UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia

Departamento de Informática

Detecção Automática de Documentos Paralelos

Por

Fátima Alexandra da Silva Bernardes

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para Obtenção do grau de Mestre em Engenharia Informática

Orientador: Doutor José Gabriel Pereira Lopes

Lisboa

2009

Agradecimentos

Um especial agradecimento a todos os meus amigos que se disponibilizaram a ajudar-me na validação manual dos documentos paralelos, aos meus pais pela paciência e à minha irmã pela força.

Resumo

A internet é uma fonte excelente de textos paralelos, sendo que dois textos são paralelos se um é tradução do outro ou ambos são traduções do mesmo texto fonte. Através da Internet, podem ser obtidos textos em diversas línguas devido ao constante crescimento do número de endereços Web multilingues. Estes textos são especialmente úteis na construção de corpora paralelos, os quais serão indispensáveis para aplicações como a Tradução Automática (baseada em exemplos, contextos ou estatística).

O objectivo nesta dissertação é a detecção automática de documentos paralelos, retirados de endereços Web multilingues, através de abordagens independentes da língua. Dos métodos estudados, foram aplicados métodos baseados nos nomes dos documentos e na proporcionalidade dos seus tamanhos, conseguindo valores de precisão entre 95% e 100%, dependendo dos corpora utilizados. De forma inovadora, utilizaram-se cognatos (palavras escritas de forma semelhante e com significado idêntico), existentes nos dois textos a comparar, para suportar a decisão sobre se os textos em análise são ou não paralelos.

Ao utilizar cognatos para estudar o seu comportamento e relevância na detecção de pares candidatos de documentos paralelos descobriu-se que, se os cognatos forem usados conjuntamente com métodos baseados em nomes de documentos e na proporcionalidade dos tamanhos dos textos, os cognatos têm um impacto evidente nos valores de precisão conseguidos pelos dois métodos anteriores. Através da identificação de cognatos consegue-se resultados de 99% para a medida f-measure em corpus com uma organização bastante rígida. Por outro lado, corpus menos organizado consegue-se obter valores de f-measure de 95,5%.

Abstract

The WEB is an excellent source for extracting parallel texts. Two texts are parallel if one is the translation of the other or both are translations of a single source document. Texts could be fetched in different languages from the internet because there are an increasing number of multilingual WEB addresses. Parallel texts are special useful to construct parallel corpora and constitute a very important resource for applications such as Machine Translation (example-, context- and statistics-based translations).

In the framework of this dissertation we aim at automatically detecting parallel documents, retrieved from multilingual Web sites, by using language independent methods. By using commonly applied filters, such as document names and lengths, we attained precision values between 95% and 100%, depending on the corpus used. Our objective of improving these values by using cognates (words that are written in close forms and have similar meanings) has shown visible impact.

In the study on the use of cognates and their relevance for the detection of candidate pairs of parallel documents it was discovered that, if cognates are used in conjunction with methods based on document names and lengths, cognates have a clear impact on the precision values: 99% for the f-measure in a corpus with a fairly rigid organization and 95.5% in less well organized corpora.

Índice de Conteúdos

Índice de Figuras

Índice de Tabelas

1. Introdução

A actual popularização da Internet e, consequentemente o seu uso generalizado, fez com que surgisse a necessidade de criar websites com traduções em várias línguas. Isto proporcionou à área de processamento da língua natural uma base de recursos inestimável de corpus paralelos, constituídos por textos paralelos ou bitextos. Os bitextos, também conhecidos por documentos paralelos, são pelo menos dois documentos em tudo iguais mas escritos em diferentes idiomas, ou seja, são traduções uns dos outros. O conjunto de vários documentos paralelos define um corpus paralelo. Estes conjuntos de textos permitem disponibilizar recursos a várias ferramentas existentes da área do Processamento da Língua Natural, como por exemplo, a aquisição automática de léxico para modelos de tradução estatística e a ligação entre vocabulários para acesso a informação multilingue ou CLIR (*cross-language information retrieval*). Por esta razão, os corpora paralelos são um bem necessário para estas ferramentas (sendo apresentada a motivação para o seu uso no capítulo 2). No entanto, a construção manual de corpus paralelos é uma tarefa praticável bastante morosa, o que torna a automatização da sua construção uma necessidade. Actualmente os documentos paralelos de um corpus podem ser extraídos através da Internet, já que esta é uma gigante fonte de recursos multilingue.

Ao proceder à extracção de documentos através da Internet, existem vários passos a executar até se conseguir um corpus paralelo. Os vários passos para a detecção de bitextos podem ser efectuados através da utilização de diferentes métodos descritos em pormenor no capítulo 0, onde é apresentado o trabalho relacionado neste âmbito. Esses métodos poderão recorrer a conhecimento sobre as línguas utilizadas nos documentos, os quais são designados por métodos dependentes da língua (capítulo 3.1) ou, ao contrário destes, pode-se recorrer a métodos independentes da língua (capítulo 3.2). Após o estudo dos trabalhos existentes, é apresentado o trabalho realizado (capítulo 0), onde a metodologia apresentada é constituída por três fases: a fase de extracção dos Web sites da Internet e a preparação dos documentos (capítulo 4.4) que serão usados na fase de detecção dos bitextos (capítulo 4.5). Para se proceder a esta detecção são usados três métodos diferentes: o método baseado em nomes dos documentos, o método baseado na proporcionalidade dos tamanhos dos ficheiros e o método baseado em cognatos. A eficiência destes métodos é calculada através da precisão recorrendo a validação manual (capítulo 4.6) o que permite discutir os resultados da aplicação destes métodos e o trabalho futuro no capítulo 0.

2. Motivação

Na área do processamento da língua natural, existem aplicações baseadas em modelos estatísticos. Estes modelos fazem uso de corpora paralelos para obterem recursos linguísticos [1], tais como sejam a aquisição de léxico, a aprendizagem computacional de modelos estatísticos de tradução ou a aquisição de vocabulário bilingue para CLIR (*cross-language information retrieval*).

Os modelos estatísticos utilizam dados de treino para aprenderem correspondências de tradução entre pares de línguas [2], não sendo necessário recorrer a dicionários bilingues e permitindo às traduções a desambiguação das palavras a traduzir [2, 3].

Devido à utilização de modelos estatísticos é crucial a existência de corpora paralelos ricos em recursos linguísticos contribuindo para o sucesso das abordagens estatísticas usadas nas aplicações multilingue [1, 4]. Devido ao cariz estatístico destas aplicações, é reconhecida a necessidade de construir corpora paralelos com quantidades consideráveis de textos, de preferência em várias línguas, com qualidade e diversidade de informação [3-5].

Outra razão apontada para o uso de modelos estatísticos é a constante evolução das línguas [6]. Por exemplo, em português a actual palavra "farmácia" num passado não muito distante, escrevia-se "pharmácia". Actualmente, em Portugal foi aprovado o novo acordo ortográfico o qual irá mudar uma parte significativa de regras na língua portuguesa. Consequentemente,

quando se começar realmente a usar extensamente estas regras, as aplicações baseadas em modelos estatísticos poderão facilmente aprender as novas regras com base em documentos escritos de acordo com o novo acordo ortográfico.

Este trabalho pretende ser uma contribuição para esta área permitindo a detecção de bitextos com o intuito de construir corpus paralelos ricos em qualidade e quantidade para fornecer léxicos variados às aplicações de tradução.

3. Trabalho Relacionado

No processo de identificação de documentos paralelos podem existir várias etapas de filtragem dos potenciais bitextos. A filtragem consiste em verificar num conjunto de documentos quais são os que podem ser paralelos entre si por aplicação de um dado método. Os métodos aplicados nestas etapas de filtragem podem ser dependentes ou independentes dos idiomas utilizados. No caso de serem dependentes do idioma a sua utilização pressupõe a existência de identificadores da língua ou dicionários, para se poder fazer a correspondência das palavras nos dois documentos. No caso dos métodos independentes dos idiomas a principal vantagem é a sua versatilidade de uso, o que permite serem utilizados para várias línguas sem qualquer tipo de conhecimento sobre as línguas de interesse.

Os métodos independentes do idioma já implementados recorrem a várias características dos documentos, tais como o nome e tamanho dos ficheiros, a estrutura HTML (caso sejam páginas Web), a existência de palavras capitalizadas, números, pontuação ou sequências de parágrafos.

Através da tabela 1 podemos ver várias combinações destes métodos implementados por vários estudos já realizados. Estes métodos são descritos mais pormenorizadamente nos próximos capítulos.

Tabela 3.1 Métodos utilizados em vários estudos.

Nome	Nome de Ficheiros	Tamanho de Ficheiros	Estrutura HTML	ALT Text	Anchor Text	Distância de Edição	Palavras Capitalizadas	Números	Pontuação	Sequência de Parágrafos	Identificação da Língua
[2, 3] PTMiner	X	X	X		X						X
[1, 5, 7] Strand	X	X	X	X	X						X
[8] PTI	X										
[9] WPDE	X	X	X	X	X	X					
[6] News Feed						X	X	X	X	X	
[10] Langlais e Patry						X	X	X	X		
[11] BITS	X		X								X
[12] Noah A. Smith		X									
[4] Noah A. Smith		X									

3.1 Métodos Dependentes do Idioma

Na classificação de possíveis documentos paralelos, os métodos dependentes da língua recorrem a dicionários e léxicos multilingues de tradução [5, 11].

O sistema BITS[11] utiliza um dicionário bilingue para descobrir pares de traduções que possibilitam o cálculo da semelhança de cada par de documentos, X e Y, segundo a fórmula:

$$semelhança\ (X,Y) = \frac{\text{Número de pares de } tokens \text{ traduzidos}}{\text{Número de } tokens \text{ existentes em X}} \quad (1)$$

Os pares de *tokens* que correspondem a traduções utilizados nesta equação são filtrados através do modelo baseado em distâncias. O modelo baseado em distâncias, segundo [11], representa graficamente a existência de pares de tradução relativamente a dois textos. Por exemplo, dados dois documentos, X e Y, o ponto (*x*,*y*) assinalado no gráfico indica que existe um par de *tokens* traduzidos nas posições *x* do documento X e *y* do documento Y. Os pontos correspondentes a traduções existentes no léxico bilingue representados graficamente irão tender a ficar próximos da diagonal principal do gráfico. Usando este método são excluídos pares de traduções, supostamente incorrectas, com base na localização nos documentos.

O sistema WPDE [9] utiliza a ferramenta *Champollion Tool Kit* para gerar pares candidatos de documentos paralelos. Esta ferramenta é um alinhador a nível da frase e é bastante dependente da quantidade de informação existente num léxico, por essa razão, o WPDE utilizou um léxico de Inglês-Chinês com 250.000 entradas. A medida de semelhança entre pares de documentos utilizada pelo WPDE é calculada com base na seguinte fórmula:

$$semelhança\ (X,Y) = \frac{\text{Número de frases alinhadas}}{\text{Total de frases existentes em X e Y}} \quad (2)$$

Patry e Langlais [10] também usam um alinhador a nível da frase, o JAPA, para a detecção de pares de documentos paralelos. Este alinhador classifica o alinhamento efectuado em cinco classes de alinhamento:

- 0-1 ou 1-0 – indica que foram acrescentadas ou removidas frases;
- 1-1 – a frase tem apenas uma correspondência entre os textos;
- 1-2 ou 2-1 – indica que a frase foi alinhada com duas frases;
- 2-2 – duas frases são alinhadas com duas frases.

Através desta classificação, definem que dois documentos são paralelos se contiverem muitos alinhamentos 1-1, enquanto que documentos não paralelos contêm vários alinhamentos 1-0 e 0-1.

De uma forma mais simples, o sistema PTIs [8] extrai todos os termos existentes em cada par de documentos contabilizando as suas frequências de ocorrência. Cada documento é representado por um vector. Cada entrada do vector corresponde a um termo traduzido no outro texto e indica a frequência de ocorrência desse termo no texto. Por exemplo, dados dois textos T1 e T2, os vectores correspondentes a cada texto seriam, respectivamente: $V_{T1} = \{w_1, w_2, \ldots w_n\}$ e $V_{T2} = \{w_1, w_2, \ldots w_n\}$, onde w_i de V_{T1} corresponde à tradução de w_i de V_{T2}.

Esta representação dos documentos em forma de vectores permite a aplicação simples e directa de medidas de semelhanças. No caso do PTI[8] é aplicado o coeficiente de Jaccard (capítulo 3.3).

3.2 Métodos Independentes do Idioma

A designação método independente do idioma aplica-se a métodos que não utilizam qualquer tipo de conhecimento da língua. Estes métodos geralmente utilizam características dos documentos que possam indicar o possível paralelismo entre dois documentos sem recorrer a qualquer tipo de conhecimento linguístico. As próximas secções deste capítulo descrevem técnicas utilizadas desta abordagem, nomeadamente através dos nomes dos ficheiros (capítulo 3.2.1) e respectivos tamanhos (capítulo 3.2.2), através da estrutura HTML de páginas Web (capítulo 3.2.3 e 3.2.4) ou ainda sobre a semelhança entre certos tipos de palavras (capítulo 3.4). Para calcular a semelhança entre palavras são definidas medidas de semelhança (também designadas por medidas de distância) no capítulo 3.3.

3.2.1 Métodos Baseados em Nomes ou URLs dos Ficheiros

Os Web designers demonstram uma tendência para dar nomes semelhantes aos ficheiros de documentos paralelos. Esta prática é extensamente utilizada para permitir manter e gerir de forma simples os Web sites. Como resultado, esta tendência natural facilita o processo de filtragem de documentos paralelos.

Os nomes dos ficheiros são normalmente constituídos por dois tipos de segmentos com significado útil no processo de filtragem. Um dos segmentos é normalmente exactamente igual aos dois nomes dos documentos paralelos e permite expressar o paralelismo entre ambos. O segundo segmento indica especificamente o idioma utilizado em cada documento. Estes segmentos que permitem caracterizar o idioma de um ficheiro são sufixos, infixos ou prefixos do nome dos ficheiros de documentos paralelos, tal como apresentado na Ilustração 3.1. Estas duas partes do nome dos ficheiros são habitualmente ligadas com hífen ‘-‘ ou com underscore ‘_’.

page_pt.html	pt_page.html	page_pt_main.html
page_en.html	en_page.html	page_en_main.html
a) Constituição de nomes de ficheiros de documentos paralelos utilizando sufixo para indicar o idioma utilizado em cada documento.	b) Constituição de nomes de ficheiros de documentos paralelos utilizando prefixo para indicar o idioma utilizado em cada documento.	c) Constituição de nomes de ficheiros de documentos paralelos utilizando infixo para indicar o idioma utilizado em cada documento.

Ilustração 3.1 - Diferentes formas de construção de nomes de ficheiros para documentos paralelos usados na Web.

