que na nossa opinião ajudaria em muito aqueles que compilam manualmente ou semi-automaticamente corpos a partir da Internet nas mais diversas línguas europeias. De facto, este trabalho provou que as MSD não só podem ser utilizadas para obter uma ideia sobre o corpo em mãos, mas também para medir, comparar e classificar diferentes conjuntos de documentos de acordo com o seu grau de similaridade e assim ajudar os investigadores a decidir se um determinado documento ou conjunto de documentos devem fazer parte de um dado corpo ou não.

Agradecimentos

Gostaríamos de agradecer à Bárbara Furtado e ao João Miguel Franco pelas correções ortográficas e gramaticais no artigo.

Hernani Costa é apoiado pela bolsa n. 317471 da REA do People Programme (Marie Curie Actions) da European Union's Framework Programme (FP7/2007-2013).

Este trabalho também é parcialmente apoiado pelo projeto de inovação para a educação TRADICOR (PIE 13-054, 2014-2015); pelo projeto de inovação para a educação NOVATIC (PIE 15-145, 2015-2017); o projeto de I&D INTELITERM (ref. n. FFI2012-38881, 2012-2015); o projeto de I&D LATEST (Ref: 327197-FP7-PEOPLE-2012-IEF); e o projeto de I&D TERMITUR (ref. n. HUM2754, 2014-2017).

Referências

Anthony, Laurence. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. http://www.laurenceanthony.net.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis* Bloomsbury Discourse. Bloomsbury Academic.

Barbaresi, Adrien. 2014. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. Em *9th Web as Corpus Workshop (WaC-9), 14th Conf. of the European Chapter of the Association for Computational Linguistics*, 1–8. Gothenburg, Sweden.

Barbaresi, Adrien. 2015. Challenges in the linguistic exploitation of specialized republishable web corpora. Em *RESAW Conf. 2015*, 53–56. Aarhus, Denmark. Short paper talk.

Baroni, Marco & Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. Em *4th Int. Conf. on Language Resources and Evaluation* LREC'04, 1313–1316.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek & Pavel Rychlý. 2006. WebBootCaT: instant domain-specific corpora to support human translators. Em *11th Annual Conf. of the European Association for Machine Translation* EAMT'06, 247–252. Oslo, Norway: The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, Lynne & Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Corpas Pastor, Gloria. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología* 5(1). 155–184.

Corpas Pastor, Gloria & Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). Em A. Beeby, P.R. Inés & P. Sánchez-Gijón (eds.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* Benjamins translation library, chap. 5, 75–107. John Benjamins Publishing Company.

Costa, Hernani. 2010. *Automatic Extraction and Validation of Lexical Ontologies from text*. Coimbra, Portugal: University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering. Tese de Mestrado.

Costa, Hernani. 2015. Assessing Comparable Corpora through Distributional Similarity Measures. Em *EXPERT Scientific and Technological Workshop*, 23–32. Malaga, Spain.

Costa, Hernani, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor & Ruslan Mitkov. 2015a. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. Em *9th Int. Workshop on Semantic Evaluation* SemEval'15, 96–101. Denver, Colorado: ACL.

Costa, Hernani, Gloria Corpas Pastor & Ruslan Mitkov. 2015b. Measuring the Relatedness between Documents in Comparable Corpora. Em *11th Int. Conf. on Terminology and Artificial Intelligence*, 29–37. Granada, Spain.

Costa, Hernani, Hugo Gonçalo Oliveira & Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. Em *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* ECAI'10, 23–29. Lisbon, Portugal.

Costa, Hernani, Hugo Gonçalo Oliveira & Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. Em *15th Portuguese Conf. on Artificial Intelligence*, vol. 7026 EPIA'11, 597–609. Lisbon, Portugal: Springer.

EAGLES. 1996. Preliminary Recommendations on Corpus Typology. Relatório técnico. EAGLES Document EAG-TCWG-CTYP/P. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

El-Khalili, Nuha H., Bassam Haddad & Haya El-Ghalayini. 2015. Language Engineering for Creating Relevance Corpus. *Int. Journal of Software Engineering and Its Applications* 9(2). 107–116.

Grishman, Ralph. 1997. Information Extraction: Techniques and Challenges. Em *Int. Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology* SCIE'97, 10–27. London, UK: Springer.

de Groc, Clement. 2011. Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. Em *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, vol. 1 WI-IAT'11, 497–498. Lyon, France: IEEE Computer Society.

Gutiérrez Florido, Rut, Gloria Corpas Pastor & Miriam Seghiri. 2013. Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. Em *Workshop on Optimizing Understanding in Multilingual Hospital Encounters, 10th Int. Conf. on Terminology and Artificial Intelligence*, Paris, France.

Harris, Zelig. 1970. Distributional Structure. Em *Papers in Structural and Transformational Linguistics*, 775–794. Dordrecht, Holland: D. Reidel Publishing Company.

Ibrahimov, Oktay, Ishwar Sethi & Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. Em *16th Int. Conf. on Pattern Recognition*, vol. 4, 285–288. IEEE Computer Society.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vít Suchomel. 2014. Finding Terms in Corpora for Many Languages with the Sketch Engine. Em *Demonstrations at the 14th Conf. of the European Chapter of the Association for Computational Linguistics*, 53–56. Gothenburg, Sweden: ACL.

Kilgarriff, Adam. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics* 6(1). 97–133.

Maia, Belinda. 2003. What are comparable corpora? Em Silvia Hansen-Schirra & Stella Neumann (eds.), *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, 27–34. Lancaster, UK.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics* 2(1). 133–152.

Salton, Gerard & Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24(5). 513–523.

Schmid, Helmut. 1995. Improvements In Part-of-Speech Tagging With an Application To German. Em *ACL SIGDAT-Workshop*, 47–50. Dublin, Ireland.

Singhal, Amit. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24(4). 35–42.

Propuesta recibida el mayo 2016 y aceptada para publicacin el julio 2016.

Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales

Social Network Multilingual Author Profiling using character and POS n-grams

Carlos-Emiliano González-Gallardo
LIA-Université d'Avignon,
GIL-Instituto de Ingeniería UNAM
carlos.gonzalez-gallardo@alumni.univ-avignon.fr

Juan-Manuel Torres-Moreno
LIA-Université d'Avignon,
École Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Azucena Montes Rendón
CENIDET
amr@cenidet.edu.mx

Gerardo Sierra
GIL-Instituto de Ingeniería UNAM
gsierram@iingen.unam.mx

Resumen

En este artículo presentamos un algoritmo que combina las características estilísticas representadas por los n-gramas de caracteres y los n-gramas de etiquetas gramaticales (POS) para clasificar documentos multilengua de redes sociales. En ambos grupos de n-gramas se aplicó una normalización dinámica dependiente del contexto para extraer la mayor cantidad de información estilística posible codificada en los documentos (emoticonos, inundamiento de caracteres, uso de letras mayúsculas, referencias a usuarios, ligas a sitios externos, *hashtags*, etc.). El algoritmo fue aplicado sobre dos corpus diferentes: los *tweets* del corpus de entrenamiento de la tarea *Author Profiling* de PAN-CLEF 2015 (Rangel et al., 2015) y el corpus de "Comentarios de la Ciudad de México en el tiempo" (CCDMX). Los resultados presentan una exactitud muy alta, cercana al 90 %.

Palabras clave

Minería de textos, Aprendizaje automático, Clasificación textual, n-gramas, Blogs, Tweets, Redes sociales

Abstract

In this paper we present an algorithm that combines the stylistic features represented by characters and POS n-grams to classify social network multilingual documents. In both n-gram groups a dynamic normalization by context was applied to extract all the possible stylistic information encoded in the documents (emoticons, character flooding, capital letters, references to other users, hyperlinks, hashtags, etc.). The algorithm was applied to two different corpus; *Author Profiling* of PAN-CLEF 2015 training tweets (Rangel et al., 2015) and the corpus of "Comments of Mexico City in time" (CCDMX). Results shows up to 90 % of accuracy.

Keywords

Text Mining, Machine Learning, Text Classification, n-grams, Blogs, Tweets, Social Networks

1 Introducción

La clasificación automática de texto se encarga de predecir de forma automática a cuál de las clases existentes pertenece un texto. Este modelo es creado a partir de un corpus etiquetado que contenga ejemplos de esas clases (Koppel et al., 2002).

A diferencia de la identificación de autor, que tiene como objetivo predecir si un texto pertenece o no a un autor específico, el perfilado de autor tiene como objetivo predecir si un texto pertenece o no a un grupo de autores que comparten ciertas características; como el género, la edad, el nivel educativo, la región geográfica, etc.

El interés por el perfilado de autor a partir de textos procedentes de Internet ha ido creciendo en los últimos años. Esto es debido a la gran cantidad de información que se produce continuamente en las redes sociales y los blogs. En marzo de 2016, Facebook reportó tener aproximadamente 1 090 millones de usuarios activos al día[1]; mientras que Twitter[2] 320 millones de usuario activos al mes.

Los documentos textuales producidos por los usuarios de estas redes, tienen características que los hacen difícilmente comparables con los textos literarios, documentales o ensayos en donde tradicionalmente el perfilado de autor es aplicado (Argamon et al., 2003, 2009); evitando así que

[1] http://www.facebook.com
[2] http://www.twitter.com

puedan ser analizados de forma similar (Peersman et al., 2011).

