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 trust (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).

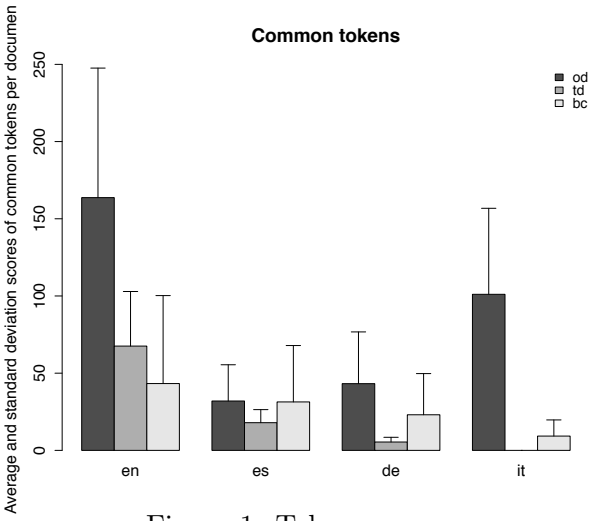

Figura 1: Tokens comuns.

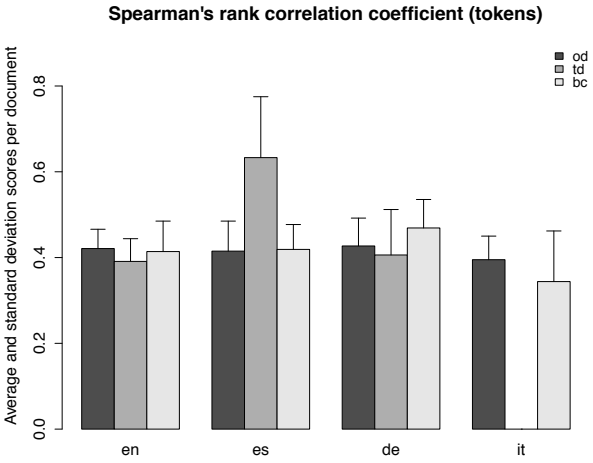

Figura 2: SCC.

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

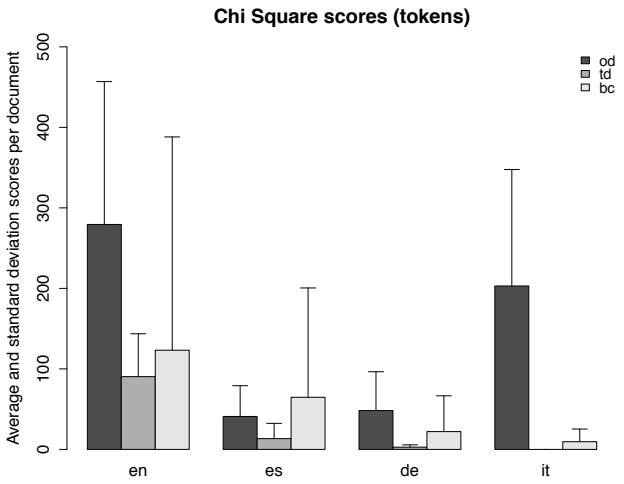

Figura 3: χ^2.

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

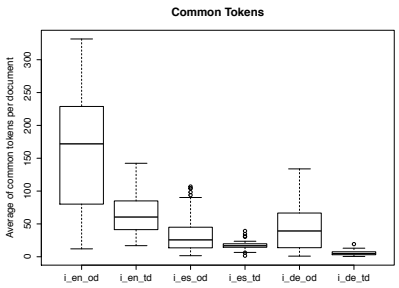

Figura 4: NTC.

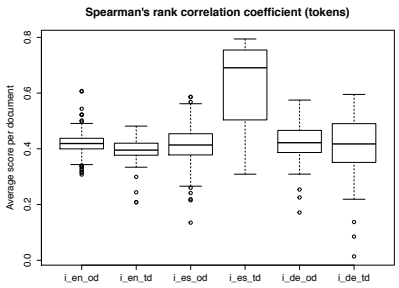

Figura 5: SCC.