Muitas vezes, os Web designers em vez de diferenciarem apenas os vários documentos paralelos através do nome do ficheiro fazem-no também através da organização de pastas. Assim, verifica-se através dos URLs a existência de pastas específicas para cada versão da língua disponível num dado Web site. Por exemplo, no Web site das Nações Unidas (http://www.un.org/) podemos encontrar um bom exemplo de aplicação desta técnica. A página *UN Millennium Development Goals*, que se encontra neste domínio, existe em seis línguas diferentes. Para cada uma dessas línguas os endereços URLs para acesso a documentos são:

Tabela 3.2 - URLs de versões em várias línguas para a mesma página

Língua	URL
Inglês	`http://www.un.org/millenniumgoals/documents.html`
Francês	`http://www.un.org/french/millenniumgoals/doc.html`
Espanhol	`http://www.un.org/spanish/millenniumgoals/documents.html`
Russo	`http://www.un.org/russian/goals/documents.html`
Árabe	`http://www.un.org/arabic/millenniumgoals/documents.html`
Chinês	`http://www.un.org/chinese/millenniumgoals/documents.htm`

No exemplo apresentado na Tabela 3.2 os nomes dos ficheiros são todos praticamente idênticos, com excepção das versões francesa e russa nas quais os nomes dos ficheiros são ligeiramente diferentes. Estes dois casos mostram que por vezes os Web designers não aplicam esta prática de forma rigorosa dando origem a algumas incoerências de nomeação dos URLs.

Outra versão das práticas usadas para a nomeação de URLs é a combinação entre os nomes dos ficheiros e a divisão por pastas dos documentos. Um exemplo desta generalização das abordagens seguidas pelos Web designers é:

Versão em Português: `http://www.site.com/folder1/pt/folder2/page_pt.html`

Versão em Inglês: `http://www.site.com/folder1/eng/folder2/page_eng.html`

Na área do Processamento da Língua Natural existem vários sistemas que tiram proveito da semelhança dos URLs para a detecção de possíveis documentos paralelos ou de sites com interesse para a construção de corpus paralelos.

O sistema Web Parallel Data Extraction (WPDE) para extracção de pares candidatos de documentos paralelos combina o uso dos nomes dos ficheiros com a medida de semelhança *edit-distance*[9] (ver capítulo 2.6) para determinar o quão parecidos eles são. O WPDE utiliza uma lista de padrões com os possíveis segmentos para cada língua. Por exemplo, para o chinês, a lista poderá ser tc, sc, tchi, schi, entre outros. Os URLs são seleccionados se contiverem um segmento existente na lista de padrões para uma dada língua. Caso contrário, o WPDE tenta encontrar o URL que seja mais semelhante segundo a métrica *edit-distance* (capítulo 3.3).

No Parallel Text Miner (PTMiner) [2] a abordagem é mais simplista. Através do nome de cada ficheiro são gerados nomes de ficheiros semelhantes através de segmentos que indicam a língua pretendida. O PTMiner tenta então encontrar os ficheiros correspondentes aos nomes gerados, se existirem são considerados como documentos paralelos. O Parallel Text Identification System (PTI) [8] utiliza o algoritmo do PTMiner mas no caso de não encontrar ficheiros com os nomes gerados, recorre a uma análise de conteúdo.

Na implementação deste trabalho este método será utilizado recorrendo apenas à medida de Levenshtein para calcular a similaridade dos URLs (capítulo 3.3) com conhecimento prévio da língua de cada documento (capítulo 4.4.3).

3.2.2 Métodos Baseados no Tamanho dos Ficheiros

Dado que documentos paralelos são traduções, o seu conteúdo deverá ser o mesmo ou ser o mais idêntico possível em línguas diferentes. Isto implica que os tamanhos dos ficheiros sejam proporcionais, embora diferentes. Esta é uma característica dos documentos paralelos usada para detectar o paralelismo entre dois textos. No entanto, se quisermos aplicar este método a um grupo de documentos sem a aplicação prévia de um outro filtro, o tamanho dos ficheiros não consegue indicar de forma razoável quais são os paralelos. Por exemplo, se tivermos os seguintes ficheiros e respectivos tamanhos: {$T_{1,L1}$ (2Kb); $T_{2,L1}$ (3Kb); $T_{3,L1}$ (3Kb)} para uma língua L1 e {$T_{1,L2}$ (3Kb)} para uma língua L2. Ao se aplicar o método baseado no tamanho dos ficheiros a este grupo de documentos, não se consegue saber se o ficheiro $T_{1,L2}$ é paralelo com $T_{2,L1}$ ou com $T_{3,L1}$. Em corpus com um grande número de documentos, este método sozinho não permite identificar facilmente pares candidatos de

documentos paralelos, pois é muito comum existirem vários documentos com tamanhos iguais ou muito semelhantes.

Apesar desta característica não poder ser usada por si só é um bom indicador em conjugação com outros métodos. Por exemplo, podemos detectar dois documentos com o método baseado em nomes dos ficheiros e, no entanto, ao comparar o tamanho dos ficheiros, se estes forem muito diferentes, isto é, se a razão entre os comprimentos na língua "A" e na língua "B" se afastarem muito da média para estes pares de línguas, podemos concluir que não são documentos paralelos.

O WPDE (Web Parallel Data Extraction) [9] apenas utiliza este método para ficheiros com mais de 40 Bytes. Ao aplicar uma primeira vez este método consegue eliminar 5% de potenciais pares paralelos. Aos restantes ficheiros calcula o quociente entre os seus tamanhos e em seguida aplica as outras técnicas para verificação do paralelismo, tais como a estrutura HTML (capítulo 3.2.3), tradução do conteúdo (capítulo 3.1) e um classificador (capítulo 3.1).

Chen e Nie ao implementarem este método mencionam o problema de estabelecer um bom intervalo de confiança para a selecção dos rácios calculados. Este intervalo tem de permitir filtrar os pares que não são paralelos sem sacrificar muitos dos que são paralelos [2]. No entanto, devido a este problema, Smith [12] sugere uma implementação baseada no modelo de regressão linear para resolver esta questão. Este método, também usado por Resnik e Smith [5], tem ainda a vantagem de conseguir diminuir o espaço de procura de forma exponencial verificando-se apenas uma perda linear de bons pares.

Este método é útil para eliminar pares identificados como paralelos mas que na realidade não o são. Por esta razão, e devido também ao facto de o método baseado nos nomes dos ficheiros não detectar por vezes que dois documentos não são paralelos quando estes têm pequenas variações nos seus nomes, o método baseado nas proporcionalidades dos ficheiros foi implementado neste trabalho com o objectivo de filtrar os resultados obtidos previamente

pelo método baseado nos nomes dos ficheiros. Para este fim, foi calculado o valor de proporção média de tamanhos dos ficheiros para os pares de línguas utilizados e foram feitos vários testes recorrendo a vários valores de tolerância (entre 10% a 40%) do valor médio de proporção entre os pares de línguas. O valor de tolerância permite aceitar pares de textos com uma proporcionalidade diferente em *x*% do valor médio da proporcionalidade para um dado par de línguas.

3.2.3 Métodos Baseados em Estrutura HTML

O método baseado na estrutura HTML para a detecção de documentos paralelos foi sugerido nos primeiros passos da implementação do sistema STRAND [1]. Segundo Resnik [1], as páginas paralelas são compostas por sequências de etiquetas HTML de forma muito idêntica, o que pode exprimir que os textos são traduções uns dos outros. De acordo com esta ideia, o algoritmo desenvolvido por Resnik compara a sequência dos *markups* em dois documentos, bem como o número de caracteres de conteúdo compreendidos entre cada *markup*, o que permite saber se os documentos são muito diferentes ou muito semelhantes.

No entanto, Chen [2] observou que documentos paralelos podem ter estruturas HTML algo diferentes. Por exemplo, isto pode acontecer caso não se utilize o mesmo editor HTML para a construção de páginas paralelas, o que poderá gerar código diferente. Esta opinião também é em parte partilhada por Ma e Liberman [11], em especial, por este filtro não considerar as páginas que são paralelas mas que possuem um aspecto diferente, e também devido ao caso inverso, ou seja, pode aceitar as que não são paralelas mas que têm um aspecto muito semelhante.

3.2.4 Métodos Baseados em Etiquetas HTML

Os elementos HTML, conhecidos por etiquetas ou *tags*, permitem estruturar um documento HTML. Através das etiquetas é possível definir ligações a outras páginas, inserir imagens, criar tabelas, entre outras inúmeras funcionalidades do HTML. O objectivo dos elementos HTML é formatar a apresentação do documento e permitir a interactividade do utilizador com o mundo Web.

As etiquetas são constituídas por atributos e pelo conteúdo a que se aplicam. De uma forma geral seguem o esquema:

```
<tagname attribute1=”xpto” … attributeN=”ypto”> content </tagname>
```

Para a extracção de documentos paralelos através da Web, Resnik [1] propôs a utilização das etiquetas `<a>` e `<img>` para descobrir potenciais pares de textos paralelos. A etiqueta `<a>`, designada por âncora, permite definir ligações através do seu atributo `href=”URL”`. Enquanto que, a etiqueta `<img>` permite mostrar imagens num documento HTML. A utilização destas etiquetas para extracção de documentos paralelos é descrita nas próximas duas secções.

3.2.4.1 Métodos Baseados em ALT Text

A etiqueta `<img>` contém o atributo `ALT` que permite definir um texto alternativo caso a imagem não seja mostrada ao utilizador. Se o *browser* não puder carregar a imagem, o texto alternativo é mostrado substituindo a imagem.

O sistema WPDE [9], através de 2000 páginas disponíveis no Web site da *Microsoft Reseach*, verificou que as imagens usadas nesse site que representam os vários idiomas existentes são quase sempre acompanhadas pelo atributo `ALT` com a discrição da língua. As palavras que a descrição `ALT` tem nestes casos são, por exemplo: english, chinese, englishversion,

chineseversion, entre outras. Assim, o sistema WPDE utiliza este atributo para detectar se um site existe em várias línguas recorrendo a uma lista predefinida de palavras que identificam as línguas de interesse. Por exemplo, dada a lista de expressões `L = {"English Version", "English", "en","Versão Portuguesa", "Português", "pt", "Español", "es"}`, e dada a seguinte parte de código HTML de uma página principal de um Web site:

```
<img src="/img/struct/languages/lang_en_white.gif" alt="[en]" />
<img src="/img/struct/languages/lang_fr_white.gif" alt="[fr]" />
<img src="/img/struct/languages/lang_it_white.gif" alt="[it]" />
<img src="/img/struct/languages/lang_lv_white.gif" alt="[lv]" />
<img src="/img/struct/languages/lang_lt_white.gif" alt="[lt]" />
<img src="/img/struct/languages/lang_hu_white.gif" alt="[hu]" />
<img src="/img/struct/languages/lang_mt_white.gif" alt="[mt]" />
<img src="/img/struct/languages/lang_nl_white.gif" alt="[nl]" />
<img src="/img/struct/languages/lang_pl_white.gif" alt="[pl]" />
<img src="/img/struct/languages/lang_pt_white.gif" alt="[pt]" />
```

Neste caso, o WPDE identificaria três imagens com expressões existentes na lista `L`. Logo, esta página tem as línguas que se pretendem e sendo o Web site seleccionado como candidato para a identificação dos seus documentos paralelos.

3.2.4.2 Métodos Baseados em Âncoras HTML

No sistema STRAND [5], Resnik propôs a utilização de um método baseado nas páginas principais de um Site, as quais permitem percorrer os URLs com mais probabilidade de conterem documentos paralelos. Muitos dos sites que utilizam páginas em várias línguas são constituídos por uma página principal, a qual é designada por página pai (*parent page*). Esta contém várias ligações para as versões do site nas línguas disponíveis. Através da procura de âncoras nos documentos HTML é possível encontrar as ligações para as diversas versões. Na literatura existente sobre o tema deste trabalho, as âncoras definem-se apenas como sendo hiperligações existentes numa página que permitem aceder a outras páginas.

Por exemplo, no Web site das Nações Unidas (http://www.un.org/) a página principal índica várias línguas através de hiperligações. Estas hiperligações permitem remeter o utilizador para páginas com a informação na língua seleccionada (Ilustração 3.2).

Ilustração 3.2 - Página Principal do Web Site da ONU contendo ligações para acesso ao site em várias línguas.

Através do código desta página, podemos verificar a definição HTML das hiperligações. O código HTML da página mostra que as hiperligações contêm a indicação da língua utilizada nas páginas para as quais remetem a ligação, por exemplo:

```
<a href="/arabic/" onMouseOver="msover('arabic')" onMouseOut="msout('arabic')">

<a href="/chinese/" onMouseOver="msover('china')" onMouseOut="msout('china')">

<a href="/french/" onMouseOver="msover('french')" onMouseOut="msout('french')">

<a href="/russian/" onMouseOver="msover('russian')" onMouseOut="msout('russian')">
```

Outra técnica muito utilizada na construção de Web sites, é utilizar em cada página uma hiperligação para as páginas equivalentes noutro idioma (Ilustração 3.3). As páginas que respeitam esta técnica são designadas por Resnik como "*sibling pages*", páginas irmãs [5].

Ilustração 3.3 - Páginas Irmãs no Site das Nações Unidas. Na base do site encontram-se ligações para as versões existentes noutras línguas da mesma página.

O sistema WPDE [9] utiliza as âncoras para obter Web sites candidatos. O método consiste em varrer todas as âncoras de um Web site. Se o Web site contiver mais do que uma âncora que indique que o Web site existe em várias línguas, o site é seleccionado como candidato. Para identificar que um Web site existe em várias línguas, o WPDE procura as âncoras que contenham expressões existentes numa lista predefinia, tal como, a lista que utiliza para o método baseado no texto da etiqueta `ALT`, descrito no capítulo 3.2.4.1. Por exemplo, dada a lista de expressões `L = {"English Version", "English", "en","Versão Portuguesa", "Português", "pt", "Español", "es"}`, e dada a seguinte parte de código HTML da página principal do Web site da Wikipedia (`http://www.wikipedia.org/`):

```
<a href="http://en.wikipedia.org/" lang="en" xml:lang="en">English</a> •
<a href="http://es.wikipedia.org/" lang="es" xml:lang="es">Español</a> •
<a href="http://eo.wikipedia.org/" lang="eo" xml:lang="eo">Esperanto</a> •
<a href="http://fr.wikipedia.org/" lang="fr" xml:lang="fr">Français</a> •
<a href="http://id.wikipedia.org/" lang="id" xml:lang="id">Bahasa Indonesia</a> •
<a href="http://it.wikipedia.org/" lang="it" xml:lang="it">Italiano</a> •
<a href="http://hu.wikipedia.org/" lang="hu" xml:lang="hu">Magyar</a> •
<a href="http://nl.wikipedia.org/" lang="nl" xml:lang="nl">Nederlands</a> •
<a href="http://ja.wikipedia.org/" lang="ja" xml:lang="ja" title="Nihongo">日本語</a> •
<a href="http://ko.wikipedia.org/" lang="ko" xml:lang="ko" title="Hangugeo">한국어</a> •
<a href="http://no.wikipedia.org/" lang="nb" xml:lang="nb">Norsk (bokmål)</a> •
<a href="http://pl.wikipedia.org/" lang="pl" xml:lang="pl">Polski</a> •
<a href="http://pt.wikipedia.org/" lang="pt" xml:lang="pt">Português</a> •
```

O método seguido pelo sistema WPDE iria decidir obter o URL `http://www.wikipedia.org/` pois detectara que as âncoras com os URLs `http://en.wikipedia.org/`, `http://es.wikipedia.org/` e `http://pt.wikipedia.org/`, continham expressões existentes na lista `L`.

3.3 Métodos Baseados em Medidas de Distância (Similaridade)

As medidas de distâncias são úteis para representar quantitativamente a similaridade entre dois objectos, podendo ser utilizadas tanto em métodos independentes do idioma como nos dependentes. No contexto apresentado, permitem medir a semelhança entre dois documentos através de um ou vários tipos de constituintes dos textos. Os constituintes que podem ser utilizados por métodos independentes da língua são por exemplo, palavras capitalizadas, números, pontuação ou sequências de parágrafos e são descritos nas próximas capítulos deste capítulo.