Dentro de las características que poseen los textos procedentes de Twitter y redes sociales, se encuentra su longitud, que es notablemente más corta (Peersman et al., 2011), el uso no estandarizado de mayúsculas y signos de puntuación, el gran número de errores ortográficos, etc.

Las redes sociales como Twitter tienen sus propias reglas y características que los usuarios explotan para expresarse y comunicarse entre sí. Estas reglas pueden ser aprovechadas para extraer una mayor cantidad de información estilística. (Gimpel et al., 2011) introducen esta idea para crear un etiquetador gramatical para Twitter. En nuestro caso, optamos por realizar una normalización dinámica dependiente del contexto. Esta normalización permite agrupar aquellos elementos que tengan la capacidad de proveer información estilística sin importar su variabilidad léxica. Esta fase ayuda al sistema de clasifición a mejorar su rendimiento.

El artículo está organizado de la siguiente manera: en la sección 2 hacemos una breve presentación del uso de n-gramas y etiquetas POS. En la sección 3 detallamos la metodología empleada en la normalización dinámica dependiente del contexto. La sección 4 presenta los corpus utilizados en el estudio. El modelo de aprendizaje es detallado en sección 5. Los diversos experimentos realizados y los resultados obtenidos son presentados en la sección 6. Para finalizar, en la sección 7 exponemos las conclusiones y algunas perspectivas de trabajo futuro.

2 N-gramas de caracteres y etiquetas gramaticales (POS)

Los n-gramas son un recurso de gran utilidad en el Procesamiento del Lenguaje Natural (PLN), ya que permiten la extracción de características de contenido y estilísticas a partir de los textos, que pueden ser utilizadas en tareas como resumen automático, traducción automática y clasificación textual.

Los n-gramas son secuencias de elementos de la unidad de información textual seleccionada (Manning & Schütze, 1999). Esta información cambia en función de la tarea a realizar y del tipo de información que se desea extraer. Por ejemplo, en traducción y resumen automático es común utilizar n-gramas de palabras y n-gramas de oraciones (Torres-Moreno, 2014; Giannakopoulos et al., 2008; Koehn, 2010). Dentro de la clasificación de texto, para la detección de plagio e identificación y perfilado de autor, los n-gramas de caracteres, palabras y etiquetas POS (*Part-of-Speech*) son utilizados (Doyle & Kešelj, 2005; Stamatatos et al., 2015; Oberreuter & Velásquez, 2013).

La unidades de información seleccionadas en este trabajo son caracteres y etiquetas POS. Con los n-gramas de caracteres se pretende extraer la mayor cantidad de elementos estilísticos posible: frecuencia de caracteres, uso de sufijos (género, número, tiempos verbales, diminutivos, superlativos, etc.), uso de signos de puntuación (frecuencia de uso, repetición), uso de emoticonos, etc. (Stamatatos, 2006, 2009).

Los n-gramas POS proporcionan información referente a la forma en que está estructurado el texto: la frecuencia de elementos gramaticales, la diversidad de estructuras gramaticales empleadas y la inteacción entre elementos gramaticales. Las etiquetas POS fueron obtenidas usando el etiquetador gramatical de Freeling[3]. Para controlar completamente el proceso de normalización y hacerlo independiente de un detector de nombres propios, preferimos realizar una normalización específica para estos corpus, en lugar de utilizar las funciones de Freeling (Padró & Stanilovsky, 2012).

Una etiqueta POS cuenta con varios niveles de detalle que permiten conocer los diferentes atributos de una categoría gramatical. En nuestro caso únicamente utilizamos el primer nivel de detalle que hace referencia a la categoría en sí misma (ver el cuadro 1).

Palabra: *versión*			
Atributo	**Código**	**Valor**	**Etiqueta**
Categoría	N	Nombre	
Tipo	C	Común	
Género	F	Femenino	
Número	S	Singular	*N*
Caso	0	-	
Género semántico	0	-	
Grado	0	-	

Cuadro 1: Etiquetado gramatical de la palabra *versión*.

3 Normalización dinámica dependiente del contexto

El léxico utilizado en las redes sociales es muy variado debido a la libertad que existe para codificar los mensajes. Para contrarrestar este he-

[3]Freeling está disponible en: `http://nlp.lsi.upc.edu/freeling/node/1`

cho, es necesario normalizar aquellos elementos que tengan la capacidad de proveer información estilística sin importar su variabilidad léxica: referencias a usuarios, ligas a sitios externos y *hashtags*. Este proceso denominado Normalización dinámica dependiente del contexto se separa en dos partes: Normalización del texto y Re-etiquetado POS:

- Normalización del texto

  Es común observar en redes sociales como Twitter las referencias a otros usuarios pertenecientes a la red. Esta referencia está determinada de la siguiente forma:

  `@nombre_de_usuario`

  La cantidad de posibles valores que se le pueden asignar a la etiqueta `nombre_de_usuario` es potencialmente infinita (dependiendo de la cantidad de usuarios de la que disponga la red social). Para evitar tanta variabilidad, decidimos normalizar este elemento con el fin de resaltar la intención de realizar una referencia a un usuario.

  Las ligas a sitios de Internet tienen un comportamiento similar; la cantidad de ligas a estos sitios también es potencialmente infinita. Lo importante y rescatable es el hecho de utilizar un enlace a un sitio externo, por lo que todas las cadenas de texto que cumplen con el patrón:

  `http[s]://liga_sitio_externo`

  también fueron normalizadas.

- Re-etiquetado POS

  Estos elementos también proveen información gramatical importante que es necesario conservar, pero los etiquetadores gramaticales convencionales son incapaces de detectar. Por ello, en nuestro trabajo las referencias a usuarios, las ligas a sitios Internet y los *hashtags* son re-etiquetados de tal forma que se mantenga la interacción de estos elementos con el resto de los elementos gramaticales (ver un ejemplo en el Anexo, cuadro 17).

Una arquitectura general del sistema es mostrada en la figura 1.

4 Conjunto de datos

Con la finalidad de realizar pruebas pertenecientes a diversos contextos, hemos utilizado córpora provenientes de dos redes sociales: Twitter y Facebook. El corpus multilingüe de entrenamiento PAN-CLEF 2015 (Twitter) se encuentra etiquetado por género, edad y rasgos de personalidad. El corpus de "Comentarios de la Ciudad de México en el tiempo" (CCDMX) (comentarios de Facebook) dispone únicamente de etiquetas de género en español.

4.1 Corpus PAN-CLEF (train) 2015

El corpus PAN-CLEF (train) 2015[4] (Rangel et al., 2015) está conformado por un total de 324 muestras distribuidas en cuatro idiomas: español, inglés, italiano y holandés. Cada una de las muestras se compone de aproximadamente 96 *tweets* (Nowson et al., 2015).

Con respecto al género, la distribución del corpus está equilibrada en los cuatro idiomas (50 % como "Mujeres" y 50 % como "Hombres").

	Muestras		**Total**
	Mujeres	**Hombres**	
Español	50	50	100
Inglés	76	76	152
Italiano	19	19	38
Holandés	17	17	34

Cuadro 2: Corpus PAN-CLEF (train) 2015, Distribución de muestras por género.

En el caso de español e inglés las muestras también se encuentran etiquetadas por grupos de edad: 18-24, 25-34, 35-49 y >50 años. En este caso el corpus no está equilibrado, siendo el grupo "25-34" el más numeroso, y el grupo ">50" el que cuenta con el menor número de muestras, en ambos idiomas. Ver cuadro 3.

Grupo		**Español**	**Inglés**
18-24	muestras	22	58
	porcentaje	22 %	38 %
25-34	muestras	46	60
	porcentaje	46 %	40 %
35-49	muestras	22	22
	porcentaje	22 %	14 %
>50	muestras	10	12
	porcentaje	10 %	8 %
	Total muestras	100	152

Cuadro 3: Corpus PAN-CLEF (train) 2015, Distribución de muestras por edad.

Para los cuatro idiomas se cuentan con etiquetas de clases pertenecientes a cinco rasgos de per-

[4]Sitio web del PAN: `http://pan.webis.de/`

sonalidad: extraversión, inestabilidad emocional, amabilidad, responsabilidad y apertura al cambio.

Cada rasgo fue anotado con un valor discreto comprendido entre [-0.5, +0.5] (ver Anexo, cuadro 18).

4.2 Corpus de Comentarios de la Ciudad de México en el tiempo (CCDMX)

El corpus CCDMX está compuesto por 5 979 comentarios en español mexicano, procedentes de la página de Facebook "La Ciudad de México en el tiempo"[5]. La longitud promedio de los comentarios es de 110 caracteres. El corpus CCDMX fue anotado manualmente en el Grupo de Ingeniería Lingüística (GIL) de la UNAM en 2014[6].

El corpus CCDMX se encuentra únicamente etiquetado por género, siendo ligeramente mayor la cantidad de comentarios pertenecientes a la clase "Hombres" (ver cuadro 4).

	Comentarios	**%**
Mujeres	2573	43 %
Hombres	3406	57 %
Total de muestras	5 979	100 %

Cuadro 4: Corpus CCDMX, Distribución de muestras por género.