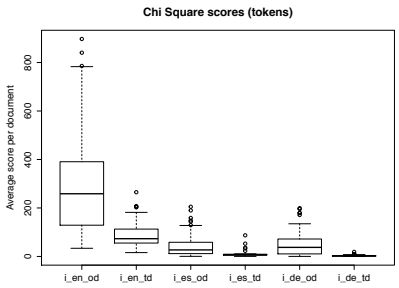

Figura 6: χ^2.

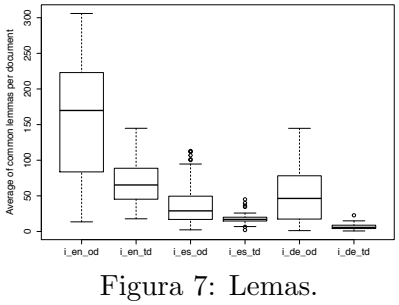

Figura 7: Lemas.

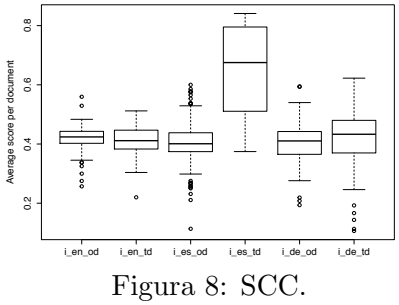

Figura 8: SCC.

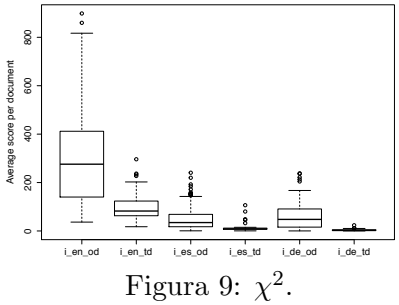

Figura 9: χ^2.

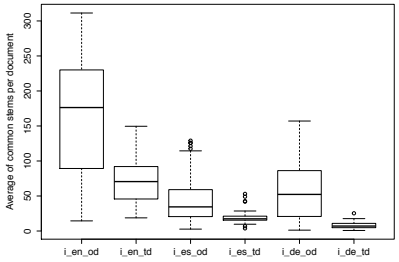

Figura 10: Stems.

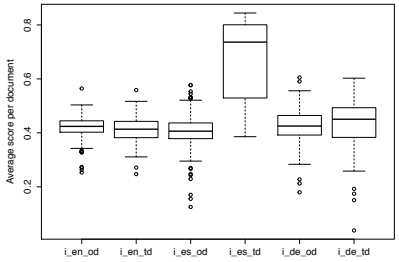

Figura 11: SCC.

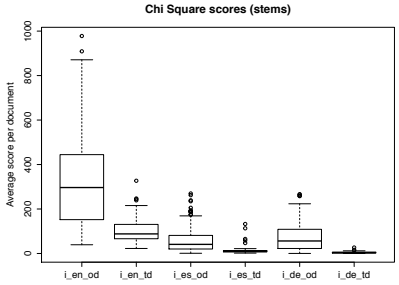

Figura 12: χ^2.

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.

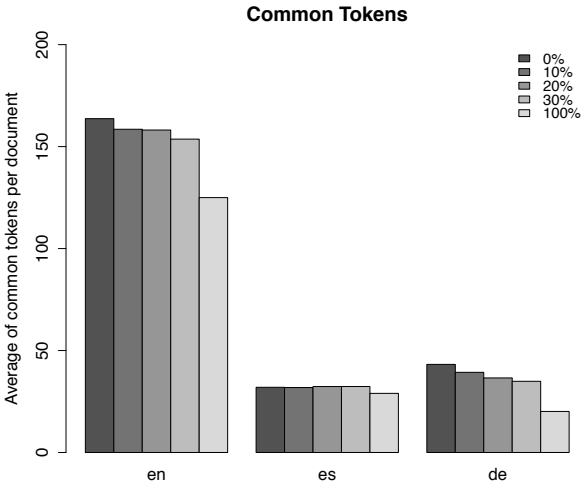

Figura 13: NTC.