A utilização destas medidas pode ser baseada em várias métricas de distâncias já extensivamente estudadas e utilizadas [6, 8, 10]. Ainda neste capítulo serão descritas apenas algumas das métricas mais utilizadas, começando com a Distância Euclidiana, o Produto Interno, passando pelo Coeficiente do co-seno, pelo Coeficiente de Dice e pelo de Coeficiente de Jaccard, terminando então com a Distância de Levenshtein e a Distância de Hamming, ambas pertencentes à classe das distâncias de edição. Pode-se encontrar em [13, 14] uma lista mais vasta de métricas, bem como um estudo de comparação de precisões alcançadas na extracção de equivalentes de tradução a partir de textos paralelos utilizando 28 métricas em [13].

Para a apresentação das métricas, consideremos dois objectos (textos, por exemplo) $A = (a_1, a_2, \ldots, a_n)$ e $B = (b_1, b_2, \ldots b_n)$ de um espaço *n*-dimensional. Estes dois objectos são vectores

em que as suas componentes contabilizam a frequência, da presença ou da ausência, ou ainda a frequência pesada de certas características dos textos (descritas no início deste capítulo, como é o caso de palavras capitalizadas, números, pontuação, …) e são utilizados para o cálculo de medidas de distância entre dois textos. Além dos constituintes já referidos, os textos podem ser representados por palavras, por sequências de palavras ou por caracteres (n-gramas de palavras ou de caracteres) que os constituem.

Dada uma sequência qualquer, um *n-grama* é uma subsequência constituída por *n* elementos dessa sequência. Esta subsequência pode ser constituída por *n* caracteres, *n* palavras ou por quaisquer outros *n* elementos da sequência base inicial, dependendo da aplicação em questão. Para o presente caso, se considerarmos a expressão "Detecção Automática de Documentos Paralelos", os bigramas ou 2-gramas de palavras, possíveis são:

```
{"Detecção Automática"; "Automática de"; "de Documentos";

"Documentos Paralelos"}.
```

Se considerarmos caracteres, os bigramas seriam:

```
{"De"; "et"; "te";"ec"; … ; "le"; "el"; "lo"; "os"}.
```

A medida de semelhança mais comum é a distância euclidiana definida pela equação (3). Embora seja bastante intuitiva e útil devido à sua simplicidade, pode classificar dois objectos como muito similares quando, na verdade, esses dois objectos não partilham nenhuma característica [8]. Este problema pode acontecer quando se tem a mesma frequência (baixa) de *tokens* em dois documentos, mas estes não têm qualquer correspondência. Desta forma, a distância euclidiana, que se baseia apenas na diferença de frequências sem ter em conta os *tokens* em comum, dará um resultado de semelhança não fidedigno em certas situações.

$$d(A,B) = \|A - B\| = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (3)$$

O produto interno, definido pela equação (4), permite a soma dos produtos dos pesos de cada componente utilizada para a medição.

Segundo Chen et. al [8], esta métrica não calcula valores que pertençam a um intervalo no qual se possa comparar e compreender o seu resultado, a não ser que se normalizem os vectores de forma a obter valores entre 0 e 1, mas neste caso a métrica do produto interno transforma-se na métrica do coeficiente do co-seno (5).

$$\sum_{i=1}^{n} a_i \cdot b_i \quad (4)$$

O coeficiente do co-seno, definido pela equação (5), calcula a similaridade através do ângulo co-seno entre os vectores normalizados dos objectos utilizados. Os documentos são representados através dos vectores e a similaridade entre ambos é dada pelo ângulo co-seno formado entre si. Quando se pretende comparar objectos no domínio textual esta é umas das métricas mais utilizadas.

$$\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} {a_i}^2 * \sum_{i=1}^{n} {b_i}^2}} \quad (5)$$

O coeficiente de Dice utiliza os constituintes escolhidos de dois documentos para calcular o grau de similaridade de dois objectos através da fórmula (6).

$$\frac{2\,|A\ \cap B|}{|A|\ +\ |B|} \qquad (6)$$

O numerador de (6) representa o dobro do número de constituintes comuns aos dois textos, enquanto que o denominador denota a soma das frequências desses constituintes em cada um dos textos. Por exemplo, dadas as palavras "*Documentos*" e "*Documents*", em português e em inglês, respectivamente, se as quisermos comparar, medindo o grau de semelhança entre elas, com base nos bigramas de caracteres e no Coeficiente de Dice, os bigramas possíveis seriam:

{"Do";"oc";"cu";"um";"me";"en";"nt";"**to**";"**os**"}

{"Do";"oc";"cu";"um";"me";"en";"nt";"**ts**"}

Assim, porque a palavra "*Documento*" tem 9 bigramas, enquanto que "*Documents*" tem 8 e os bigramas comuns a ambos são 7, com esta informação o valor do coeficiente de Dice é:

$$\frac{2\ *\ 7}{9+8} = 0{,}82$$

Ou seja, através desta métrica "*Documentos*" e "*Documents*" têm uma semelhança de 82%.

O coeficiente de Jaccard é muito semelhante ao anterior. No entanto, utiliza a cardinalidade do grupo de características que intercepta os objectos a dividir pela cardinalidade da união dessas mesmas características (7).

$$\frac{|A \cap B|}{|A \cup B|} \quad (7)$$

Utilizando o exemplo anterior, e também os bigramas de caracteres para as mesmas palavras "*Documentos*" e "*Documents*", o coeficiente de Jaccard indicaria uma semelhança de 70%, dado que:

$$\frac{7}{9+8-7} = 0{,}7$$

Existe ainda uma classe de distâncias designada por distâncias de edição. Este tipo de distâncias não mede exactamente a semelhança entre objectos mas sim a dissemelhança. O facto que torna estas medidas em funções de dissemelhança é porque na realidade o que elas contabilizam são as diferenças entre as sequências a comparar, baseando-se nas operações de inserção, remoção ou substituição necessárias para transformar uma sequência noutra. Por exemplo, dadas as palavras "acquired" e "adquire", teremos duas edições:

1. "`acquired`" → "`adquired`" (substituição de 'c' por 'd').

2. "`adquired`" → "`adquire`" (remoção do 'd' final).

Para o exemplo utilizado nas medidas de semelhança anteriores teríamos apenas uma edição:

1. "`documents`" → "`documentos`" (inserção 'o')

ou, por outro lado, teríamos:

2. "`documentos`" → "`documents`" (remoção 'o').

As duas distâncias de edição mais utilizadas são a de Levenshtein e a de Hamming. A distância de Levenshtein considera as três operações de edição possíveis. Por esta razão, é

vista como generalização da distância de Hamming que apenas contabiliza substituições e que, por isso, apenas funciona com sequências do mesmo comprimento.

Apesar destas medidas serem de dissemelhança, podem ser facilmente convertidas para medidas de semelhança. Por exemplo, dadas duas palavras *A* e *B*, em que *A* tem *#A* caracteres e *B* tem *#B* caracteres, a medida de semelhança calculada através da métrica de Levenshtein, *Levenshtein(A,B)*, será:

$$1 - \frac{Levenshtein(A, B)}{máx(\#A, \#B)} \tag{8}$$

Deste modo, para os exemplos anteriores temos:

$$1 - \frac{Levenshtein(\ acquired, adquire\ )}{máx(8, 9)} = 1 - \frac{3}{9} = 0.78$$

$$1 - \frac{Levenshtein(\ documents, documentos\ )}{máx(9, 10)} = 1 - \frac{1}{10} = 0.90$$

3.4 Métodos Baseados em Palavras Cognatas

Palavras cognatas são palavras que têm uma semelhança ortográfica elevada e têm o mesmo significado. As palavras cognatas são palavras com a mesma origem histórica e devido à evolução das sociedades podem ter modificado a sua forma ortográfica ligeiramente, mantendo no entanto o seu significado (nível semântico). Por esta razão, as palavras cognatas podem ser facilmente encontradas entre pares de línguas distintas desde que partilhem origens e o mesmo alfabeto. Para línguas que não utilizem o mesmo alfabeto pode-se proceder à transliteração dos alfabetos para o reconhecimento de cognatos. A Tabela 3.3 apresenta alguns exemplos de palavras cognatas para português e para inglês:

Tabela 3.3 - Exemplos de palavras cognatas entre português e inglês.

Português	Inglês
Portugal	Portugal
Lisboa	Lisbon
Documento	Document
Social	Social
Parlamento	Parliament
Europa	Europe
Representação	Representation

Devido à evolução das palavras a nível semântico, existem palavras semelhantes na forma mas que têm significados completamente diferentes e que, por esse motivo, não são cognatas. Aliás, são designadas por falsos cognatos ou falsos amigos (*false friends*). Por exemplo, "*Livraria*" em português é um falso cognato de "*Library*" em inglês, pois as duas palavras têm significados diferentes apesar da semelhança de escrita. "*Livraria*" em português seria traduzido por "*Bookstore*" em inglês, enquanto que "*Library*" em inglês seria traduzido por "*Biblioteca*" em português.

Ao comparar dois textos podem ser identificados vários cognatos. Nas Ilustração 3.4 e Ilustração 3.5, estão identificados vários cognatos entre os dois textos apresentados (não de forma exaustiva, mas apenas a titulo ilustrativo). Através destes dois textos, com a identificação de alguns cognatos entre eles, percebe-se que os cognatos permitem identificar características de possíveis textos paralelos.

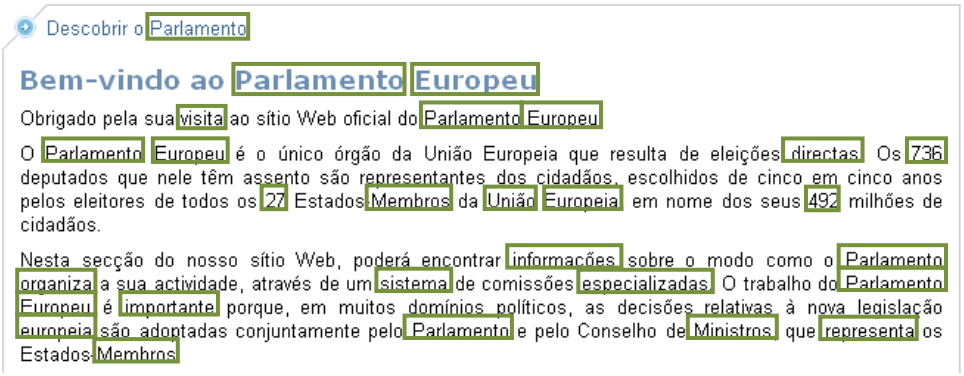

Ilustração 3.4 - Texto em português para a identificação de cognatos com o texto em inglês da ilustração 3.5.

Ilustração 3.5 - Texto em português para a identificação de cognatos com o texto em inglês da ilustração 3.4.

Devido aos cognatos permitirem obter características comuns entre dois textos são usados para várias tarefas no processamento de língua natural. Simard et. al [15], sugeriram aplicar os cognatos no processo de alinhamento de documentos usando o algoritmo de Gale e Church [16] como base. Seguiram-se vários estudos e outras técnicas baseadas em cognatos, mas sempre referentes ao processo de alinhamento dos documentos.

Este trabalho baseia-se em cognatos para um dos métodos de filtragem de documentos paralelos de forma independente da língua (capítulo 4.5.3). Os cognatos são extensivamente utilizados para o alinhamento, existindo vários estudos que comprovam a sua utilidade em tarefas multilingues [17]. No entanto, no processo de detecção de documentos paralelos ainda não foram utilizados de forma a mostrar o seu potencial utilizando o seu conceito lato. Isto porque na verdade já foram utilizados cognatos como filtros de documentos paralelos, mas não no sentido generalizado do seu conceito, mas sim utilizando apenas o conceito de homógrafos como descrito no capítulo 3.4.1.1.

Palavras cognatas podem ser descobertas através de métodos baseados em *n-gramas*. Michel Simard et al. [15] utilizaram os quatro primeiros caracteres de palavras para a detecção de cognatos, apesar de se ter revelado um bom método tem o problema de ser sensível ao início das palavras e não detectar as variações no final [14]. Por exemplo, no caso das palavras "*conservative"* e "*conseil"*, em inglês e em francês respectivamente, seriam consideradas como cognatas e não o são [17]. Depois de Michel Simard et al. [15], foi sugerido um método semelhante mas que se baseia em *4-gramas* contíguos numa sequência, sem ser necessariamente os quatro primeiros caracteres [18]. Através deste segundo método é construído um gráfico para cada par de documentos, em que cada posição (*x*,*y*) representa um ponto caso os documentos partilhem nessas posições um *4-gramas* comum. Ou seja, na posição *x* do texto T_1 existe uma sequência de quatro caracteres iguais à existente na posição *y* do texto T_2. Como este *4-gramas* de caracteres não têm necessariamente frequência 1, a representação gráfica (Dot-plot) dá origem a múltiplos pontos representados, sendo visível uma linha diagonal representada que é depois tratada para serem determinados pontos de alinhamento, no processo de alinhamento de textos paralelos.

Para fazer o alinhamento de textos paralelos, António Ribeiro [14] utiliza possíveis cognatos e tokens homógrafos com idênticas frequências para fazer uma regressão linear sobre os

pontos possíveis de alinhamento. Através do método do histograma das distâncias entre a posição actual desses pontos e a posição esperada, determinada pela regressão linear, procede-se a uma primeira filtragem de maus alinhadores. Uma segunda filtragem de maus alinhadores é feita utilizando o método das bandas de confiança.

3.4.1.1 Métodos Baseados na Semelhança de Documentos Medida Através da Ocorrência de *Tokens* Homógrafos

Numa perspectiva conservadora, diz-se que duas palavras são homógrafas quando se escrevem da mesma forma mas têm significados diferentes, pronunciando-se na maioria das vezes de forma diferente. Por exemplo: colher (substantivo) e colher (verbo), jogo (substantivo) e jogo (verbo) ou Sede (de lugar) e sede (avidez). Contudo, não existe palavra alguma que não tenha mais de um significado, apesar de se pronunciar e escrever da mesma forma. "Portugal" tem um significado de lugar em frases como "nasceu em Portugal", "vive em Portugal", "vai para Portugal", etc. A mesma palavra tem um significado de agente em frases como "Portugal decretou...", "Portugal decidiu...", "Portugal exigiu...", "Portugal venceu a Turquia", etc.

No âmbito deste trabalho, e num contexto multilingue, os homógrafos podem também ser vistos, num sentido mais lato, ou como palavras ou como números que não são traduzidos entre línguas. São exemplo disto, nomes próprios como "Loyd's Bank", acrónimos como "IBM", datas como "2008", abreviaturas como "Ltd", entre outros.

De forma a tirar partido dos homógrafos, estudos já realizados utilizaram-nos como características dos textos que pretendem comparar palavras capitalizadas, números e sinais de pontuação para a detecção de textos paralelos. Estes métodos são apresentados nos capítulos seguintes.

3.4.1.2 Métodos Baseados em Palavras Capitalizadas

As palavras capitalizadas referem-se a palavras com a primeira letra em maiúscula e que não iniciem uma frase, ou seja, geralmente são nomes próprios usados no meio de frases.