5 Modelo de aprendizaje

Para los experimentos utilizamos un modelo clásico de aprendizaje supervisado usando *Support Vector Machines* (SVM) (Vapnik, 1998), que ha mostrado ser robusto y eficaz en diversas tareas de PLN.

En particular, para realizar los experimentos empleamos el paquete Python *SciKit Learn*[7], usando un *kernel* lineal LinearSVC (Pedregosa et al., 2011), que produjo empíricamente los mejores resultados.

5.1 Características utilizadas

Las ventanas de *n*-gramas de caracteres y etiquetas POS contempladas fueron generadas con una longitud de 1 a 3 unidades. De esta forma, por ejemplo, la palabra *"versión"* está representada por los siguientes *n*-gramas de caracteres:

{*v, e, r, s, i, ó, n, _v, ve, er, rs, si, ió, ón, n_, _ve, ver, ers, rsi, sió, ión, ón_*}

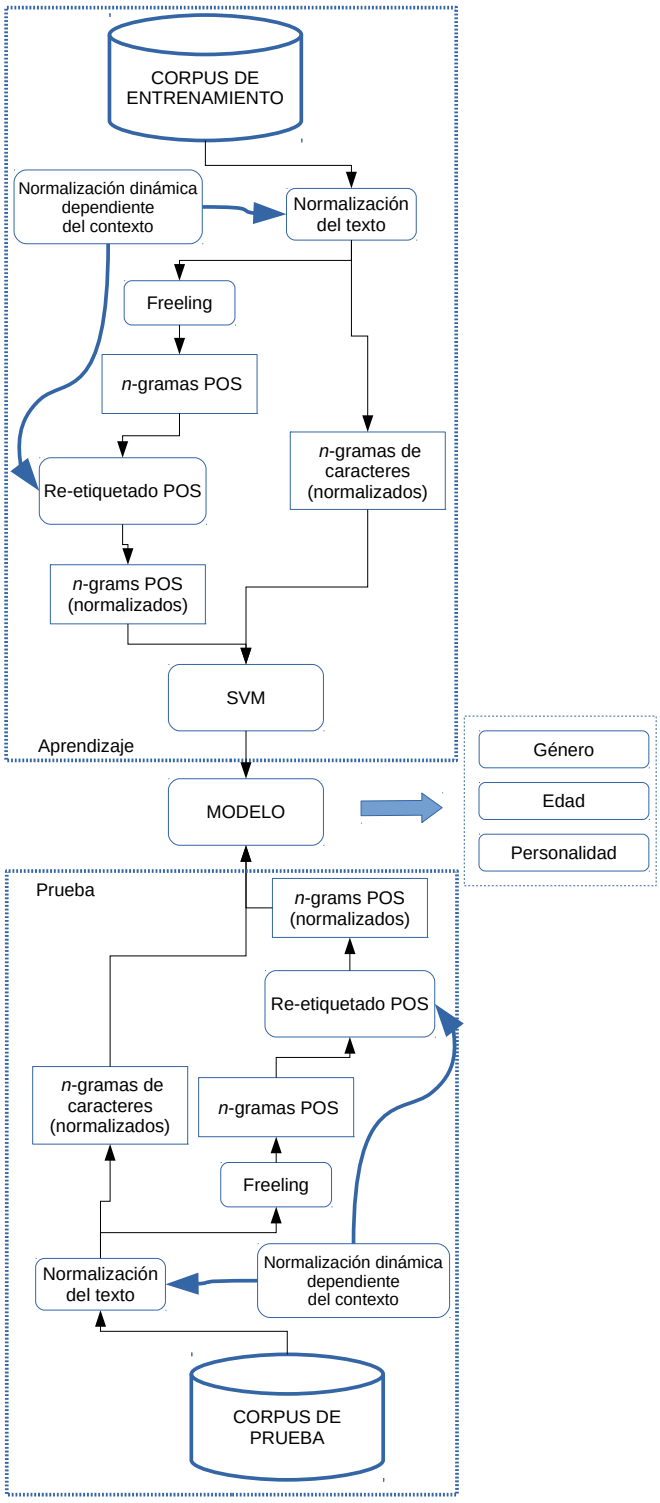

Figura 1: Arquitectura general del sistema de clasificación.

[5]Sitio web del blog: http://www.facebook.com/laciudaddemexicoeneltiempo

[6]Este corpus puede ser solicitado en el sitio web del GIL, en http://corpus.unam.mx

[7]Disponible en el sitio: http://scikit-learn.org

Y la secuencia de etiquetas POS

```
REF@USERNAME V C D N P V REF#LINK
```

está representada por los siguientes n-gramas POS:

```
{REF@USERNAME, V, C, D, N, P, V,
REF#LINK, REF@USERNAME V, V C, C D, D
N, N P, P V, V REF#LINK, REF@USERNAME
V C, V C D, C D N, D N P, N P V, P V
REF#LINK}
```

Una escala lineal de frecuencia es utilizada en todos los casos con excepción de los n-gramas POS para los textos en español, en donde se aplica una función logarítmica del tipo:

$$\log_2(1 + frecuencia) \quad (1)$$

que permite evitar una desviación en los cálculos debido a las grandes frecuencias.

5.2 Protocolo experimental

Cuatro experimentos fueron realizados con el corpus PAN-CLEF (train) 2015, uno por cada idioma. El 70 % de las muestras fue utilizado para entrenar el modelo de aprendizaje y el 30 % durante su evaluación.

Por otro lado, tres experimentos fueron realizados con el corpus CCDMX.

- En primer lugar, el 100 % de los comentarios fueron utilizados como muestras de prueba, utilizando el modelo de aprendizaje generado con las muestras en español de entrenamiento del corpus PAN-CLEF (train) 2015.
- Para el segundo experimento, se crearon muestras de 50 comentarios, juntando así 121 muestras que fueron probadas utilizando el mismo modelo de aprendizaje que el primer experimento.
- Finalmente, el tercer experimento consitió en tomar el 70 % de las 121 muestras para entrenar el modelo de aprendizaje y el 30 % para probar su desempeño.

6 Resultados

Para evaluar el desempeño del sistema en ambos corpus, varias medidas clásicas fueron implementadas:

La exactitud (Ex), precisión (Pr), cobertura (Co) y valor-F (F_1) (Manning & Schütze, 1999) fueron medidos en el corpus CCDMX para evaluar la predicción de género.

En el corpus PAN-CLEF (train) 2015, las mismas medidas fueron utilizadas para evaluar la predicción de género (español, inglés, italiano y holandés) y la edad (español e inglés).

Finalmente, para la evaluación de los rasgos de personalidad en el corpus PAN-CLEF (train) 2015, la medida RMSE (Rangel et al., 2015) fue utilizada.

6.1 Resultados sobre el corpus PAN-CLEF (train) 2015

Los cuadros 5 a 12 presentan los resultados multilingües obtenidos sobre el corpus PAN-CLEF (train) 2015.

Los casos para el experimento en italiano (tabla 9) y para el experimento en holandés (tabla 11) ameritan ser explicado. Las medidas de evaluación reportan 1 en prácticamente todos los casos; esto es debido a que la cantidad de muestras existentes eran muy pocas para italiano y holandés.

Pensamos que valdría la pena probar con una mayor cantidad de datos para validar los resultados en estos dos idiomas.

Español
Las pruebas se realizaron sobre 30 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.929	0.867	0.897	0.900
Mujeres	0.875	0.93	0.902	
18-24	0.750	1	0.857	0.800
25-34	0.750	0.875	0.807	
35-49	1	0.667	0.800	
>50	1	0.500	0.667	

Cuadro 5: Corpus PAN-CLEF (train) 2015, Resultados género y edad (español).

Rasgo	**RMSE**
Extraversión	0.106
Inestabilidad emocional	0.128
Amabilidad	0.158
Responsabilidad	0.164
Apertura al cambio	0.138
Promedio	0.139

Cuadro 6: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (español).

Inglés
Las pruebas se realizaron sobre 46 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.826	0.826	0.826	0.826
Mujeres	0.826	0.826	0.826	
18-24	0.895	0.944	0.919	0.848
25-34	0.789	0.833	0.810	
35-49	0.800	0.667	0.727	
>50	1	0.750	0.857	

Cuadro 7: Corpus PAN-CLEF (train) 2015, Resultados género y edad (inglés).

Rasgo	**RMSE**
Extraversión	0.182
Inestabilidad emocional	0.182
Amabilidad	0.150
Responsabilidad	0.123
Apertura al cambio	0.162
Promedio	0.160

Cuadro 8: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (inglés).

Italiano
Las pruebas se realizaron sobre 12 muestras.

	Pr	**Co**	F_1	**Ex**
Hombres	1	1	1	1
Mujeres	1	1	1	

Cuadro 9: Corpus PAN-CLEF (train) 2015, Resultados género (italiano).

Rasgo	**RMSE**
Extraversión	0.065
Inestabilidad emocional	0.194
Amabilidad	0.091
Responsabilidad	0.100
Apertura al cambio	0.112
Promedio	0.112

Cuadro 10: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (italiano).