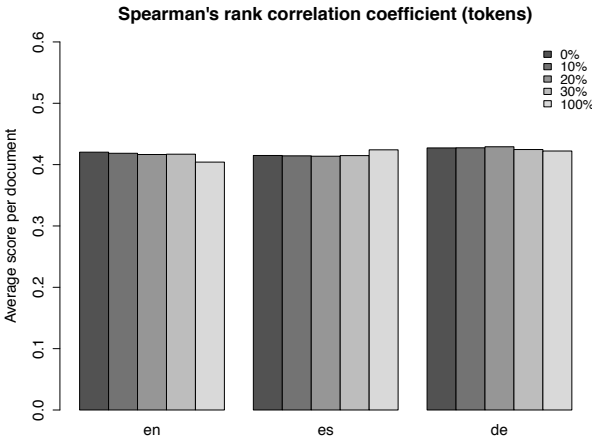

Figura 14: SCC.

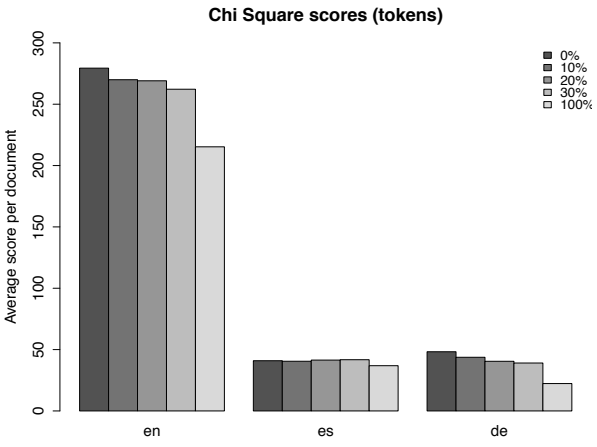

Figura 15: χ^2.

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.

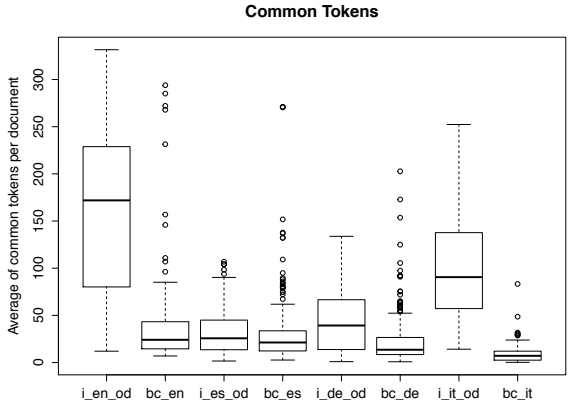

Figura 16: NTC.

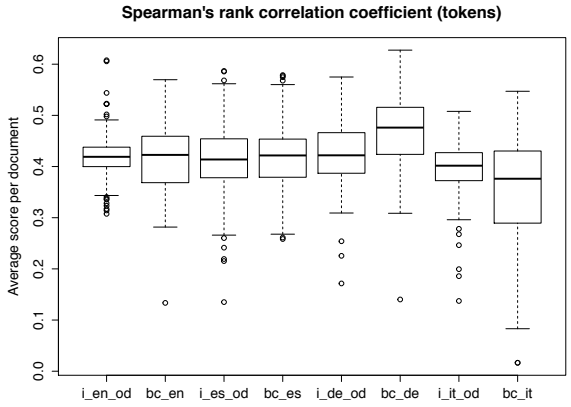

Figura 17: SCC.