Este método pode ter alguns problemas dependendo da língua em que se trabalhe e de muitas outras circunstâncias. Nem sempre um nome próprio permanece com a mesma grafia em duas línguas diferentes. É o caso da palavra *Lisboa*, que é traduzida para inglês por *Lisbon*, ou *Portugal* que é traduzido para Checo por *Portugalsko*, *Portugalska*, etc., dependendo do caso em que esteja a ser utilizado. Outros exemplos deste tipo de problema é a escrita dos adjectivos que expressam nacionalidades ser feita em maiúsculas em inglês e, em alemão, todos os nomes são capitalizados. Por este motivo, muitas vezes, a capitalização por si só não é suficiente para identificar nomes próprios.

Patry e Langlais [10] seguiram a abordagem utilizada por Nadeau e Foster [6] para a identificação de palavras capitalizadas. A aproximação seguida por ambos utiliza o coeficiente do co-seno (5) como medida de semelhança entre os textos. Para concretizar esta abordagem, são extraídos de cada documento vectores que representam um dado recurso do texto. Neste caso em concreto, os vectores são construídos através das palavras capitalizadas e são utilizados no cálculo do coeficiente do co-seno (5) para a avaliação da similaridade entre os documentos.

3.4.1.3 Métodos Baseados em Números

Os números existentes num documento permitem referir datas, quantidades, medidas, enumerações, entre outras coisas. Este tipo de referência é algo que geralmente não muda com as traduções o que torna também os números um bom elemento de comparação entre textos paralelos.

Nadeau e Foster [6] aplicaram este método para detecção de documentos paralelos num corpus constituído por notícias, utilizando a medida de similaridade do co-seno (5). Verificaram que este método consegue obter uma precisão de 100% e um *recall* de 85%. O que significa que nunca são detectados falsos positivos. No entanto, existem 15% de falsos negativos.

3.4.1.4 Métodos Baseados em Pontuação

Estes métodos apenas têm em conta a pontuação que seja aplicada da mesma forma nas línguas em questão. Por exemplo, as vírgulas são usadas de forma diferente no Inglês em relação ao Português e noutras línguas. Por esta razão não devem ser consideradas para efeitos de detecção de documentos paralelos de forma independente da língua.

No entanto, ao se considerar apenas sinais de pontuação como aspas, parênteses curvos e rectos, e sequências de parágrafos, em princípio, como estes não dependem tanto da construção gramatical da língua e a sua localização nos dois textos é praticamente a mesma, pode-se usar este método segundo uma abordagem independente da língua.

Este método pode ser usado recorrendo a uma das medidas de similaridade. Nadeau e Foster [6] através da medida do co-seno (5) aplicada a este método verificaram um *recall* de 100%, no entanto a precisão obteve um valor baixo, cerca de 30%. O que indica que este método

selecciona demasiados falsos positivos, ou seja, são classificados como textos paralelos demasiados pares, mais do que os que realmente existem. Isto também nos indica que existem muitos textos a partilharem frequências de pontuação muito semelhantes. Segundo estes dados [6], métodos baseados em pontuação não permitem reflectir características suficientemente fortes para a identificação de documentos paralelos.

Os métodos baseados em homógrafos podem ser conjugados de forma tirar mais partido da semelhança entre possíveis documentos paralelos. Nadeau e Foster [6] aplicaram a combinação do método baseado em palavras capitalizadas com o método baseado em números verificando desta forma um aumento na precisão e no recall.

4. Trabalho Realizado

4.1 Introdução

O objectivo deste trabalho é descobrir os textos paralelos existentes na Web em determinados endereços. Para este fim foram implementados três métodos de filtragem para documentos que poderão eventualmente ser paralelos: o método baseado em nomes de ficheiros (capítulo 3.2.1), o método baseado na proporcionalidade dos tamanhos dos ficheiros (capítulo 3.2.2) e o método baseado em cognatos (capítulo 3.4).

Convém relembrar que os métodos baseados em cognatos são extensamente usados em processos como o alinhamento de textos paralelos mas não na fase de detecção do paralelismo de dois documentos. O objectivo da utilização destes métodos neste trabalho é conseguir aumentar a precisão do processo de descoberta de bitextos.

4.2 Definição dos Corpus Usados

Actualmente a Internet contém vários corpora multilingues disponíveis através de vários Web sites. Assim, para se obter corpora relevantes para a experimentação da implementação, foram escolhidos três Web sites principais tendo em conta dois critérios. O primeiro critério de escolha foi a necessidade dos corpora conterem línguas conhecidas (Português, Espanhol, Francês e Inglês) de forma a se poder validar os pares candidatos a documentos paralelos encontrados. O segundo critério assentou na estrutura de ficheiros dos websites para a utilização dos métodos baseados no nome dos ficheiros. Para se poder utilizar este método é necessário que exista alguma estruturação no Web site de forma a que os bitextos possam ser identificados. Tendo estes critérios em conta, a escolha recaiu sobre os Web sites do Eur-Lex[1], do Vaticano[2] e das Nações Unidas[3].

O Eur-Lex é constituído por 23 línguas e foram extraídos deste site 14268 documentos representando cerca de 368MegaByte. Este site é extremamente bem comportado, dir-se-ia que poucos documentos apresentam dificuldades na detecção de que são, ou não, paralelos.

O Web site do Vaticano contém apenas 8 idiomas e não é tão paralelo quanto o Web site do Eur-Lex, isto significa que existe um número significativo de páginas que não têm tradução correspondente noutros idiomas. Este caso acontece com maior ênfase nas páginas em Italiano devido a esta ser a língua oficial do site e por isso existem mais páginas escritas nesta língua. O corpus retirado do Web site do Vaticano é constituído por um total de 55595 documentos representando 877MegaByte.

O terceiro corpus foi retirado do Web site das Nações Unidas sendo constituído por 6 línguas e foi extraído deste site um total de 877MegaByte contendo 43105 documentos.

[1] http://eur-lex.europa.eu/
[2] http://www.vatican.va/
[3] http://www.un.org/

Na Tabela 4.1 são apresentados o número de documentos de cada uma das línguas utilizadas nos Corpus descritos anteriormente. Os números apresentados nesta tabela foram obtidos através do identificador de língua implementado e discutido no capítulo 4.4.3. Os número apresentados na Tabela 4.1 são inferiores aos realmente existentes e disponibilizados pelos Web sites. Isto deve-se ao facto de se ter utilizado um browser para a conversão dos ficheiros de HTML para TXT utilizando a técnica copiar-colar (capítulo 4.4.2). Ao se utilizar um browser para esta conversão, fez com que muitas páginas não tenham sido carregadas pelo browser fazendo com que este retornasse erros informativos desta falha pelo que as páginas nestas condições foram descartadas.

Tabela 4.1 – Número de documentos em 5 línguas existentes em cada um do corpus utilizado no trabalho. O símbolo – significa que o idioma não se encontra disponível para o corpus em causa.

Corpus	Português	Espanhol	Francês	Inglês
Euro-Lex	283	294	338	273
Vaticano	4222	5352	4036	4867
Nações Unidas	--	623	1849	244

4.3 Metodologia

O processo usado para a identificação dos documentos paralelos de um corpus é constituído pelos seguintes passos:

1. Transferência para disco local do corpus pretendido através da utilização de um crawler (capítulo 4.4.1).
2. Conversão do HTML para texto puro[4] dos documentos transferidos. O texto puro é caracterizado inexistência de etiquetas HTML, sendo apenas constituído pelo conteúdo da informação existente no documento HTML sem qualquer formatação (capítulo 4.4.2).
3. Identificação da língua dos documentos (capítulo 4.4.3).
4. Aplicação do método baseado em nomes ao Corpus (capítulo 4.5.1).
5. Aplicação do método baseado na proporcionalidade dos tamanhos (capítulo 4.5.2).
6. Aplicação do método baseado em cognatos (capítulo 4.5.3).

Os passos 1, 2 e 3 constituem a fase de preparação do corpus para posteriormente aplicar a fase de detecção automática de bitextos constituída pelos restantes passos (Ilustração 4.1). O funcionamento de cada uma destas fases é descrito nas próximas secções deste capítulo.

[4] Neste contexto, foi decidido usar o termo "texto puro" como tradução do termo "*plain text*" em Inglês.

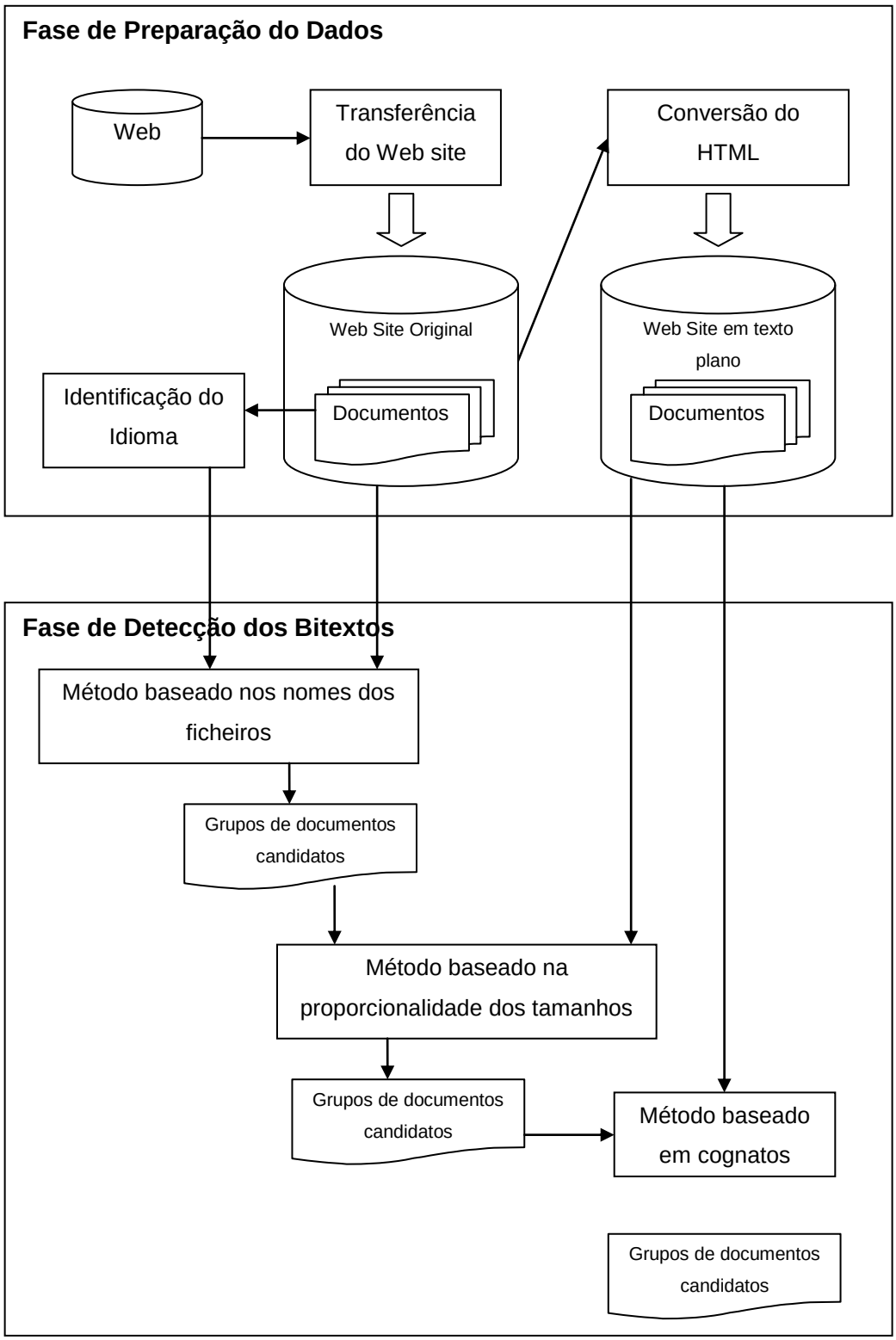

Ilustração 4.1 - Diagrama de componentes da detecção automática dos documentos paralelos.

4.4 Fase de Preparação

A fase de preparação têm como objectivo preparar o corpus para posteriormente ser utilizado na fase de detecção de candidatos a documentos paralelos. A fase de preparação é constituída por três passos. Inicialmente o corpus é obtido através de um Web site (capítulo 4.4.1) e os documentos HTML extraídos são convertidos para formato .txt (capítulo 4.4.2). Tendo feito a transferência do corpus é identificada a língua de cada documento (capítulo 4.4.3).

4.4.1 Transferência do Corpus

Para proceder à transferência de um Corpus foi utilizado o programa GNU Wget[5]. Esta ferramenta permite fazer transferências de ficheiros através dos vários protocolos Internet mais usados, incluindo HTTP, HTTPS e FTP.

O GNU Wget opera através da linha de comando permitindo ser invocado através de outros programas.

Uma das funcionalidades de interesse do GNU Wget é a recursividade das transferências, permitindo desta forma, percorrer um Web Site seguindo as ligações existentes e a estrutura de directorias.

Através desta funcionalidade de recursividade do GNU Wget foram transferidos apenas documentos HTML existentes nos Web sites de interesse para este trabalho (capítulo 4.2). A cópia mantida localmente obtida através da transferência destes Web Sites mantém a estrutura de directorias original, o que permite usar localmente os caminhos relativos dos ficheiros como se fossem URLs originais. Esta característica será útil para o método baseado nos nomes dos ficheiros (capítulo 3.2.1).

[5] Este programa pode ser encontrado em http://www.gnu.org/software/wget/.

4.4.2 Conversão dos Documentos HTML

Depois da transferência do corpus, tal como indicado no capítulo anterior, os documentos existentes são convertidos do formato HTML para ficheiros .txt, sendo removido qualquer tipo de informação inerente à linguagem HTML. Apenas é guardado o conteúdo apresentado por cada documento.

Para proceder a esta tarefa foi implementado um script que abre as páginas Web num browser e através dos comandos copiar e colar (*copy-paste)* obtém o conteúdo de cada documento apresentado num browser guardando-o no respectivo ficheiro .txt. Esta técnica permite obter informação que é gerada por scripts do lado do cliente (por exemplo, Javascript ou Flash) que de outra forma se perderia.

Tendo os documentos neste formato já sem a estrutura HTML é executada a normalização dos mesmos. Esta normalização consiste colocar todas as letras em minúsculas e em colocar um espaço entre palavras, ou entre números, e a pontuação. Por exemplo, dada a seguinte frase onde os espaços estão assinalados com traço:

"Gosto_de_dizer._Direi_melhor:_gosto_de_palavrar_(Fernando_Pessoa_-_1082)".

Depois de aplicada a normalização ficaria:

"gosto_de_dizer_._direi_melhor_:_gosto_de_palavrar_(_fernando_pessoa_-_1982_)".

Este processo é necessário para os métodos que consideram o conteúdo dos documentos. Caso contrário, a pontuação iria ser contabilizada o que pode dar origem a resultados não previstos, uma vez que as regras de pontuação variam de língua para língua. Este caso é aplicável neste trabalho pelo método baseado em cognatos (capítulo4.5.3).

4.4.3 Identificação do Idioma dos Documentos

Para a detecção de documentos paralelos a identificação dos idiomas de cada texto existente num corpus é necessária de forma a poderem ser utilizados apenas os textos nos idiomas de interesse. Muitos autores utilizam a identificação do idioma como um processo de filtragem para descartar documentos.