Holandés
Las pruebas se realizaron sobre 10 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.833	1	0.901	0.900
Mujeres	1	0.800	0.889	

Cuadro 11: Corpus PAN-CLEF (train) 2015, Resultados género (holandés).

Rasgo	**RMSE**
Extraversión	0.118
Inestabilidad emocional	0.161
Amabilidad	0.145
Responsabilidad	0.032
Apertura al cambio	0.118
Promedio	0.139

Cuadro 12: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (holandés).

6.2 Laboratorio de evaluación PAN-CLEF 2015

En 2015 se llevó a cabo el treceavo laboratorio de evaluación organizado por PAN-CLEF[8]. La tarea de perfilado de autor consistió en predecir el género, la edad y 5 rasgos de personalidad de usuarios de Twitter a partir de los *tweets* emitidos.

El corpus de entrenamiento corresponde al corpus descrito en la sección 4.1, mientras que el corpus de prueba se encuentra constituido por 142 muestras en inglés, 88 en español, 36 en italiano y 32 en holandés (Rangel et al., 2015). Estos dos corpus constituyen el conjunto de datos oficial.

El método propuesto ne este artículo se posiciona en segundo lugar *(gonzalesgallardo15)* de la tabla general de resultados descrita en (Rangel et al., 2015). Un extracto de la misma se muestra en el cuadro 13.

Lugar	**Equipo**	**Global**
1	alvarezcarmona15	0.8404
2	gonzalesgallardo15	0.8346
3	grivas15	0.8078
4	kocher15	0.7875
5	sulea15	0.7755
...	...	...
19	bayot15	0.6178

Cuadro 13: Extracto de la tabla de resultados en (Rangel et al., 2015).

[8]Sitio web del PAN-CLEF 2015: `http://pan.webis.de/clef15/pan15-web/index.html`

6.3 Resultados sobre el corpus CCDMX

El primer experimento realizado con este corpus pretende descubrir qué tanto repercute la diferencia en el tamaño de las muestras de entrenamiento y prueba. La fase de entrenamiento fue realizada con el 70 % de las muestras en español del corpus PAN-CLEF (train) 2015. Hay que recordar que una muestra de este corpus está compuesta por aproximadamente 100 *tweets*.

Se probaron las 5 979 muestras disponibles del corpus CCDMX. Los resultados se muestran en el cuadro 14.

	Pr	**Co**	F_1	**Ex**
Hombres	0.598	0.631	0.614	0.549
Mujeres	0.474	0.439	0.456	

Cuadro 14: corpus CCDMX, Resultados experimento 1.

En el segundo experimento se optó por generar muestras de 50 comentarios, que representan un compromiso razonable entre el número de muestras y número de caracteres por muestra (aproximadamente 5 000 caracteres).

Un total de 121 muestras fueron probadas con un modelo de aprendizaje entrenado con el 70 % de las muestras en español del corpus PAN-CLEF (train) 2015. Los resultados son ligeramente mejores que en el experimento anterior, pero el cambio de dominio parece repercutir en gran medida el desempeño del sistema (ver cuadro 15).

	Pr	**Co**	F_1	**Ex**
Hombres	0.657	0.942	0.774	0.686
Mujeres	0.818	0.346	0.486	

Cuadro 15: Corpus CCDMX, Resultados experimento 2.

Por último, un tercer experimento fue realizado sobre este corpus. De las 121 muestras, el 70 % fue utilizado para entrenar el modelo de aprendizaje y el 30 % para medir su desempeño.

Estos últimos resultados obtenidos son mucho mejores que los anteriores, reafirmando la hipótesis de que el cambio de dominio afecta en gran medida el desempeño del sistema presentado (ver cuadro 16).

	Pr	**Co**	F_1	**Ex**
Hombres	0.950	0.900	0.924	0.920
Mujeres	0.880	0.940	0.909	

Cuadro 16: Corpus CCDMX, Resultados experimento 3.

7 Conclusiones y trabajo futuro

El uso de n-gramas de caracteres y n-gramas de etiquetas POS, como lo muestra los resultados, es una buena opción en textos densos debido a su capacidad de extracción de información.

En el caso de n-gramas de caracteres, fue posible extraer emoticonos, exageración de signos de puntuación (inundamiento de caracteres), uso de letras mayúsculas y todo tipo de información emocional codificada en los *tweets* y en los comentarios de Facebook.

Con los n-gramas de etiquetas POS, para el español y el inglés fue posible capturar los subconjuntoss más representativos de dos y tres elementos gramaticales. En el caso del italiano y el holandés se pudieron capturar los elementos gramaticales más frecuentes.

El algoritmo de clasificación presentado muestra ser bastante eficaz para detectar el genero, aunque un poco menos adecuado en las tareas de clasificación de la edad. Una idea interesante a desarrollar en un trabajo futuro podría ser la traducción de los emoticonos usados en las redes sociales en términos que puedan ser procesados con los mismos algoritmos de este artículo. Así la frase:

"Estoy muy feliz :) :)"

cuyas etiquetas gramaticales son:

"V R A F F F F"

sería procesada como:

```
V R A EMOT#H_SMILE EMOT#H_SMILE
```

Pensamos que esta estrategia podría mejorar aún más los resultados del sistema de clasificación.

Otro estudio en el corpus CCDMX podría consistir en agrupar el conjunto de comentarios en grupos de tamaños variables, por ejemplo: $1, 2, 4, 8, \ldots, n2^n$ comentarios y medir su impacto en el desempeño del algoritmo.

El enfoque multilingüe del algoritmo da la oportunidad de ser aplicado en tareas que involucren la detección de género o edad en opiniones dentro de redes sociales (Cossu et al., 2014, 2015).

Agradecimientos

Este trabajo fue parcialmente financiado por el proyecto CONACyT-México No. 215179 "Caracterización de huellas textuales para el análisis forense". Igualmente agradecemos el financiamiento del proyecto Europeo CHISTERA CALL - ANR: *Access Multilingual Information opinionS (AMIS)*, (Francia - Europa).

Referencias

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

Argamon, Shlomo, Moshe Koppel, James Pennebaker & Jonathan Schler. 2009. Automatically profiling the author of anonymous text. *Communications of the ACM* 52(2). 119–123.

Cossu, Jean-Valère, Eric SanJuan, Juan-Manuel Torres-Moreno & Marc El-Bèze. 2015. Multidimensional reputation modeling using microblog contents. En F. Esposito, O. Pivert, M.-S. Hacid, W. Z. Rás & S. Ferilli (eds.), *Foundations of Intelligent Systems: 22nd International Symposium, ISMIS 2015*, 452–457. Springer.

Cossu, Jean-Valère, Rocio Abascal-Mena, Alejandro Molina, Juan-Manuel Torres Moreno & Eric SanJuan. 2014. Bilingual and Cross Domain Politics Analysis. *Research in Computing Science* 1(85). 9–19.

Doyle, Jonathan & Vlado Kešelj. 2005. Automatic Categorization of Author Gender via N-Gram Analysis. En *6th Symposium on Natural Language Processing, SNLP*, n/a.

Giannakopoulos, George, Vangelis Karkaletsis & George Vouros. 2008. Testing the use of n-gram graphs in summarization sub-tasks. En *Text Analysis Conference*, 158–167.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 42–47. ACL.

Koehn, Philipp. 2010. *Statistical machine translation*. New York, NY, USA: Cambridge University Press 1st edn.

Koppel, Moshe, Shlomo Argamon & Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4). 401–412.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Nowson, Scott, Julien Perez, Caroline Brun, Shachar Mirkin & Claude Roux. 2015. XRCE Personal Language Analytics Engine for Multilingual Author Profiling—Notebook for PAN at CLEF 2015. En L. Cappellato, N. Ferro, G. Jones & E. SanJuan (eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*, vol. 1391 CEUR Workshop Proceedings, .

Oberreuter, Gabriel & Juan D. Velásquez. 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications* 40(9). 3756–3763.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, ELRA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Machine Learning Research* 12. 2825–2830.

Peersman, Claudia, Walter Daelemans & Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. En *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37–44. ACM.

Rangel, F., P. Rosso, M. Potthast, B. Stein & W. Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. En Cappellato L., Ferro N., Gareth J. & San Juan E. (eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*, online.

Stamatatos, Efstathios. 2006. Ensemble-based Author Identification Using Character N-grams. En *3rd International Workshop on Text-based Information Retrieval*, 41–46.

Stamatatos, Efstathios. 2009. A Survey of Modern Authorship Attribution Methods. *American Society for information Science and Technology* 60(3). 538–556.

Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso & Benno Stein. 2015. Overview of the pan/clef 2015 evaluation lab. En *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 518–538. Springer.

Torres-Moreno, Juan-Manuel. 2014. *Automatic text summarization*. London: Wiley-Sons.

Vapnik, Vladimir N. 1998. *Statistical learning theory*. New York: Wiley-Interscience.

Anexo

En este anexo presentamos algunos ejemplos de normalización dinámica, y una distribución de muestras por rasgos de personalidad en el corpus PAN-CLEF 2015.

tweet original	@username creo que esta versión la supera... ...http://t.co/peOlOweM Lo va petar en la #feriaJaen2012
tweet normalizado	@us creo que esta versión la supera... ...htt Lo va petar en la #feriaJaen2012
tweet original (POS)	F N V C D N P V N N V V S D F N
tweet normalizado (POS)	REF@USERNAME V C D N P V... ...REF#LINK N V V S D REF#HASHTAG

Cuadro 17: Normalización dinámica dependiente del contexto.