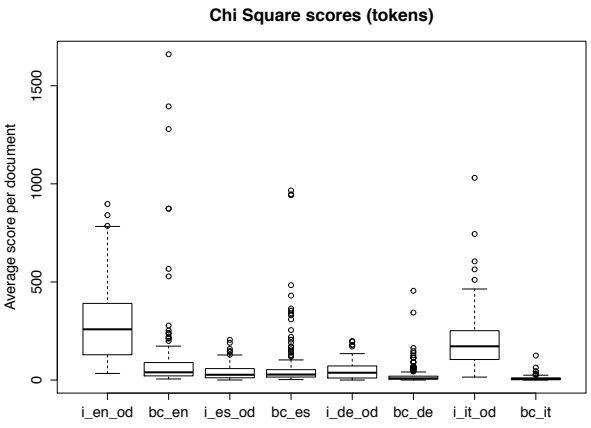

Figura 18: χ^2.

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.

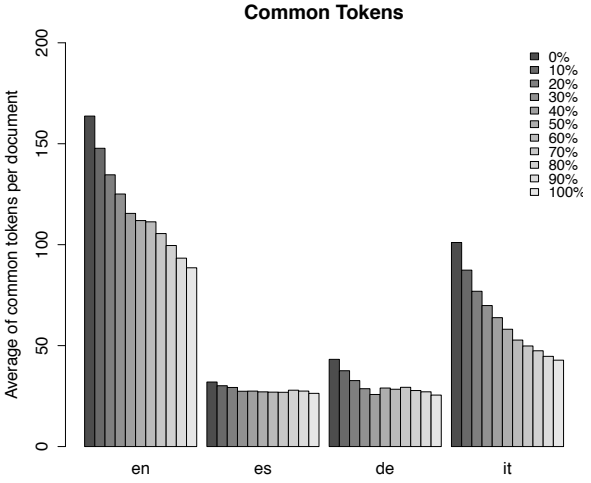

Figura 19: NTC.

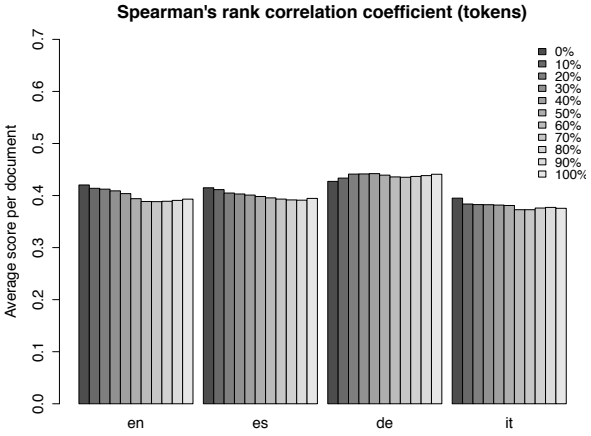

Figura 20: SCC.

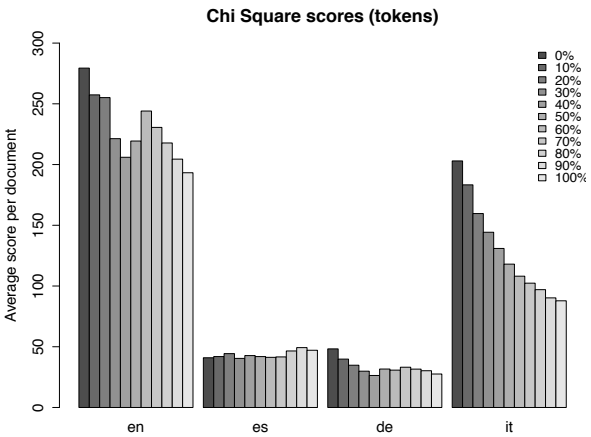

Figura 21: χ^2.

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.