A abordagem seguida para executar esta tarefa inclui verificação do código HTML e, se necessário, através do URL do documento. Para isso é necessária a existência de uma lista de segmentos associados a cada língua, por exemplo:

```
ingles = {"/english/", "/en/", "_english", "_en", "_english_", "_en_",
         "english_", "en_", "-en", "_uk", "_uk_", "uk_"}

portugues = {"/pt/", "_pt", "_pt_", "pt_", "/po/", "_po", "po_", "_po_", "-
         po", "-pt"};
```

Estas listas contêm padrões usados habitualmente na definição das línguas através do HTML ou nos URLs, tal como foi discutido no capítulo 3.2.1., e permitem a identificação dos idiomas de cada texto de forma simples e com resultados satisfatórios. Uma vantagem da utilização destas listas é a possibilidade de acrescentar mais idiomas à aplicação de forma simples e prática.

A abordagem implementada começa por analisar as marcas `<html lang=”XX”>` e `<meta name=”dc.language” content=”XX”>`. Se uma destas marcas HTML contiver um dos elementos definidos na lista de segmentos então o idioma do documento está encontrado. Caso não seja possível identificar a língua do documento recorrendo às marcas HTML a procura passa a ser efectuada exclusivamente através do URL do documento.

Tal como foi referido no capítulo 4.4.1 sobre a Transferência do Corpus, o sistema de directorias original é mantido em disco local preservando desta forma o URL dos documentos. Assim, o caminho relativo do documento é análogo ao URL e mantém as mesmas propriedades que são aproveitadas para a detecção da língua dos documentos em conjunto com as listas de segmentos. Para cada documento, se o caminho relativo contiver um dos segmentos da lista de idiomas então a língua é identificável.

Depois de se ter aplicado estes dois métodos, caso não seja possível atribuir uma língua ao documento então será atribuída a identificação "UNLISTED[6]" ao documento.

Apesar de esta abordagem ser simples apresenta alguns problemas inerentes ao modo com que os Web sites são construídos.

Um dos problemas mais comuns ocorre devido aos programas que são usados para construir Web sites, os IDEs (Ambientes de Desenvolvimento Integrado). A maioria destes programas assume como língua padrão o Inglês e coloca-o como idioma na informação HTML. Outro problema é relativo à própria organização do Web site que por vezes não existe e torna a identificação difícil através dos URLs.

O primeiro problema pode ser contornado ignorando a identificação do documento como inglês, ou seja, assume-se que a língua ainda não foi identificada e recorre-se ao URL para despiste da língua. Assim, se através da análise do URL for detectada outra língua é esta que será atribuída ao documento, caso contrário mantém o inglês como o idioma atribuído ao ficheiro. Esta abordagem demonstrou ser razoável e na grande maioria dos casos é a mais correcta, embora tivessem sido identificados alguns exemplos em que foi atribuída uma língua diferente do inglês e efectivamente o inglês seria a língua correctamente atribuída.

[6] No contexto deste trabalho, "UNLISTED" significa língua não listada ou não existente no programa.

Também se verificou que, por vezes, na construção dos Web sites, o idioma do documento é identificado através de meta-informação (ex.: `<meta name="dc.language" content="pt">`) e sem que seja actualizada a informação predefinida pelo programa utilizado (`<html lang="en">`), ou vice-versa. Nestas situações, considerou-se que a língua do documento seria a definida como diferente de inglês. Esta escolha foi devida ao facto de os programas assumirem mais uma vez o inglês como língua padrão e se existir outra língua definida será porque alguém a acrescentou ao código HTML, logo terá mais peso a consideração humana na escolha do idioma.

Mais uma vez esta abordagem apresentou-se eficiente na identificação da língua, excepto no site do Vaticano em que foram encontrados documentos identificados como italianos, quando na verdade, estavam escritos noutra língua. A razão para isto ter acontecido foi provavelmente a reutilização das páginas em italiano modificando apenas o conteúdo dos documentos para o texto na outra língua. Este tipo de situação não poderá ser identificado através desta abordagem precisando recorrer a outras técnicas não implementadas, por exemplo, através de classificadores que considerem o conteúdo dos documentos e que identifiquem a língua em que estão escritos[19].

4.5 Fase de Detecção de Documentos Paralelos

Esta fase é constituída por três filtros aplicados sequencialmente ou isoladamente. Cada um dos filtros pretende obter grupos de documentos do corpus que sejam paralelos de acordo com os seus critérios intrínsecos, apurando com maior precisão os resultados obtidos pelo filtro aplicado antes dele.

Ao aplicar-se um filtro, são gerados grupos de possíveis documentos paralelos. Estes grupos poderão ser posteriormente passados ao filtro seguinte, de forma a fornecerem a informação inferida pelo filtro anterior, para serem novamente trabalhados. Em todo o processo da detecção de possíveis documentos paralelos apenas são considerados os documentos com mais de 2kB.

Tal como indicado na Ilustração 4.1, a fase de detecção de documentos paralelos inicia-se com a aplicação do filtro através do método baseado nos nomes dos ficheiros (capítulo 4.5.1). Este filtro irá gerar grupos constituídos por documentos candidatos a textos paralelos, passando esta informação ao filtro seguinte que funciona com base no método da proporcionalidade dos tamanhos dos documentos. Este segundo filtro irá trabalhar sobre os grupos de documentos previamente gerados pelo primeiro filtro, descartando os pares de textos que não estiverem de acordo em termos da proporcionalidade de tamanhos (capítulo 4.5.2). Tendo já sido aplicados estes dois filtros anteriores, é utilizado o filtro através do reconhecimento dos possíveis cognatos dos textos em análise (capítulo 4.5.3) dos grupos já identificados como paralelos.

4.5.1 A Interface

Para obter os melhores resultados da aplicação dos filtros apresentados no capítulo anterior foi desenvolvido uma pequena interface onde é possível o utilizador indicar os métodos que pretende usar de acordo com os respectivos parâmetros. O utilizador tem de indicar o corpus que pretende usar e definir o par de línguas em que pretende descobrir documentos paralelos. A interface implementada é constituída por duas áreas, tal como apresentado na Ilustração 4.2. A área do lado direito permite seleccionar o corpus e o par de línguas através da lista "*Corpora*" existente e através da lista "*Pair Language to Use"* (a opção "*All Lang*" permite indicar que se pretende utilizar todos os pares de línguas indicados na lista).

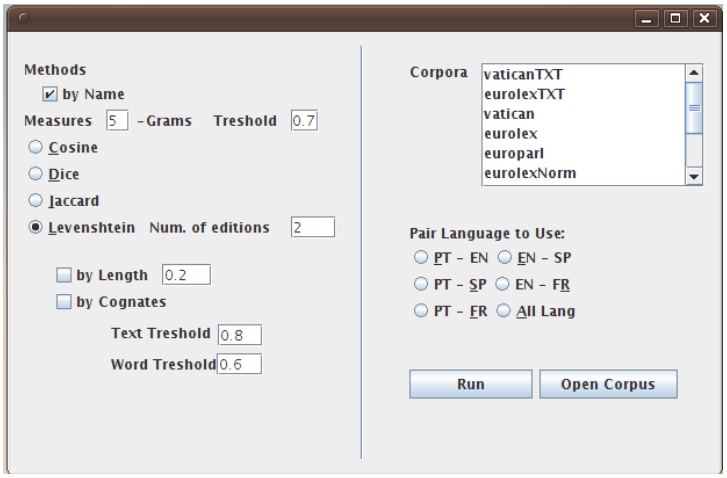

Ilustração 4.2 - Interface da aplicação onde o utilizador tem de indicar o corpus a usar bem como o par de línguas e métodos a serem aplicados.

Na área esquerda da interface devem ser seleccionados os métodos a utilizar no corpus seleccionado bem como os seus parâmetros. São disponibilizados três métodos: método baseado em nomes dos ficheiros ("*by Name")*, método baseado na proporcionalidade dos tamanhos ("*by Length*") e método baseado em cognatos ("*by Cognates*"). Cada um destes métodos necessita de parâmetros que devem ser indicados na interface e que serão depois usados na respectiva implementação (descritas nas próximas capítulos).

Para o método baseado nos nomes dos ficheiros terá de ser indicada uma das distâncias a serem aplicadas (capítulo 4.5.2). No caso das distâncias de semelhança o utilizador tem de introduzir os *n*-grams que devem ser usados e o valor mínimo de semelhança entre os nomes

dos ficheiros, designado por "*threshold*". Se for seleccionada a distância de Levensthein deverá ser introduzido o número máximo de edições permitidas para aceitação da semelhança.

Para usar o método baseado na proporcionalidade dos tamanhos deve ser introduzido o valor de tolerância da proporcionalidade dos tamanhos dos ficheiros (capítulo 4.5.3).

Por último, no método baseado em cognatos devem ser indicados dois parâmetros, o valor mínimo da semelhança entre palavras e o valor mínimo de semelhança dos vectores de textos (mais detalho no capítulo 4.5.4).

Nas próximas secções serão apresentadas as implementações dos métodos indicados anteriormente com referência em mais detalhe da utilização dos respectivos parâmetros.

4.5.2 Filtro através do Método Baseado em Nomes dos Documentos

Este método utiliza uma distância de semelhança indicada pelo utilizador (capítulo 4.5.1) para descobrir os possíveis documentos paralelos tirando partido da boa organização do sistema de directorias em que os Web sites são estruturados, tal como enunciado no capítulo 3.2.1 e no capítulo 4.4.3. Esta abordagem pode ser usada devido aos documentos que são traduções terem URLs muito idênticos e apenas apresentam pequenas variações como as que são apresentadas na lista de segmentos.

Para se obter grupos de textos paralelos através deste método todos os documentos são comparados com todos os outros documentos que não tenham a mesma língua e que ainda não tenham sido identificados como paralelos de outro documento. Quando se calcula o grupo de bitextos para um dado documento numa língua "A" com todos os documentos existentes no corpus numa língua "B" apenas é seleccionado o documento da língua "B" que tiver maior semelhança de URL com o documento da língua "A". Isto porque, é considerado que não existem documentos com mais do que uma tradução na mesma língua e no mesmo Web site.

A semelhança entre os URLs dos documentos é calculada através de uma medida de semelhança escolhida pelo utilizador. Caso esse valor esteja acima de um valor de referência

então os dois documentos são considerados textos paralelos e são colocados no mesmo grupo de documentos se constituírem o par mais semelhante entre as duas línguas. A medida de semelhança e o valor de referência a serem aplicados neste método são indicados pelo utilizador.

Apesar da interface implementada apresentar várias medidas de semelhança, para a discussão de resultados apenas foi utilizada a medida de Levensthein na utilização deste método, já que esta é uma medida precisa para a medição de pequenas variações entre os nomes dos ficheiros. Por exemplo, os nomes dos documentos do corpus do Eur-Lex apresentam diferenças de dois caracteres entre pares paralelos, ou seja, duas edições.

4.5.3 Filtro através do Método Baseado na Proporcionalidade dos Tamanhos dos Documentos

O objectivo principal deste método é através de cada grupo de bitextos, previamente calculados com o filtro anterior (baseado nos nomes dos ficheiros), detectar possíveis documentos do grupo que não sejam paralelos, tal como enunciado no capítulo 3.2.2.

O princípio deste método baseia-se, tal como o seu nome indica, na proporcionalidade dos tamanhos dos documentos. Para dois documentos serem tradução um do outro, ambos têm de ter tamanhos proporcionais, caso contrário não são considerados paralelos.

Para este fim, foi calculado um valor médio aproximado para a proporcionalidade entre documentos escritos em Português, Inglês, Espanhol e Francês através do corpus existente do Eurolex, tirando partido do seu alto grau de paralelismo entre os seus documentos. Os valores encontrados são apresentados na Tabela 4.2 e os valores devem ser lidos apenas na horizontal. Por exemplo, a proporcionalidade dos tamanhos de documentos entre Português e Inglês é de

1,104 e não de 0,922, este segundo valor é a proporcionalidade entre Inglês e Português. Os valores apresentados deveriam ser complementares entre si, mas não o são. Isto porque no corpus utilizado podem existir documentos que não estejam traduzidos em todas as línguas.

Tabela 4.2 - Tabela de proporcionalidades entre pares de línguas.

Idiomas	Português	Inglês	Francês	Espanhol
Português	--	1,104	1,095	1,071
Inglês	0,922	--	0,877	0,884
Francês	0,954	1,194	--	--
Espanhol	0,968	1,162	--	--

Tendo estes valores de referência, o método baseado na proporcionalidade dos tamanhos é aplicado aos grupos de candidatos a documentos paralelos anteriormente gerados pelo filtro baseado nos nomes dos ficheiros. Para cada par de documentos existentes num grupo paralelo este método calcula o valor de proporcionalidade dos tamanhos e compara-o com o valor de referência anteriormente calculado. Se o valor de proporcionalidade for igual ao valor de referência com uma diferença máxima de *T*%, o par de documentos permanece como paralelos, caso contrário, são considerados não paralelos e são removidos do grupo. O valor de *T* é dado pelo utilizador (capítulo 4.5.1).

4.5.4 Filtro através do Método Baseado em Cognatos

A abordagem seguida para implementar este método é muito semelhante à descrita por Ma e Liberman para o sistema BITS [11] (capítulo 3.1). A principal diferença, é que a abordagem seguida neste trabalho utiliza o conceito de cognato na sua generalidade, ao contrário de Ma e Liberman que apenas usam possíveis cognatos na sua versão mais estrita, os homógrafos (tais como datas, nomes, acrónimos e números), que normalmente não são traduzidos. Outra diferença é a forma com que os documentos são comparados. No sistema BITS é utilizado um modelo baseado em distâncias, enquanto que neste trabalho será seguida a abordagem através de frequências enunciada por Chen et al. [8].

No processo de descoberta de possíveis cognatos, não são usadas palavras com um tamanho inferior a três caracteres. Estas palavras não são relevantes para expressar a caracterização individual de um texto, porque palavras como preposições, artigos ou conjunções (em português – e, o, a, dos, das, em, que…; em inglês – and, the, of, for…) são extensamente utilizadas em qualquer tipo de texto.

A implementação deste método percorre individualmente cada um dos grupos de candidatos de documentos paralelos previamente gerados pelos filtros anteriores (capítulos 4.5.2 e 4.5.3). Os documentos que constituem um grupo são comparados entre si ao nível da palavra para a detecção de cognatos.

Seja C um corpus constituído por *x* grupos, $C = \{G_1, G_2, \ldots, G_x\}$. Cada grupo G_i é constituído por dois possíveis documentos paralelos, ou seja $G_i = \{d_1, d_2\}$ onde d_1 e d_2 são documentos em duas línguas diferentes.

Cada documento d_1 existente num grupo G_x é comparado com d_2 de G_x através da identificação de cognatos em ambos os textos. Ao serem identificados os cognatos existentes entre d_1 e d_2 são criadas as representações vectoriais, v_1 e v_2, das frequências dos cognatos em cada um dos textos. Cada posição destes vectores corresponde a cognatos existentes nos dois

textos, ou seja, v_1 e v_2 contêm as frequências das ocorrências do grupo de cognatos $\{c_a, \{c_{b1}, c_{b2...}\}\}$ onde c_a pertence a d_1 e $\{c_{b1}, c_{b2...}\}$ são palavras cognatas de c_a existentes em d_2. Por exemplo, dado {acquired, {adquirido, adquiridos}} significa que num texto em inglês a palavra "acquired" é cognata das duas palavras "adquirido" e "adquiridos" num texto em português. No entanto, as frequências são contabilizadas no seu total, ou seja, supondo que a palavra "acquired" ocorre-se uma vez no texto em inglês e as palavras "adquirido" e "adquiridos" ocorressem uma vez, cada uma, no texto em português. Neste caso, na posição da palavra "acquired" no vector de frequências do documento em inglês estaria indicado o valor 1, enquanto que, na posição correspondente a "adquirido" e "adquiridos" no vector de frequências do documento em português estaria o valor 2 (Ilustração 4.3).