		Rango								
Idioma	Rasgo	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
Español	Extraversión	3		5	5	28	32	9	9	9
	Inestabilidad emocional	2	10	25	9	12	19	10	10	2
	Amabilidad		3	16	6	16	40	12	2	5
	Responsabilidad		2		21	7	20	12	21	17
	Apertura al cambio			7	10	37	15	9	14	8
Inglés	Extraversión	1	4	10	17	41	37	20	13	9
	Inestabilidad emocional	11	5	22	9	19	37	19	18	12
	Amabilidad	5	2	12	19	44	46	13	7	4
	Responsabilidad		1	4	30	38	27	33	12	7
	Apertura al cambio			2	1	47	39	23	19	21
Italiano	Extraversión				8	13	9		3	5
	Inestabilidad emocional		1	3	3	8	4	12	5	2
	Amabilidad			1	3	11	9	7		7
	Responsabilidad				3	18	6	5	6	
	Apertura al cambio				1	14	9	2	7	5
Holandés	Extraversión				3	5	11	7	6	2
	Inestabilidad emocional		1	5	3	3	4	6	8	4
	Amabilidad		2	1	5	10	10	2	4	
	Responsabilidad			2	4	15	6	5	2	
	Apertura al cambio					4	11	4	12	3

Cuadro 18: Corpus PAN-CLEF 2015, Distribución de muestras por rasgos de personalidad.

Projetos, Apresentam-Se!

Propuesta recibida el mayo 2016 y aceptada para publicacin el julio 2016.

Propuesta de clasificación de un banco de voces con fines de identificación forense

A proposal to classify a voice database for forensic identification

Fernanda López-Escobedo
Universidad Nacional Autónoma de México
flopeze@unam.mx

Julián Solórzano-Soto
Universidad Nacional Autónoma de México
jsolorzanos@iingen.unam.mx

Resumen

En este artículo se presenta el proyecto que se desarrolla para proponer una clasificación de un banco de voces con fines de identificación forense. Se expone la manera en que la información lingüística puede ser utilizada en una base de datos para reducir el número de falsos positivos y falsos negativos que resultan cuando se llevan a cabo comparaciones automatizadas para la identificación forense de voz. En particular, se abordan los fenómenos fonéticos que se han propuesto para realizar una clasificación de un banco de voces en este nivel de la lengua. A partir de esta información se describe cómo construir un modelo de base de datos y el tipo de búsquedas que se espera lograr.

La propuesta de generar descriptores lingüísticos para la clasificación de un banco de voces pretende ser una metodología que permita coadyuvar en la impartición de justicia en México y otros países de habla hispana.

Palabras clave

sociolingüística, identificación forense de voz, variación, banco de voces, español

Abstract

This article describes the project developed to classify a voice database for forensic identification. It exposes how the linguistic information can be used in a database to reduce the number of false positives and false negatives when an automatized forensic voice comparison is carried out. In particular, the phonetic phenomena that have been proposed for classifying a voice database in this level of language are reported. From this information it describes how to build a database model and the type of search to be achieved.

The proposal to create linguistic descriptors to classify a voice database is intended as a methodology to assist in the administration of justice in Mexico and other Spanish-speaking countries.

Keywords

sociolinguistics, forensic voice identification, variation, voice database, spanish

1 Introducción

Actualmente, los laboratorios de análisis forense de voz de los servicios periciales en México cuentan con bancos de voces que están organizados sin tomar en cuenta las variantes lingüísticas de cada hablante. Las voces se clasifican únicamente por género y número de averiguación previa o carpeta de investigación.

Por ejemplo, Aguilar (2011), quien fue Director General de la Coordinación de Servicios Periciales de la la Procuraduría General de la República (PGR) por más de 15 años, menciona que dicha institución cuenta con cuatro bases de datos: huellas dactilares, identificación balística, genética forense y análisis de voz. Respecto a esta última comenta que el reconocimiento se realiza de manera automática con base en las características acústicas de la persona a partir de ondas sonoras, pero no hace referencia a ningún aspecto lingüístico de la voz.

Recientemente, Morrison et al. (2016) realizó una encuesta para conocer cómo las agencias criminales de diferentes partes del mundo llevan a cabo el reconocimiento forense de voz. Uno de los resultados que reporta es que los sistemas automáticos de comparación forense de voz que más se utilizan son BATVOX, seguido de IKAR Lab. En México, las agencias de investigación que cuentan con análisis forense de voz utilizan alguno de estos dos sistemas. Por ejemplo, la PGR cuenta con el sistema BATVOX, mientras que la Procuraduría General de Justicia de la Ciudad de México cuenta con el sistema IKAR Lab.

Ambos sistemas tienen la capacidad de clasificar las muestras de voz asociándoles determinados atributos, lo que facilitaría la aplicación de los resultados obtenidos en este proyecto. No se cuenta con documentación que señale la manera en que las instituciones mencionadas utilizan las ventajas que ofrecen estos sistemas para potenciar las bases de datos con las que cuentan, pues

se trata de información confidencial. Sin embargo, los expertos que en ellas laboran expresan que las muestras de voz no se clasifican a partir de información lingüística. Prueba de ello es que los peritajes que realizan no contienen una descripción lingüística de la voz que se está analizando, y no existen expertos lingüistas laborando en dichas instituciones. El perfil del perito que realiza análisis de voz en México es el de un ingeniero experto en análisis de audio.

Morrison et al. (2016) añade que de las 44 agencias que respondieron la encuesta, 28 cuentan con una base de datos. Agrega que la mayoría contienen grabaciones de sospechosos pero pocas (menos de la mitad) cuentan además con voces de personas condenadas o procesadas por algún delito, así como de grabaciones de la población. Dicho autor también reporta las características técnicas con las que se realizan las grabaciones que están contenidas en las bases de datos (llamadas telefónicas, entrevistas con sospechosos, audios de internet, etc.), el número de lenguas incluidas y el número de hablantes. Sin embargo, no se menciona nada respecto a la clasificación de estas bases de datos a partir de información social o lingüística.

2 Reconocimiento forense de voz

Una de las aplicaciones del área de reconocimiento automático de hablantes es la del reconocimiento en el contexto forense. El método que se utiliza consiste en comparar una muestra de voz de un hablante desconocido (dubitada) contra un conjunto de referencia compuesto por muestras de voz de hablantes conocidos (indubitadas); es decir, una base de datos. No obstante, su aplicación en el ámbito forense presenta situaciones complejas. Rose (2002) hace la diferencia entre una base de datos que representa un conjunto de referencia cerrado y una que representa un conjunto abierto. La diferencia se encuentra en que, mientras en el conjunto cerrado se sabe que el hablante de la muestra dubitada es uno de los hablantes del conjunto de referencia, en el conjunto abierto el hablante de la muestra dubitada puede o no estar en el conjunto de referencia. La primera situación representa una tarea más sencilla para el reconocimiento pues es posible determinar que el hablante de la muestra dubitada será el que presente la distancia más pequeña respecto al conjunto de muestras indubitadas. En el ámbito forense esta situación será poco común. Los sistemas automáticos de comparación forense de voz cuentan con un conjunto abierto de muestras de voz indubitadas. Incluso, algunas veces parte de estas muestras de voz son también dubitadas (por ejemplo, llamadas telefónicas interceptadas). En este último caso, si al realizar la comparación la muestra dubitada resulta similar a una muestra también dubitada del conjunto de referencia, el resultado es útil para reunir información forense de inteligencia pero no para identificar a un hablante, como lo señala Ramos (2007).

Otra característica del reconocimiento automático de hablantes es la dependencia del texto. Rose (2002) menciona que existen sistemas que son dependientes del texto y sistemas que no lo son. Nuevamente, la segunda situación resulta más compleja y es la que se presenta en el ámbito forense. Por ejemplo, las diferencias acústicas entre dos muestras de voz de un mismo hablante diciendo la palabra “mesa”, serán menores a las que presenta en dos muestras de voz produciendo palabras diferentes “mesa” y “silla”.

Ramos (2007) menciona que la producción de habla es un proceso complejo pues depende de muchas variables que incluyen factores sociolingüísticos como nivel educativo, el contexto lingüístico, diferencias dialectales, así como cuestiones fisiológicas. El estado del arte en los sistemas automáticos de comparación forense de voz o reconocimiento de hablantes se ha concentrado en desarrollar sistemas que analizan la señal acústica para realizar la comparación entre la muestra de voz dubitada y el conjunto de referencia. Dichos sistemas se caracterizan por realizar la comparación contra un conjunto de referencia abierto y por realizar el reconocimiento independientemente del texto. No obstante, y a pesar de que la señal acústica proporciona información sobre distintos factores, sus características están directamente relacionadas con cuestiones fisiológicas acerca de la configuración del tracto vocal.

El proyecto que aquí se presenta pretende incluir información sobre diferencias dialectales con el fin de robustecer los sistemas automáticos de comparación forense de voz. Actualmente no existe un análisis lingüístico previo que permita realizar la confronta únicamente contra una parte del conjunto de referencia que se caracterice por contener voces que compartan características lingüísticas con la muestra de voz dubitada. Contar con información que permita generar grupos en el conjunto de referencia de muestras de voz reduciría el número de falsos positivos y falsos negativos que se generan con las confrontas o comparaciones automatizadas.

Por otro lado, debido a que en muchas ocasiones las grabaciones dubitadas se encuentran en formatos digitales comprimidos, las propiedades acústicas se ven muy afectadas, lo que impi-

de llevar a cabo la comparación. La información sobre las variantes lingüísticas que caracterizan las distintas voces contribuiría no solamente a la clasificación del conjunto de referencia, sino también a reforzar los dictámenes que actualmente se realizan pues se podrían llevar a cabo comparaciones no automatizadas basadas en parámetros lingüísticos, aún y cuando las características del formato de audio no lo permitan.

El objetivo de este trabajo consiste en describir los parámetros lingüísticos que se han definido en el nivel fonético para la creación de catálogos de descriptores lingüísticos en un banco de voces. Es importante resaltar que el objetivo de este proyecto es generar el conocimiento para clasificar un banco de voces que sirve como conjunto de referencia en una comparación automatizada, y no así generar el banco de voces, pues como ya se mencionó, las agencias de investigación en México ya cuentan con grandes bases de datos de voz.

3 División dialectal del español

Desde 1918 con la publicación del *Manual de Pronunciación española* de Navarro hasta el día de hoy, los estudios lingüísticos sobre la pronunciación del español han tenido grandes avances. El *Manual de Pronunciación española* tenía por objeto "describir breve y sencillamente la pronunciación española" pues "sabido es que la lengua española presenta importantes diferencias de pronunciación, no sólo entre los diversos países en que se habla, sino entre las regiones de un mismo país, y frecuentemente entre las comarcas y lugares de una misma región" (1918, pg. 5).

Moreno Fernández (1993, pg. 15) señala que "los individuos, al hablar entre sí, son capaces de distinguir a los que pertenecen a su misma comunidad de los que son ajenos a ella. Los límites de una comunidad pueden ser locales, regionales, nacionales o incluso supranacionales y sus miembros generalmente conocen o intuyen el alcance de la conducta lingüística que los caracteriza. Esto nos llevaría al concepto de dialecto: los hablantes pueden sentirse miembros de una comunidad dialectal". Sin embargo, "aunque una persona tenga conciencia de su pertenencia a una comunidad, también es capaz de identificar dentro de ella variantes internas de carácter geolingüístico o sociolingüístico, así como de reconocer cuáles son los usos más prestigiosos de su variedad y apreciar la relación histórica de su habla con otra".

El primer lingüista en proponer una división dialectal de México fue Henríquez Ureña en 1921. Propone seis grandes zonas dialectales que divide en:

- El territorio hispánico de los Estados Unidos,
- El norte de la República Mexicana,
- La altiplanicie del Centro, donde se ubica la Ciudad de México,
- Las tierras calientes de la costa oriental, en particular Veracruz y Tabasco,
- La península de Yucatán, donde ejerce influencia el maya,
- La América Central, comenzando en el Estado mexicano de Chiapas.

"Lo que a juicio de Henríquez Ureña caracteriza como zona dialectal a cada una de las regiones geográficas enumeradas es el vocabulario, pues reconoce que no hay uniformidad fonética en ninguna de ellas" (Moreno de Alba, 1998, pg. 157). La principal crítica que se le hace a Henríquez Ureña es que se trata de una clasificación basada únicamente en cuestiones léxicas. En este sentido resulta pertinente resaltar que para determinar una zona dialectal se deben tomar en cuenta las características presentes en diferentes niveles: el nivel fonético que estudiará los sonidos de la lengua; el nivel morfológico que estudiará la estructura de las palabras, su constitución interna y sus variaciones; el nivel sintáctico que estudiará la manera en la que se combinan las palabras; y el nivel léxico que estudiará el uso de determinados vocablos. En este trabajo se presentan únicamente los descriptores lingüísticos correspondientes al nivel fonético; no obstante, el objetivo del proyecto es proponer la clasificación de una base de datos con descriptores en todos los niveles, sirviendo el nivel fonético como ejemplo de lo que se espera lograr.

Al trabajo de Henríquez Ureña siguieron otros como *La pronunciación del español de América: Ensayo histórico-descriptivo* de Canfield (1962), *El problema de la división del español americano en zonas dialectales* de Rona (1964) y *Dialectología hispanoamericana: Teoría, descripción, historia* de Zamora & Guitart (1982). Además, se llevaron a cabo investigaciones cuyo objetivo era describir el habla de algunas zonas de la República Mexicana, por ejemplo el trabajo de Matluck (1951) sobre *La pronunciación en el español del Valle de México, Fonología del español hablado en la Ciudad de México: Ensayo de un método sociolingüístico* de Perissinotto (1975), *El habla de Tepotzotlán* de Cortichs de Mora (1951), *El habla de Guanajuato* de Boyd-Bowman (1960),

El español de Jalisco de Cárdenas (1967), entre otros.

Sin embargo, el estudio más completo que se ha hecho sobre el español de México es el que comienza el lingüista Lope Blanch (1970) con el objetivo de "iniciar una serie de investigaciones, conducentes a reunir los datos lingüísticos –fonéticos, gramaticales y léxicos- necesarios para determinar cuáles son las principales modalidades dialectales existentes hoy en el país" (Lope Blanch, 1970, pg. 4). "Los datos recogidos presentaron gran abundancia, variedad y riqueza y en consecuencia el proyecto inicial de delimitar las zonas dialectales fue superado y se transformó en el levantamiento de un Atlas general del español de México" (Espejo, 1998, pg. 119). Se estudiaron 193 localidades y se analizaron al menos 3 informantes por localidad, de distinto nivel sociocultural, edad y sexo. Para el levantamiento de los datos se realizaron cuestionarios y entrevistas, algunas de las cuales se grabaron en cintas magnetofónicas. La información recabada se registró en mapas que dan cuenta de la variedad lingüística del español de México.

A pesar de que los datos recopilados en dicho trabajo resultan de gran utilidad en el estudio de las variantes dialectales del español de México, los años han transcurrido y con ello se debe asumir un cambio lingüístico que probablemente ya no esté reflejado en los datos recopilados hace casi más de cuarenta años.

En el trabajo que aquí se presenta se realizó una revisión bibliográfica de los fenómenos fonéticos que se han estudiado, principalmente en el español de México, y de un modo más general en el español de América y en el español peninsular, con el fin de determinar cuáles resultan pertinentes para la clasificación del banco de datos. Una de las principales referencias ha sido el *Atlas lingüístico de México*, sin embargo somos conscientes de la necesidad de corroborar los datos a través del levantamiento de grabaciones en las diferentes regiones a las que se hace referencia.

3.1 Descriptores fonéticos

Los fenómenos que se proponen como descriptores lingüísticos en el nivel fonético son 17:

- Articulación apicoalveolar de /s/
- Articulación fricativa de /t͡ʃ/
- Asibilación de róticas
- Aspiración de /s/ implosiva
- Aspiración de /x/
- Aspiración y/o velarización de /f/
- Bilabialización de /f/
- Confusión de líquidas
- Debilitamiento y/o pérdida de /d/ intervocálica y final de sílaba
- Nasalización vocálica
- Realización fricativa de /j/
- Seseo y ceceo
- Variante oclusiva de /b, d, g/ en posición intervocálica
- Velarización de nasal implosiva
- Velarización de róticas
- Vocalización de líquidas
- Yeísmo

Resulta interesante resaltar que en el caso de los fenómenos que afectan a las vocales únicamente se ha incluido el de la nasalización. A pesar de que existen otros fenómenos que se han estudiado ampliamente como la diptongación de hiatos, el debilitamiento vocálico o el cierre vocálico, son fenómenos que podrían resultar complejos en el análisis auditivo hecho por un perito en acústica forense.

Este último criterio ha resultado fundamental en la elección de los fenómenos pues el objetivo es clasificar un banco de voces con fines forenses. Por lo tanto, los usuarios de dicho banco, y a su vez responsables de clasificar las voces, no serán expertos en lingüística sino peritos en acústica forense que, en la mayoría de los casos, tienen una formación en el área de ingeniería. Así, los fenómenos fonéticos que servirán como descriptores lingüísticos deben ser fácilmente identificables en un análisis auditivo por una persona no experta en fonética, condición que consideramos cumplen los 17 fenómenos que aquí se presentan.