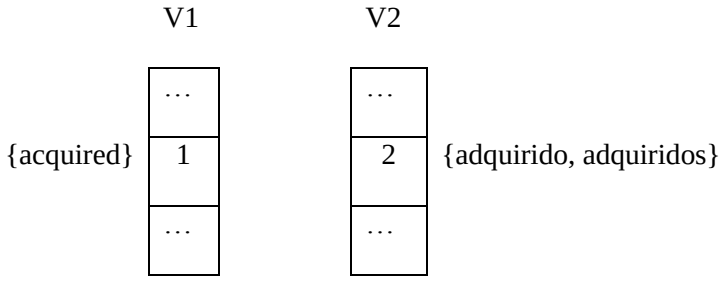

Ilustração 4.3- Vectores de frequências de palavras cognatas entre dois documentos.

Se os vectores de frequências forem semelhantes, pode-se concluir que os documentos são paralelos, caso contrário, os documentos de G_x passam a não serem considerados como possíveis documentos paralelos. Para calcular a semelhança entre dois vectores, usou-se o coeficiente do co-seno (capítulo 3.3).

Estes passos são descritos pelo algoritmo em pseudocódigo da Ilustração 4.4.

```
1. for each Gi in C
2.          wordList = extractWordsFrom(d1)
4.          v[][] = FindCognatesBetween(wordlist, d2)
5.          measure = cosine(v[1], v[2])
6.          if(measure < treshold)
7.                removeGroup(Gi)
```

Ilustração 4.4 - Pseudocódigo do algoritmo do método baseado em cognatos.

A função `FindCognatesBetween(wordList, d2)` (Linha 4 da Ilustração 4.4) é o motor deste algoritmo fazendo a extracção dos cognatos entre dois textos e devolvendo os vectores de frequências, v_1 e v_2, de cada um dos textos. Esta função recebe como argumentos uma lista de palavras pré-extraídas de um dos textos (`wordlist`) e um segundo documento (`d2`). A função percorre `d2` e compara, através da distância de Levenshtein normalizada (capítulo 3.3), cada uma das palavras de `d2` com as existentes na `wordlist` dada. Em cada comparação, se a distância de Levenshtein normalizada entre as duas palavras for superior a um dado valor ("*Word threshold*" introduzido pelo utilizador – capítulo 4.5.1) então essas palavras são consideradas cognatas e as frequências de ocorrências dessas palavras são contabilizadas para a construção dos vectores v_1 e v_2. No final, é aplicado o coeficiente do co-seno aos vectores construídos para o cálculo da semelhança entre as frequências de ocorrência nos respectivos textos dos cognatos identificados. Os documentos são considerados paralelos se o resultado do coeficiente do co-seno for superior ao valor introduzido pelo utilizador no campo "*Text threshold*" da interface.

4.6 Fase de Avaliação Manual do Corpus

Após ter-se identificado os possíveis documentos paralelos de forma automática é necessário avaliar a eficiência desse processo. Para este fim são normalmente usadas as medidas de precisão, *recall* e *f-measure* [1, 3-6, 8, 10, 11].

A partir de um conjunto de documentos D, em que D = $\{d_1, d_2, \ldots, d_n\}$ constitui um corpus paralelo, pode-se definir as seguintes variáveis:

- Seja **X** o número de documentos *identificados* por uma aplicação como paralelos;
- Seja **Y** o número de documentos *correctamente identificados* como paralelos por essa aplicação;
- Seja N o *número real* de documentos paralelos existente no corpus.

As fórmulas para a precisão (P), o *recall* (R) e o *f-measure* (F) são, respectivamente:

$$P = \frac{Y}{X} \quad (9)$$

$$R = \frac{Y}{N} \quad (10)$$

$$F = \frac{2 * P * R}{P + R} \quad (11)$$

Apesar de existirem outras medidas que permitem calcular a eficiência dos algoritmos usados em sistemas de tradução automática, Turian e tal. [20] demonstram que o cálculo do *f-measure* é o mais simples de entender e o mais consensual na prática devido ao seu extenso uso nesta área.

Neste trabalho estas medidas foram determinadas para os corpus do EuroLex e das UN mas não para o corpus do Vaticano. Para este último não foi possível calcular o *recall* por ser um corpus com demasiados documentos. Para grandes quantidades de documentos torna-se difícil o cálculo do parâmetro N, o qual é necessário para determinar o *recall* e consequentemente a *f-measure*, já que este usa o *recall* na sua fórmula.

No cálculo do valor do *recall*, o parâmetro N é o que tem um impacto realista na medição da eficiência do resultado de uma aplicação. No entanto, para ser um valor preciso, real e de confiança, é necessário recorrer à identificação manual dos documentos paralelos existentes num corpus. Por esta razão, o seu cálculo é difícil para grandes quantidades de documentos tornando-se um processo muito demorado necessitando de recursos humanos indisponíveis para a realização deste trabalho.

4.6.1 A Amostra

Para proceder ao cálculo das medidas apresentadas no capítulo anterior é necessário determinar o valor de Y e N (equação 9 e 10). Caso o corpus para o qual se pretende calcular a precisão, o *recall* e o f-measure, contiver um número de documentos muito superior a 300, é escolhida uma amostra aleatória constituída por 274 documentos que tenham sido identificados como paralelos (corresponde ao valor de X – equação 9). Caso contrário, a amostra utiliza todos os pares identificados como paralelos (corresponde ao valor de N – equação 10).

A amostra foi dividida em 15 grupos de 20 pares de documentos e cada grupo foi entregue a uma pessoa que validou cada par de textos como "válido", se o par de textos correspondia a uma tradução, e como "inválido" para os pares que não correspondiam a tradução.

No final foram recolhidos os vários grupos e contabilizados os pares de textos com a indicação "válido" para obter o valor de Y e proceder ao cálculo do valor da precisão do corpus.

4.6.2 A Aplicação

Para ajudar no processo de validação dos pares candidatos de documentos paralelos foi desenvolvida uma aplicação (Ilustração 4.5) que mostra, de cada vez, um par de documentos que são apresentados, lado a lado, a um utilizador para este os validar, ou não, como paralelos. Esta aplicação permite guardar as validações para posteriormente serem contabilizadas e fornecerem os valores necessários para o cálculo da precisão.

A aplicação implementada para este efeito permite visualizar pares de documentos identificados (ou não) como paralelos de forma paralela. Assim, o utilizador pode visualizar dois documentos em simultâneo (1) com opção de *scroll* simultâneo dos textos (2). Se os textos forem paralelos, o utilizador deverá premir o botão "*Válido(s)*", caso contrário, deverá optar pelo botão "*Inválido(n)*" (3). Existem ainda três botões (4) para as seguintes funcionalidades: "*Definir Directoria com Ficheiros*", "*Guardar(g)*" e "*Fechar(f)*". O primeiro botão permite definir a localização dos documentos a apresentar e os outros dois botões permitem guardar o resultado da avaliação e fechar o programa. Tendo guardado os resultados obtidos e fechado normalmente o programa, o validador pode continuar noutra sessão o trabalho de validação que, por qualquer motivo, não acabou numa só sessão.

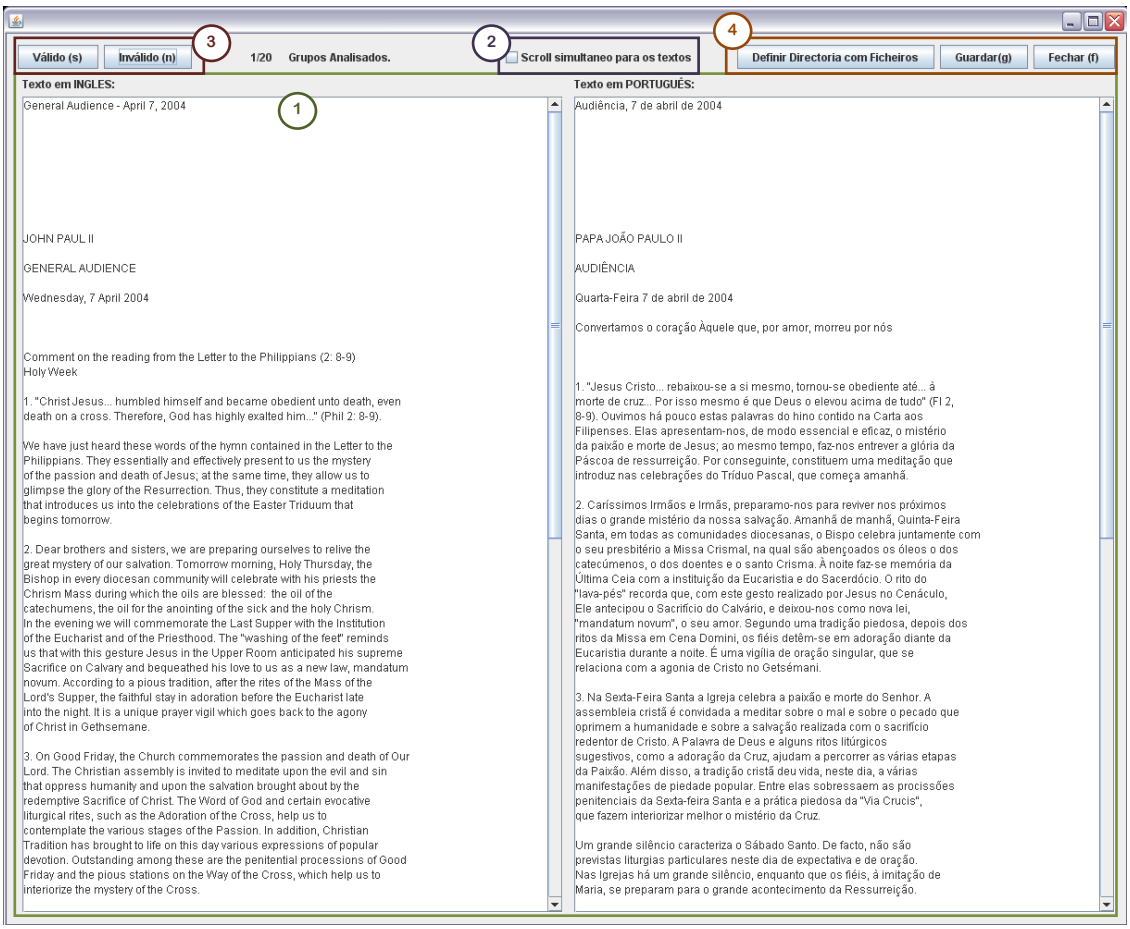

Ilustração 4.5 - Aplicação para validação manual dos pares candidatos de textos paralelos.

5. Resultados

Para a avaliação da eficiência dos filtros usados, foi calculado o valor da precisão e da f-measure (capítulo 4.6) para os corpus do Eur-Lex, das Nações Unidas (UN) e do Vaticano (capítulo 4.2). O valor de recall (capítulo 4.6) foi calculado apenas para os corpus Eur-Lex e UN.

Cada um destes corpus foi submetido a quatro testes. No primeiro teste apenas foi usado o filtro baseado nos nomes dos ficheiros (1. Nomes), o segundo teste foi usado o mesmo filtro anterior juntamente com o filtro baseado nas proporcionalidades dos tamanhos dos ficheiros (2. Nomes e Tamanhos). O terceiro teste incluiu os dois filtros anteriores seguido da aplicação do filtro baseado em cognatos (3. Nomes, Tamanhos e Cognatos). Por fim, o último teste utilizou os filtros baseados nos nomes dos ficheiros e os cognatos (4. Nomes e Cognatos).

O uso destes filtros apenas detectou no máximo 271 pares candidatos de documentos paralelos no corpus Eur-lex. No corpus das UN foram detectados 83 pares candidatos de documentos paralelos. Enquanto que, no corpus do Vaticano detectou um máximo de 3519 pares candidatos a documentos paralelos. Por esta razão, para o cálculo dos valores da precisão, no caso do corpus do Vaticano foi utilizado uma amostra de 274 pares candidatos de documentos paralelos como descrito no capítulo 4.6.1., ao contrário dos corpus do Eur-lex e das UN onde foram usados todos os pares encontrados.

Cada um dos testes foi executado com vários parâmetros de forma a estudar a influência dos parâmetros nos vários corpora. Estes parâmetros são introduzidos pelo utilizador na interface da aplicação (capítulo 4.5.1). No primeiro teste (1. Nomes) foram utilizados como parâmetro do número de edição máximo a aplicar nos nomes dos ficheiros os valores 2, 4, 6 e 8 (Anexo 0). Os melhores resultados obtidos foram com duas edições para o corpus do Vaticano e Euro-Lex, e quatro edições para o corpus das UN. Os resultados obtidos para estes parâmetros são apresentados

na Tabela 5.1. A escolha destes resultados, em comparação com os apresentados em anexo 7, recaíram sobre os valores mais altos de Recall e precisão, pois reflectem um valor maior de f-measure.

Da análise da Tabela 5.1 podemos verificar que o corpus com melhores resultados é o Euro-Lex, o que já era esperado uma vez que os nomes dos ficheiros são muito bem construídos, tal como discutido no capítulo 4.5.2.

O corpus do Vaticano não consegue obter tão bons resultados como o do Euro-Lex, mas tem uma precisão elevada perto dos 100%. Este facto deve-se principalmente à nomeação dos ficheiros não ser tão rígida como acontece no corpus do Euro-Lex.

Em relação ao corpus das UN os valores apresentados para a precisão e para o recall indicam que o método dos nomes identifica alguns falsos positivos devido à precisão ser mais baixa que o recall.

Tabela 5.1 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{271}{271} = 1.000$	$\frac{271}{271} = 1.000$	1.000	$\frac{63}{89} = 0.708$	$\frac{63}{63} = 1.0$	0.803	$\frac{276}{289} = 0.955$
ENSP	$\frac{266}{266} = 1.000$	$\frac{266}{266} = 1.000$	1.000	$\frac{87}{97} = 0.897$	$\frac{87}{87} = 1.0$	0.946	$\frac{273}{287} = 0.951$
PTFR	$\frac{282}{282} = 1.000$	$\frac{282}{282} = 1.000$	1.000	---	---	---	$\frac{281}{288} = 0.976$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1.000	---	---	---	$\frac{280}{290} = 0.966$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1.000	---	---	---	$\frac{270}{286} = 0.944$

A próxima tabela (Tabela 5.2) apresenta os melhores resultados da aplicação do método baseado em nomes de ficheiros seguido do método baseado na proporcionalidade dos ficheiros. Esta tabela apresenta os resultados com número de edições igual a dois e uma taxa de tolerância de 40% para as proporcionalidades dos tamanhos dos textos (capítulo 4.5.3). Em anexo são apresentadas as tabelas análogas à Tabela 5.2 para taxas de tolerância de 0.1, 0.2 e 0.4 (anexo 7).