3.2 Información diatópica y diastrática de los fenómenos

Una vez identificados los fenómenos se elaboró una breve descripción de su realización y se hizo una tabla de cada uno de ellos con información sobre variantes diatópicas y/o diastráticas según se ha descrito en las diferentes referencias bibliográficas. Consideramos importante que los fenómenos describan algunas de estas variantes o ambas pues es la manera en la que se puede caracterizar un individuo no sólo para clasificar su voz en la base de datos, sino también para poder realizar una confronta lingüística, en el caso

Variante diatópica	Variante diastrática	Notas
México: es más frecuente en el noroeste del país; en menor medida, se registra en la región occidental del centro de México (Moreno de Alba, 1994; Lope Blanch, 1990)	Está presente en todos los niveles sociales pero es más frecuente en las mujeres (Lope Blanch, 1990)	
España (Andalucía)	Propia del nivel sociocultural bajo (Jiménez, 1999)	En Cádiz coexisten tanto la africada como la fricativa; sin embargo, entre los hablantes cultos, la segunda es rechazada (Jiménez, 1999)
Chile	Propia del nivel sociocultural bajo (Vivanco, 1999)	Motivó el surgimiento de una nueva variante (Vivanco, 1999)
Panamá: en la capital se registra la variante totalmente fricativa (Aleza & Enguita, 2002)	Es más frecuente en el nivel medio; y parece ser una marca distintiva de sexo (Aleza & Enguita, 2002)	
Puerto Rico: en el área metropolitana se registra la variante fricativa (López, 1983; Navarro, 1948)	Es más frecuente en las mujeres (López, 1983)	Algunos estudios indican que es un fenómeno propio de las generaciones jóvenes (Cedergren, 1973); y otros aseguran lo contrario (López, 1983)

Cuadro 1: Descripción del fenómeno articulación fricativa de /t͡ʃ/

de que las propiedades acústicas de la grabación estén muy afectadas.

Por ejemplo, en el caso del fenómeno "articulación fricativa de /t͡ʃ/" se cuenta con una descripción general que se narra de la siguiente manera: el español cuenta con un solo fonema africado, el linguopalatal sordo /t͡ʃ/. Este se caracteriza por su articulación doble: oclusión + fricción. Generalmente, el período de oclusión es mayor que el de la fricción; sin embargo, hay ciertas áreas dialectales donde se registran variantes (Quilis, 1999). "En la medida en que prevalezca el momento oclusivo sobre el fricativo se tratará de un sonido más tenso; si pasa lo contrario y el momento fricativo predomina, la articulación será menos tensa, más relajada" (Moreno de Alba, 1994, pg. 193).

Respecto a la caracterización diátopica y diastrática del fenómeno se cuenta con la siguiente información. En México este fenómeno se atribuye como característico del habla del noroeste del país: Baja California, Baja California Sur, Sonora, Chihuahua y una parte de Durango (Mendoza, 2004). No obstante, también se registra como poco frecuente en otras regiones de México, por ejemplo: Tacámbaro, Michoacán; Cihuatlán, Jalisco; Tepic y Acaponeta, Nayarit; la región norte de Veracruz (Moreno de Alba, 1994), entre otros.

En el Atlas lingüístico de México (mapa 35) el fenómeno tiene el mayor índice en Baja California Sur (60 %) y Guaymas, Sonora (50 %); mientras que el resto de los estados del noroeste presentan un índice menor: Baja California Norte (7.5 %), Chihuahua (20 %), Durango (5 %), etc. Otras localidades de la costa del Pacífico registran lo siguiente: Tepic (7.5 %), Tuxpan (30 %) y Acaponeta (30 %).

Aproximadamente, el 75 % de los informantes que registraron la /t͡ʃ/ fricativa fueron mujeres de todos los niveles escolares: analfabeto, semianalfabeto, medio, semiculto y culto.

Asimismo, se cuenta con información sobre este fenómeno en España y otras partes de América Latina. En Andalucía se registra el fenómeno de fricativización conforme se desciende en la escala social o se reduce la edad de los hablantes. Es decir: la articulación africada es más frecuente y prestigiosa. Por lo tanto, penetra en los estratos altos y medios (94 % y 90 %, respectivamente); mientras que en el nivel sociocultural bajo solo alcanza 68 % (Jiménez, 1999). Asimismo, la variante en cuestión es una marca que distingue entre sexos: el alófono fricativo es casi exclusivo del habla masculina (Moya & Weiderman, 1995). En Cádiz coexisten tanto la africada como la fricativa y su índice de aceptación es similar; sin embargo, entre los hablantes cultos la segunda realización no es aceptada.

En Chile, la articulación fricativa es propia de los niveles socioculturales bajos. Por ende, ha sido ampliamente estigmatizada. Este hecho motivó el surgimiento de una nueva variante, como respuesta al rechazo del alófono fricativo. La variante emergente se caracteriza por ser una realización muy marcada de la africada: está ligera-

mente más adelantada que /t͡ʃ/, tiene un periodo de oclusión más extenso (Vivanco, 1999).

En Panamá, en la capital se registra la variante totalmente fricativa (Vaquero, 1996). Los estudios concluyen que los hombres adoptan, en mayor medida que las mujeres, la articulación africada. Es decir: son la mujeres quienes tienden más a fricativizar el fonema (39.5 %) que los hombres (13.2 %) (Cardona, 2010). Asimismo, la articulación fricativa es más común en la clase media que en otras clases sociales (Aleza & Enguita, 2002).

En Puerto Rico, en el área metropolitana se presenta la variante adherente (con predominio del elemento fricativo) (López, 1983; Navarro, 1948). Esta variante es más frecuente en el sexo femenino, como ocurre en Panamá; asimismo, el fenómeno no es propio de las generaciones jóvenes (López, 1983).

Una vez descrita la información fonética del fenómeno, ésta se acomoda en una tabla como la que se muestra en el Cuadro 1. A partir de la información contenida en la tabla se definen los atributos de identidad en la base de datos, así como la asociación que existirá entre las diferentes entidades.

4 Diseño de la base de datos

El diseño de la base de datos debe responder a la necesidad de clasificar las grabaciones del banco de voces de manera que éstas puedan ser recuperadas de acuerdo a una consulta que incluya los descriptores mencionados anteriormente junto con su información diatópica y diastrática.

El análisis lingüístico previo de la voz dubitada debe servir para reducir el conjunto de referencia a un grupo de muestras de voz que comparten características lingüísticas. Por lo tanto, debe ser posible hacer una consulta que permita especificar los descriptores lingüísticos presentes en la grabación dubitada, para así restringir el conjunto de referencia al hacer la comparación automatizada.

Como cada descriptor lingüístico en la base de datos estará acompañado de su información diatópica y diastrática, las grabaciones que finalmente serán mostradas al usuario serán aquellas que contengan todos los descriptores solicitados y además dichos descriptores presenten similitud en sus atributos tales como país, región, género y nivel social.

4.1 Modelo entidad-relación

En la figura 1 se muestra el modelo entidad relación propuesto para el diseño de la base de datos que cumple con las características anteriores.

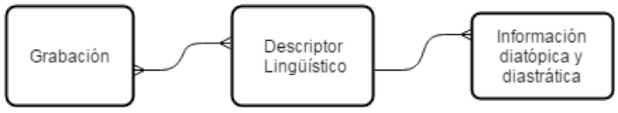

Figura 1: Diagrama Entidad-Relación.

En este modelo se tiene que una grabación puede estar asociada a uno o más descriptores lingüísticos, mientras que a su vez un descriptor linguístico puede estar asociado a varias grabaciones (relación muchos a muchos). De esta manera se puede tener que, por ejemplo, la grabación A presente "nasalización vocálica" y "articulación fricativa de /t͡ʃ/", mientras que la grabación B presente "articulación fricativa de /t͡ʃ/" y "vocalización de líquidas".

Por su parte, cada descriptor linguístico deberá tener asociada su respectiva información diatópica y diastrática. Esta información está representada por la entidad identificada en la figura como "información diatópica y diastrática", la cual consiste en una combinación de valores para atributos previamente definidos (ej. país, región, género y nivel social).

Siguiendo las buenas prácticas de un modelo relacional, es recomendable que el conjunto de países, regiones, géneros y niveles sociales sea también representado cada uno por una entidad, de manera que la combinación de valores de estos atributos sean referencias a un catálogo.

4.2 Tablas de la base de datos

A continuación se presenta la propuesta de la estructura de las tablas de la base de datos a nivel más concreto que el diagrama entidad-relación.

Primeramente, en el Cuadro 2 se muestra la información que funge como catálogo: país, región, género y nivel social. Las tablas país y región están ligadas puesto que una región debe pertenecer necesariamente a un país. Esto hace posible que en la tabla de la información diatópica y diastrática, baste con hacer referencia a la región. Todas las entradas de esta tabla son asignadas con un ID único de manera que pueden ser identificadas unívocamente en las demás tablas.

En el Cuadro 3 se muestra la tabla que representa a la entidad "descriptor lingüístico". En ella se asigna un ID para hacer referencia a cada descriptor. También es posible, de considerarse conveniente, agregar un tercer campo con una descripción del fenómeno, para que esta información pueda ser mostrada al usuario del sistema al

género	
id	*nombre*
1	H
2	M

nivel	
id	*nombre*
1	Bajo
2	Medio
3	Alto

país	
id	*nombre*
1	México
2	Panamá
3	Chile
4	Puerto Rico
5	España

región		
id	*pais_id*	*nombre*
1	1	Noroeste
2	1	Region occidental del centro
3	5	Andalucía
4	2	Capital
5	4	Área metropolitana

Cuadro 2: Tablas catálogo de la base de datos.

descriptor	
id	*nombre*
1	Articulación fricativa de /t͡ʃ/
2	Nasalización vocálica
3	Debilitamiento y pérdida de /d/ intervocálica y final de sílaba
4	Confusión de líquidas
5	Aspiración y velarización de /f/
6	Bilabialización de /f/

Cuadro 3: Tabla de los descriptores lingüsíticos.

informacion_dia				
id	*descriptor_id*	*region_id*	*genero_id*	*nivel_id*
1	1	1	2	1
2	1	1	2	2
3	1	1	2	3
4	1	2	2	1
5	1	2	2	2
6	1	2	2	3
7	1	3	n/a	1
8	1	4	n/a	2
9	1	5	2	n/a

Cuadro 4: Tabla de la información diatópica y diastrática.

momento de ser utilizado.