A aplicação do método baseado na proporcionalidade dos tamanhos dos ficheiros depois do método baseado em nomes dos ficheiros apresenta resultados positivos em relação ao corpus das UN. Isto deve-se ao facto de conseguir restringir melhor os valores da precisão (apresentado um aumento da precisão) sem mexer nos valores do recall em relação ao método aplicado anteriormente (com valores apresentados na Tabela 5.1).

Em relação aos corpus do Euro-Lex apresenta uma ligeira descida no recall para os pares de línguas Inglês-Francês (ENFR) e Português-Francês (PTFR). Este resultado deve-se à perda de um par de documentos paralelos para Inglês-Francês e de dois pares de documentos paralelos em Português-Francês. Esta perda é esperada já que o estudo das proporcionalidades entre os tamanhos dos documentos baseia-se na média, logo poderão existir, e de facto existem, casos que tenham uma diferença de tamanhos muito superiores, ou inferiores, ao da média estudada.

No caso do corpus do Vaticano verifica-se um aumento da precisão à custa da perda de alguns pares de documentos paralelos em todas as línguas. Esta perda de bitextos reflecte uma descida do valor do recall, apesar de não ser possível calculá-lo podemos verificar isso porque o número de documentos identificados como pares candidatos de documentos paralelos é inferior aos que foram identificados através do método anterior (Tabela 5.2).

Tabela 5.2 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% do método.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{270}{270} = 1.000$	$\frac{270}{271} = 0.996$	0.998	$\frac{63}{83} = 0.759$	$\frac{63}{63} = 1.000$	0.863	$\frac{271}{280} = 0.968$
ENSP	$\frac{266}{266} = 1.000$	$\frac{266}{266} = 1.000$	1.000	$\frac{87}{95} = 0.916$	$\frac{87}{87} = 1.000$	0.956	$\frac{270}{282} = 0.957$
PTFR	$\frac{280}{280} = 1.000$	$\frac{280}{282} = 0.993$	0.996	---	---	---	$\frac{267}{272} = 0.982$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1.000	---	---	---	$\frac{273}{279} = 0.950$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1.000	---	---	---	$\frac{266}{280} = 0.950$

Depois da aplicação dos métodos baseados nos nomes dos ficheiros e na proporcionalidade dos tamanhos, foi aplicado o método baseado em cognatos. Para cada um dos valores de tolerância da proporcionalidade dos tamanhos (10%, 20% e 40%) foi aplicado o método

baseado em cognatos com valores de semelhança entre palavras de 60%, 70%, 80% e 100%, e com valores de semelhança entre os vectores de frequências dos cognatos de 70% e 80%.

Com a aplicação do método baseado em cognatos verificou-se que a escolha dos parâmetros é bastante dependente dos pares de línguas que se estão a utilizar (todos os resultados estão apresentados nas tabelas do anexo 7). No entanto, na maioria das vezes verifica-se uma perda significativa de pares de documentos paralelos. Por exemplo, para Português-Francês o corpus do Vaticano, com semelhança a 80% entre palavras, 70% de semelhança entre os vectores de frequência e 20% de tolerância entre as proporcionalidades dos tamanhos (Tabela 8.19), consegue apresentar uma precisão de 100%, mas apenas são identificados 83 de 300 pares de documentos paralelos, o que sugere um recall muito baixo. Já o par de línguas Inglês-Francês no corpus do Vaticano para uma semelhança de 60% entre palavras, uma tolerância de 80% de semelhança entre os vectores de frequências e uma tolerância de 10% na proporcionalidade dos tamanhos (Tabela 8.13) apresenta uma precisão de 100%, embora apenas identifique 47 pares de documentos como sendo paralelos.

No geral, a aplicação do método baseado em cognatos aumenta a precisão da identificação dos pares de documentos paralelos, ficando muito próxima dos 100%. No entanto, este aumento implica uma descida dos valores de recall. Os resultados completos podem ser consultados no anexo 7.

O método baseado em cognatos aplicado depois do método baseado nos nomes dos ficheiros não mostra resultados de 100% como os resultados onde foram aplicados o método baseado na proporcionalidade dos tamanhos, no entanto, consegue identificar significativamente mais pares de documentos paralelos. Ou seja, demonstra maior recall.

Através das tabelas do anexo 7 onde são apresentados os resultados referentes ao método baseado em cognatos pode-se verificar que à medida que se aumenta a semelhança entre palavras para a detecção de cognatos maior é o f-measure conseguido. Este resultado demonstra que o recall e a precisão aumentam com a semelhança entre palavras. No entanto, este comportamento pode dever-se ao facto da implementação realizada não fazer reconhecimento de falsos cognatos, ou seja, para semelhança mais baixa entre palavras são introduzidos um maior número de falsos cognatos na detecção de possíveis documentos paralelos. Isto faz com que os valores do recall e da precisão baixem e demonstra a necessidade de realizar um estudo com detecção de falsos cognatos para demonstrar o seu verdadeiro impacto nesta abordagem.

6. Conclusões e Trabalho Futuro

Os resultados obtidos foram muito positivos em relação ao método baseado em cognatos demonstrando que estes são uma característica relevante na detecção automática de documentos paralelos. Em corpus com uma boa organização e com traduções fiáveis, consegue-se uma precisão e recall de 100%. Já em corpora mais mal estruturados e com traduções menos rígidas, como o corpus das UN, os valores apresentados chegaram a 93% e 98% para a precisão e o recall, respectivamente.

A aplicação dos métodos baseados nos nomes dos ficheiros e em cognatos demonstra valores de recall mais altos com uma perda muito baixa do valor da precisão mostrando que seria útil aperfeiçoar o algoritmo de descoberta de cognatos de forma a ser mais eficiente e rápido. Uma solução poderá passar pela implementação de classificadores que permitam aprender cognatos e falsos cognatos entre pares de línguas. Ao se acrescentar a descoberta de falsos cognatos prevê-se a obtenção de resultados mais exactos e com muito menos ruído no seu cálculo.

Devido à implementação do método de identificação da língua não ter qualquer conhecimento sobre as línguas que detecta, existem documentos em que não é atribuído qualquer idioma. Estes documentos ficam à partida excluídos do processo de detecção de documentos paralelos. Sendo que podem ter um efeito negativo para o cálculo do valor do recall. De forma a se colmatar esta situação, futuramente deve ser utilizado um identificador de língua que, de preferência, possa ser treinado para qualquer conjunto de línguas.

No trabalho realizado por Zhang[9] é apresentada uma tabela comparativa entre vários sistemas (Tabela 6.1). Na tentativa de comparar os resultados obtidos nesta dissertação com

os de outros trabalhos, concluiu-se que para essa comparação seria necessário ter-se a mesma implementação ou os mesmos corpora de forma a poderem ser tiradas conclusões relevantes e com fundamento. Os resultados obtidos neste trabalho são distintos nos três corpora utilizados, o que demonstra que as características desses corpora são cruciais para este tipo de comparação. Contudo, os resultados obtidos no trabalho descrito nesta dissertação são da mesma gama dos valores obtidos com o sistema WPDE, a nível de precisão e recall (f-measure), feitas as devidas ressalvas no que se refere aos corpora utilizados e às línguas alvo.

Tabela 6.1 – Resultados sumarizados de vários sistemas implementados por outros autores (tabela retirada e adaptada de [9]).

	Precisão	**Recall**
PTMiner	90%	--
STRAND	98%	61%
PTI	93%	96%
WPDE	95%	97%

7. Bibliografia

1. Resnik, P., *Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text*, in *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. 1998, Springer-Verlag.
2. Chen, J. and J.-Y. Nie, *Parallel Web Text Mining for Cross-Language IR*, in *In In Proc. of RIAO*. 2000. p. 62-77.
3. Nie, J.-Y. and J. Cai, *Filtering Noisy Parallel Corpora of Web Pages*, in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on 2001*. 2001: Tucson, AZ, USA. p. 453-458.
4. Smith, N.A., *From words to corpora: recognizing translation*, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. 2002, Association for Computational Linguistics.
5. Resnik, P. and N.A. Smith, *The Web as a parallel corpus.* Comput. Linguist., 2003. **29**(3): p. 349-380.
6. Nadeau, D. and G. Foster. *Real-Time Identification of Parallel Text from Bilingual Newsfeed*. in *Proceedings of the Computational Linguistic in the North-East (CLINE'2004)*. 2004. Montréal, Québec, Canada.
7. Resnik, P., *Mining the Web for bilingual text*, in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999, Association for Computational Linguistics: College Park, Maryland.
8. Chen, J., R. Chau, and C.-H. Yeh, *Discovering parallel text from the World Wide Web*, in *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32*. 2004, Australian Computer Society, Inc.: Dunedin, New Zealand. p. 157-161.

9. Zhang, Y., et al. *Automatic Acquisition of Chinese-English Parallel Corpus from the Web*. in *28th European Conference on Information Retrieval (ECIR)*. 2006: Springer Berlin / Heidelberg.
10. Patry, A. and P. Langlais. *Automatic Identification of Parallel Documents With Light or Without Linguistic Resources*. in *AI 2005*. 2005: Springer Berlin / Heidelberg.
11. Ma, X. and M. Liberman, Y., *BITS: A Method for Bilingual Text Search over the Web*, in *Machine Translation Summit VII*. 1999, Kent Ridge Digital Labs: National University of Singapore.
12. Smith, N., A., *Detection of Translational Equivalence*, in *Computer Science Department*. 2001, University of Maryland College Park.
13. Ribeiro, A., J.G.P. Lopes, and J. Mexia, *Extracting Equivalents from Aligned Parallel Texts: Comparison of Measures of Similarity*, in *Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI: Advances in Artificial Intelligence*. 2000, Springer-Verlag: Atibaia, SP, Brazil. p. 339-349.
14. Ribeiro, A.M.B., *Parallel Texts Alignment for Extraction of Translation Equivalents*, in *Departamento de Informática*. 2002, Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa: Lisboa. p. 148.
15. Simard, M., G.F. Foster, and P. Isabelle, *Using cognates to align sentences in bilingual corpora*, in *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2*. 1993, IBM Press: Toronto, Ontario, Canada.
16. Gale, W.A. and K.W. Church, *A program for aligning sentences in bilingual corpora*, in *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. 1991, Association for Computational Linguistics: Berkeley, California.
17. Melamed, I.D., *Bitext maps and alignment via pattern recognition.* Comput. Linguist., 1999. **25**(1): p. 107-130.
18. Church, K.W., *Char_align: a program for aligning parallel texts at the character level*, in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. 1993, Association for Computational Linguistics: Columbus, Ohio.
19. Reis, J.V.P.d., *Automatic Language Identification in Text*. 2008, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.
20. Turian, J.P., L. Shen, and I.D. Melamed. *Evaluation of Machine Translation and its Evaluation*. in *Proceedings of MT Summit IX*. 2003. New Orleans, USA.

8. Anexos

8.1.1 Tabelas de resultados

Aplicação do método baseado nos nomes dos ficheiros.

Tabela 8.1 - Resultados obtidos dos diferentes corpus utilizando o método baseado em nomes de ficheiros com distância de edição igual a 2.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{271}{271} = 1.000$	$\frac{271}{271} = 1.000$	1	$\frac{57}{79} = 0.722$	$\frac{57}{63} = 0.905$	0.803	$\frac{276}{289} = 0.955$
ENSP	$\frac{266}{266} = 1.000$	$\frac{266}{266} = 1.000$	1	$\frac{76}{83} = 0.916$	$\frac{76}{87} = 0.874$	0.895	$\frac{273}{287} = 0.951$
PTFR	$\frac{282}{282} = 1.000$	$\frac{282}{282} = 1.000$	1	---	---	---	$\frac{281}{288} = 0.976$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1	---	---	---	$\frac{280}{290} = 0.966$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1	---	---	---	$\frac{270}{286} = 0.944$

Tabela 8.2 - Resultados obtidos dos diferentes corpus utilizando o método baseado em nomes de ficheiros com distância de edição igual a 4.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{268}{272} = 0.985$	$\frac{268}{271} = 0.989$	0.987	$\frac{63}{89} = 0.708$	$\frac{63}{63} = 1.000$	0.829	$\frac{272}{285} = 0.954$
ENSP	$\frac{264}{270} = 0.978$	$\frac{264}{266} = 0.992$	0.985	$\frac{87}{95} = 0.916$	$\frac{87}{87} = 1.000$	0.956	$\frac{268}{282} = 0.950$
PTFR	$\frac{247}{282} = 0.876$	$\frac{247}{282} = 0.876$	0.876	---	---	---	$\frac{274}{281} = 0.975$
PTSP	$\frac{268}{283} = 0.947$	$\frac{268}{283} = 0.947$	0.947	---	---	---	$\frac{272}{282} = 0.965$
PTEN	$\frac{259}{270} = 0.959$	$\frac{259}{262} = 0.989$	0.974	---	---	---	$\frac{262}{278} = 0.942$

Tabela 8.3 - Resultados obtidos dos diferentes corpus utilizando o método baseado em nomes de ficheiros com distância de edição igual a 6.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{267}{272} = 0.982$	$\frac{267}{271} = 0.985$	0.983	$\frac{63}{89} = 0.708$	$\frac{63}{63} = 1.000$	0.829	$\frac{264}{277} = 0.953$
ENSP	$\frac{264}{270} = 0.978$	$\frac{264}{266} = 0.992$	0.985	$\frac{87}{97} = 0.897$	$\frac{87}{87} = 1.000$	0.946	$\frac{264}{278} = 0.950$
PTFR	$\frac{230}{283} = 0.813$	$\frac{230}{282} = 0.816$	0.814	---	---	---	$\frac{268}{275} = 0.975$
PTSP	$\frac{260}{283} = 0.919$	$\frac{260}{283} = 0.919$	0.919	---	---	---	$\frac{259}{269} = 0.963$
PTEN	$\frac{257}{273} = 0.941$	$\frac{257}{262} = 0.981$	0.961	---	---	---	$\frac{256}{272} = 0.948$

Tabela 8.4 - Resultados obtidos dos diferentes corpus utilizando o método baseado em nomes de ficheiros com distância de edição igual a 8.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{267}{272} = 0.982$	$\frac{267}{271} = 0.985$	0.983	$\frac{63}{89} = 0.708$	$\frac{63}{63} = 1.000$	0.829	$\frac{264}{277} = 0.953$
ENSP	$\frac{264}{270} = 0.978$	$\frac{264}{266} = 0.992$	0.985	$\frac{87}{97} = 0.897$	$\frac{87}{87} = 1.000$	0.946	$\frac{264}{278} = 0.950$
PTFR	$\frac{230}{283} = 0.813$	$\frac{230}{282} = 0.816$	0.814	---	---	---	$\frac{268}{275} = 0.975$
PTSP	$\frac{257}{283} = 0.908$	$\frac{257}{283} = 0.908$	0.908	---	---	---	$\frac{256}{266} = 0.962$
PTEN	$\frac{257}{273} = 0.941$	$\frac{257}{262} = 0.981$	0.961	---	---	---	$\frac{256}{272} = 0.948$

Aplicação dos métodos baseados nos nomes dos ficheiros e na proporcionalidade dos tamanhos.