Finalmente, en el Cuadro 4 se muestra la tabla referente a la entidad "información diatópica y diastrática". Esta tabla contiene un registro por cada combinación de atributos que pueda presentar un descriptor lingüístico. El Cuadro 4 presenta cómo quedaría la información contenida en el Cuadro 1 almacenada dentro de la base de datos. Cada una de sus filas representa una combinación de atributos para un fenómeno. Por ejemplo, la primera fila tiene los valores $(1, 1, 2, 1)$, los cuales representan el fenómeno lingüístico, la región, el género y el nivel social. Esta fila representa la información sobre el descriptor 1 (correspondiente a la "articulación fricativa de /t͡ʃ/" según el ID que aparece en la tabla de descriptores lingüísticos), que está presente en la región 1 (correspondiente a la región Noroeste de México), en el género 2 (mujeres), y en el nivel social 1 (bajo). Aparecen ocho filas más, cada una de las cuales representa otra combinación de atributos que existe para ese descriptor lingüístico.

4.3 Búsquedas en la base de datos

Bajo este esquema se pueden hacer búsquedas de la siguiente naturaleza. Supóngase que se determina que una cierta grabación dubitada posee los descriptores lingüísticos 1 y 2 ("articulación fricativa de /t͡ʃ/" y "nasalización vocálica"). El sistema debe buscar en la base de datos, en la tabla de "información diatópica y diastrática", todos los registros o filas cuya columna descriptor_id tenga el valor 1 o 2. Como resultado se obtienen varias tuplas de atributos correspondientes a las filas de las tablas: $(1, 1, 2, 1)$, $(1, 1, 2, 2)$ y $(1, 1, 2, 3)$ serán algunas de las tuplas que se extraen de la tabla presentada en el Cuadro 4. El sistema deberá buscar que las tuplas que corresponden al descriptor 2 coincidan lo más posible con las tuplas del descriptor 1, en cuanto a región, género y nivel social. La intersección entre los dos conjuntos de tuplas describe las características que han de contener el grupo de las grabaciones del conjunto de referencia contra las cuales resultará lógico hacer la comparación automatizada, en lugar de realizarla contra todo el banco de voces.

5 Conclusiones

El proyecto que aquí se presenta pretende ser la base teórica para generar una herramienta que contribuya a mejorar los sistemas automáticos de comparación forense de voz. Actualmente dichos sistemas únicamente analizan las características

sonoras de la muestra de voz. No obstante, identificar los fenómenos lingüísticos que describen una muestra de voz desde diferentes niveles de la lengua permitiría generar grupos de grabaciones dentro del conjunto de referencia, y con ello aumentar la asertividad de las comparaciones automatizadas.

Asimismo, la información generada en este proyecto será la base teórica para desarrollar herramientas que permitan automatizar la clasificación del conjunto de referencia. Por ejempo, en el caso de los 17 fenómenos fonéticos propuestos es posible entrenar a un sistema automático de reconocimiento de voz que permita identificar aquellos que estén presentes en una nueva muestra de voz sin necesidad de que el experto realice dicha tarea.

Se ha descrito el problema que se pretende abordar y solucionar a través del proyecto "Propuesta de clasificación de un banco de voces con fines de identificación forense". Asimismo, se ha hecho un breve recorrido por algunos de los estudios más reconocidos que se han realizado en el ámbito de la dialectología y la sociolingüística.

Con el fin de comprender la manera en la que se propone la clasificación de un banco de voces a partir de descriptores lingüísticos, se han expuesto los 17 fenómenos que, en el nivel fonético, se considera pueden ser fácilmente identificables auditivamente por una persona que no sea experta en fonética y que podrían servir para entrenar a un sistema automático de voz. A pesar de que la descripción se hace únicamente a partir de estos 17 fenómenos, el proyecto considera la definición de fenómenos en todos los niveles de la lengua.

Finalmente, se describe la manera en la que la información lingüística es utilizada por una base de datos, no sólo para clasificar y generar grupos de voces con los cuales hacer una comparación automatizada; sino, para realizar descripciones lingüísticas o perfiles lingüísticos de las voces dubitadas, aún y cuando el formato de las grabaciones no lo permite.

Con este proyecto se espera contribuir en la generación e implementación de metodologías, protocolos y técnicas que doten a los laboratorios y personal de las instituciones de procuración de justicia en México de herramientas automáticas que coadyuven en la impartición de justicia.

Referencias

Aguilar, Miguel Oscar. 2011. Bases de datos criminalísticos en la procuraduría general de la república. En Sergio García & Olga Islas de González (eds.), *La situación actual del sistema penal en México. XI Jornadas sobre Justicia Penal*, 445–450.

Aleza, Milagros & José M. Enguita. 2002. *El español de américa: aproximación sincrónica.* Valencia: Tirant lo Blanch.

Boyd-Bowman, Peter. 1960. *El habla de guanajuato.* México: UNAM.

Canfield, Delos Lincoln. 1962. *La pronunciación del español de américa. ensayo histórico-descriptivo.* Bogotá: Instituto Caro y Cuervo.

Cardona, Mauricio. 2010. Fonética del español de panamá. En Miguel A. Quesada (ed.), *El español hablado en América Central. Nivel fonético*, Frankfurt: Vervuert.

Cortichs de Mora, Estrella. 1951. *El habla de tepotzotlán.* México: UNAM.

Cárdenas, Daniel. 1967. *El español de jalisco. contribución a la geografía lingüística hispanoamericana.* Madrid: Consejo Superior de Investigaciones Científicas.

Espejo, María Bernarda. 1998. Reseña a atlas lingüístico de méxico. En *THESAURUS*, vol. Tomo lIII 1, Centro Virtual Cervantes.

Jiménez, Rafael. 1999. *El andaluz.* Madrid: Arco/libros.

Lope Blanch, Juan M. 1970. Las zonas dialectales de méxico. proyecto de delimitación. *Nueva Revista de Filología Hispánica* 19(1). 1–11.

Lope Blanch, Juan M. 1990. *Atlas lingüístico de méxico.* México: El Colegio de México – UNAM – FCE.

López, Humberto. 1983. *Estratificación social del español de san juan de puerto rico.* México: UNAM.

Matluck, José. 1951. *La pronunciación en el español del valle de méxico.* México: Edición de autor.

Mendoza, José E. 2004. *Notas sobre el español del noroeste.* México: El Colegio de Sinaloa.

Moreno de Alba, José G. 1994. *La pronunciación del español en méxico.* México: El Colegio de México.

Moreno de Alba, José G. 1998. *El español en américa.* México: Fondo de Cultura Económica.

Moreno Fernández, Francisco. 1993. Las áreas dialectales del español americano: historia de un problema. *La división dialectal del español de América* 11–38.

Morrison, Geofrey Stewart, Farhan Hyder Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs & Caroline Goemans Dorny. 2016. INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263. 92–100.

Moya, Juan A. & Emilio Weiderman. 1995. La 'ch' fricativa en granada: un sonido del habla masculina. En *Actas del XII Congreso de la Asociación Internacional de Hispanistas*, vol. I, Centro Virtual Cervantes.

Navarro, Tomás. 1918. *Manual de pronunciación española*. Madrid: Consejo Superior de Investigaciones Científicas.

Navarro, Tomás. 1948. *El español en puerto rico. contribución a la geografía lingüística hispanoamericana*. Río Piedras: Editorial de la Universidad de Puerto Rico.

Perissinotto, Giorgio. 1975. *Fonología del español hablado en la ciudad de méxico. ensayo de un método sociolingüístico*. México: El Colegio de México.

Quilis, Antonio. 1999. *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Ramos, Daniel. 2007. *Forensic evaluation of the evidence using automatic speaker recognition systems*: Universidad Autónoma de Madrid. Tesis Doctoral.

Rona, José Pedro. 1964. El problema de la división del español americano en zonas dialectales. En *Presente y futuro de la lengua española I (Actas de la Asamblea de Filología del I Congreso de Instituciones Hispánicas)*, vol. I, 215–226. Madrid: Ediciones Cultura Hispánica.

Rose, Philip. 2002. *Forensic speaker identification*. Taylor & Francis.

Vaquero, María. 1996. *El español de américa i: Pronunciación*. Madrid: Arco/libros.

Vivanco, Hiram. 1999. Análisis fonético acústico de una pronunciación de 'ch' en jóvenes del estrato social medio-alto y alto de santiago de chile. *Boletín de Filología de la Universidad de Chile* XXXVII. 1257–1269.

Zamora, Munné Juan C. & Jorge M. Guitart. 1982. *Dialectología hispanoamericana. teoría, descripción, historia*. Salamanca: Colegio de España.

http://www.linguamatica.com/

linguaMATICA