Tabela 8.5 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{209}{209} = 1.000$	$\frac{209}{271} = 0.771$	0.871	$\frac{42}{49} = 0.857$	$\frac{42}{63} = 0.667$	0.750	$\frac{191}{193} = 0.990$
ENSP	$\frac{244}{244} = 1.000$	$\frac{244}{266} = 0.917$	0.957	$\frac{50}{52} = 0.962$	$\frac{50}{87} = 0.575$	0.720	$\frac{112}{114} = 0.982$
PTFR	$\frac{229}{229} = 1.000$	$\frac{229}{282} = 0.812$	0.896	---	---	---	$\frac{95}{95} = 1.000$
PTSP	$\frac{278}{278} = 1.000$	$\frac{278}{283} = 0.982$	0.991	---	---	---	$\frac{179}{181} = 0.989$
PTEN	$\frac{237}{237} = 1.000$	$\frac{237}{262} = 0.905$	0.950	---	---	---	$\frac{160}{168} = 0.952$

Tabela 8.6 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{269}{269} = 1.000$	$\frac{269}{271} = 0.993$	0.996	$\frac{58}{71} = 0.817$	$\frac{58}{63} = 0.921$	0.866	$\frac{257}{262} = 0.981$
ENSP	$\frac{264}{265} = 1.000$	$\frac{264}{266} = 0.996$	0.998	$\frac{78}{83} = 0.940$	$\frac{78}{87} = 0.897$	0.918	$\frac{233}{239} = 0.975$
PTFR	$\frac{271}{271} = 1.000$	$\frac{271}{282} = 0.961$	0.980	---	---	---	$\frac{177}{181} = 0.978$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1.000	---	---	---	$\frac{253}{257} = 0.984$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1.000	---	---	---	$\frac{239}{250} = 0.956$

Tabela 8.7 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 30% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{269}{269} = 1.000$	$\frac{269}{271} = 0.993$	0.996	$\frac{59}{73} = 0.808$	$\frac{59}{63} = 0.937$	0.868	$\frac{265}{272} = 0.974$
ENSP	$\frac{265}{265} = 1.000$	$\frac{265}{266} = 0.996$	0.998	$\frac{87}{95} = 0.916$	$\frac{87}{87} = 1.000$	0.956	$\frac{263}{273} = 0.963$
PTFR	$\frac{275}{275} = 1.000$	$\frac{275}{282} = 0.975$	0.987	---	---	---	$\frac{226}{231} = 0.978$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1.000	---	---	---	$\frac{267}{272} = 0.982$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1.000	---	---	---	$\frac{261}{270} = 0.967$

Tabela 8.8 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{270}{270} = 1.000$	$\frac{270}{271} = 0.996$	0.998	$\frac{63}{83} = 0.759$	$\frac{63}{63} = 1.000$	0.863	$\frac{271}{280} = 0.968$
ENSP	$\frac{266}{266} = 1.000$	$\frac{266}{266} = 1.000$	1.000	$\frac{87}{95} = 0.916$	$\frac{87}{87} = 1.000$	0.956	$\frac{270}{282} = 0.957$
PTFR	$\frac{280}{280} = 1.000$	$\frac{280}{282} = 0.993$	0.996	---	---	---	$\frac{267}{272} = 0.982$
PTSP	$\frac{283}{283} = 1.000$	$\frac{283}{283} = 1.000$	1.000	---	---	---	$\frac{273}{279} = 0.950$
PTEN	$\frac{262}{262} = 1.000$	$\frac{262}{262} = 1.000$	1.000	---	---	---	$\frac{266}{280} = 0.950$

Aplicação dos métodos baseados nos nomes dos ficheiros, na proporcionalidade dos tamanhos e em cognatos.

Tabela 8.9 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{135}{135} = 1.000$	$\frac{135}{271} = 0.498$	0.665	$\frac{29}{32} = 0.906$	$\frac{29}{63} = 0.492$	0.638	$\frac{88}{89} = 0.989$
ENSP	$\frac{108}{109} = 0.991$	$\frac{108}{266} = 0.406$	0.576	$\frac{48}{48} = 1.000$	$\frac{48}{87} = 0.552$	0.711	$\frac{66}{67} = 0.985$
PTFR	$\frac{163}{163} = 1.000$	$\frac{163}{282} = 0.578$	0.733	---	---	---	$\frac{56}{56} = 1.000$
PTSP	$\frac{260}{260} = 1.000$	$\frac{260}{283} = 0.919$	0.958	---	---	---	$\frac{172}{172} = 1.000$
PTEN	$\frac{130}{130} = 1.000$	$\frac{130}{262} = 0.496$	0.663	---	---	---	$\frac{86}{90} = 0.956$

Tabela 8.10 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{166}{166} = 1.000$	$\frac{166}{271} = 0.613$	0.760	$\frac{31}{35} = 0.886$	$\frac{31}{63} = 0.492$	0.633	$\frac{138}{139} = 0.993$
ENSP	$\frac{172}{173} = 0.994$	$\frac{172}{266} = 0.647$	0.784	$\frac{48}{49} = 0.980$	$\frac{48}{87} = 0.552$	0.706	$\frac{87}{88} = 0.989$
PTFR	$\frac{205}{205} = 1.000$	$\frac{205}{282} = 0.727$	0.842	---	---	---	$\frac{73}{73} = 1.000$
PTSP	$\frac{274}{274} = 1.000$	$\frac{274}{283} = 0.968$	0.984	---	---	---	$\frac{175}{177} = 0.989$
PTEN	$\frac{180}{180} = 1.000$	$\frac{180}{262} = 0.687$	0.814	---	---	---	$\frac{110}{115} = 0.957$

Tabela 8.11 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{181}{181} = 1.000$	$\frac{181}{271} = 0.668$	0.801	$\frac{34}{39} = 0.872$	$\frac{34}{63} = 0.540$	0.667	$\frac{175}{177} = 0.989$
ENSP	$\frac{178}{179} = 0.994$	$\frac{178}{266} = 0.669$	0.780	$\frac{46}{48} = 0.958$	$\frac{46}{87} = 0.529$	0.682	$\frac{91}{93} = 0.978$
PTFR	$\frac{209}{209} = 1.000$	$\frac{209}{282} = 0.741$	0.851	---	---	---	$\frac{83}{83} = 1.000$
PTSP	$\frac{276}{276} = 1.000$	$\frac{276}{283} = 0.975$	0.987	---	---	---	$\frac{176}{177} = 0.994$
PTEN	$\frac{187}{187} = 1.000$	$\frac{187}{262} = 0.714$	0.833	---	---	---	$\frac{134}{139} = 0.964$

Tabela 8.12 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{187}{187} = 1.000$	$\frac{187}{271} = 0.690$	0.817	$\frac{32}{37} = 0.865$	$\frac{32}{63} = 0.508$	0.640	$\frac{181}{183} = 0.995$
ENSP	$\frac{183}{184} = 0.994$	$\frac{183}{266} = 0.688$	0.813	$\frac{47}{49} = 0.959$	$\frac{47}{87} = 0.540$	0.691	$\frac{95}{97} = 0.979$
PTFR	$\frac{209}{209} = 1.000$	$\frac{209}{282} = 0.741$	0.851	---	---	---	$\frac{82}{82} = 1.000$
PTSP	$\frac{277}{277} = 1.000$	$\frac{277}{283} = 0.979$	0.989	---	---	---	$\frac{176}{177} = 0.994$
PTEN	$\frac{184}{184} = 1.000$	$\frac{184}{262} = 0.702$	0.825	---	---	---	$\frac{140}{145} = 0.966$

Tabela 8.13 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{105}{105} = 1.000$	$\frac{105}{271} = 0.387$	0.558	$\frac{20}{22} = 0.909$	$\frac{20}{63} = 0.317$	0.470	$\frac{47}{47} = 1.000$
ENSP	$\frac{83}{84} = 0.988$	$\frac{83}{266} = 0.312$	0.474	$\frac{39}{39} = 1.000$	$\frac{39}{87} = 0.448$	0.619	$\frac{33}{34} = 0.971$
PTFR	$\frac{126}{126} = 1.000$	$\frac{126}{282} = 0.447$	0.618	---	---	---	$\frac{40}{40} = 1.000$
PTSP	$\frac{219}{219} = 1.000$	$\frac{219}{283} = 0.774$	0.873	---	---	---	$\frac{163}{163} = 1.000$
PTEN	$\frac{96}{96} = 1.000$	$\frac{96}{262} = 0.366$	0.536	---	---	---	$\frac{59}{62} = 0.952$

Tabela 8.14 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{124}{124} = 1.000$	$\frac{124}{271} = 0.458$	0.628	$\frac{21}{23} = 0.913$	$\frac{21}{63} = 0.333$	0.488	$\frac{94}{95} = 0.989$
ENSP	$\frac{132}{133} = 0.992$	$\frac{132}{266} = 0.496$	0.661	$\frac{42}{42} = 1.000$	$\frac{42}{87} = 0.483$	0.651	$\frac{64}{65} = 0.985$
PTFR	$\frac{172}{172} = 1.000$	$\frac{172}{282} = 0.610$	0.758	---	---	---	$\frac{65}{65} = 1.000$
PTSP	$\frac{249}{249} = 1.000$	$\frac{249}{283} = 0.880$	0.936	---	---	---	$\frac{169}{170} = 0.994$
PTEN	$\frac{146}{146} = 1.000$	$\frac{146}{262} = 0.557$	0.715	---	---	---	$\frac{84}{87} = 0.966$

Tabela 8.15 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{156}{156} = 1.000$	$\frac{156}{271} = 0.576$	0.731	$\frac{26}{29} = 0.897$	$\frac{26}{63} = 0.413$	0.566	$\frac{146}{148} = 0.986$
ENSP	$\frac{144}{145} = 0.993$	$\frac{144}{266} = 0.541$	0.700	$\frac{41}{42} = 0.976$	$\frac{41}{87} = 0.471$	0.635	$\frac{76}{77} = 0.987$
PTFR	$\frac{192}{192} = 1.000$	$\frac{192}{282} = 0.681$	0.810	---	---	---	$\frac{78}{78} = 1.000$
PTSP	$\frac{268}{268} = 1.000$	$\frac{268}{283} = 0.947$	0.973	---	---	---	$\frac{171}{172} = 0.994$
PTEN	$\frac{158}{158} = 1.000$	$\frac{158}{262} = 0.603$	0.752	---	---	---	$\frac{116}{121} = 0.959$

Tabela 8.16 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 10% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{169}{169} = 1.000$	$\frac{169}{271} = 0.624$	0.768	$\frac{26}{30} = 0.867$	$\frac{26}{63} = 0.413$	0.559	$\frac{164}{166} = 0.988$
ENSP	$\frac{156}{156} = 1.000$	$\frac{156}{266} = 0.586$	0.739	$\frac{43}{45} = 0.956$	$\frac{43}{87} = 0.494$	0.651	$\frac{83}{85} = 0.976$
PTFR	$\frac{195}{195} = 1.000$	$\frac{195}{282} = 0.691$	0.817	---	---	---	$\frac{78}{78} = 1.000$
PTSP	$\frac{269}{269} = 1.000$	$\frac{269}{283} = 0.951$	0.975	---	---	---	$\frac{171}{172} = 0.994$
PTEN	$\frac{163}{163} = 1.000$	$\frac{163}{262} = 0.622$	0.767	---	---	---	$\frac{119}{123} = 0.967$

Tabela 8.17 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{173}{173} = 1.000$	$\frac{173}{271} = 0.638$	0.779	$\frac{41}{49} = 0.837$	$\frac{41}{63} = 0.651$	0.732	$\frac{123}{126} = 0.976$
ENSP	$\frac{126}{127} = 0.992$	$\frac{126}{266} = 0.474$	0.641	$\frac{73}{74} = 0.986$	$\frac{73}{87} = 0.839$	0.907	$\frac{152}{155} = 0.981$
PTFR	$\frac{174}{174} = 1.000$	$\frac{174}{282} = 0.617$	0.763	---	---	---	$\frac{99}{102} = 0.971$
PTSP	$\frac{265}{265} = 1.000$	$\frac{265}{283} = 0.936$	0.967	---	---	---	$\frac{246}{248} = 0.992$
PTEN	$\frac{135}{135} = 1.000$	$\frac{135}{262} = 0.515$	0.680	---	---	---	$\frac{123}{129} = 0.953$

Tabela 8.18 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{215}{215} = 1.000$	$\frac{215}{271} = 0.793$	0.885	$\frac{42}{50} = 0.840$	$\frac{42}{63} = 0.667$	0.744	$\frac{138}{139} = 0.993$
ENSP	$\frac{193}{194} = 0.995$	$\frac{193}{266} = 0.726$	0.839	$\frac{73}{75} = 0.973$	$\frac{73}{87} = 0.839$	0.901	$\frac{192}{195} = 0.985$
PTFR	$\frac{230}{230} = 1.000$	$\frac{230}{282} = 0.816$	0.899	---	---	---	$\frac{129}{131} = 0.985$
PTSP	$\frac{279}{279} = 1.000$	$\frac{279}{283} = 0.986$	0.992	---	---	---	$\frac{249}{253} = 0.984$
PTEN	$\frac{195}{195} = 1.000$	$\frac{195}{262} = 0.744$	0.853	---	---	---	$\frac{164}{171} = 0.959$

Tabela 8.19 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{234}{234} = 1.000$	$\frac{234}{271} = 0.863$	0.926	$\frac{45}{54} = 0.833$	$\frac{45}{63} = 0.714$	0.769	$\frac{175}{177} = 0.989$
ENSP	$\frac{198}{199} = 0.995$	$\frac{198}{266} = 0.748$	0.854	$\frac{72}{75} = 0.960$	$\frac{72}{87} = 0.828$	0.889	$\frac{91}{93} = 0.978$
PTFR	$\frac{237}{237} = 1.000$	$\frac{237}{282} = 0.748$	0.856	---	---	---	$\frac{83}{83} = 1.000$
PTSP	$\frac{281}{281} = 1.000$	$\frac{281}{283} = 0.993$	0.996	---	---	---	$\frac{176}{177} = 0.994$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{134}{139} = 0.964$

Tabela 8.20 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{242}{242} = 1.000$	$\frac{242}{271} = 0.893$	0.943	$\frac{43}{54} = 0.796$	$\frac{43}{63} = 0.683$	0.735	$\frac{239}{244} = 0.980$
ENSP	$\frac{203}{204} = 0.995$	$\frac{203}{266} = 0.767$	0.866	$\frac{73}{76} = 0.961$	$\frac{73}{87} = 0.839$	0.896	$\frac{202}{207} = 0.976$
PTFR	$\frac{242}{242} = 1.000$	$\frac{242}{282} = 0.858$	0.924	---	---	---	$\frac{156}{160} = 0.975$
PTSP	$\frac{282}{282} = 1.000$	$\frac{282}{283} = 0.996$	0.997	---	---	---	$\frac{250}{253} = 0.988$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{210}{216} = 0.972$

Tabela 8.21 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{130}{130} = 1.000$	$\frac{130}{271} = 0.480$	0.649	$\frac{26}{33} = 0.788$	$\frac{26}{63} = 0.413$	0.542	$\frac{68}{70} = 0.971$
ENSP	$\frac{101}{102} = 0.990$	$\frac{101}{266} = 0.380$	0.549	$\frac{57}{58} = 0.983$	$\frac{57}{87} = 0.655$	0.786	$\frac{85}{87} = 0.977$
PTFR	$\frac{131}{131} = 1.000$	$\frac{131}{282} = 0.601$	0.751	---	---	---	$\frac{62}{64} = 0.969$
PTSP	$\frac{184}{184} = 1.000$	$\frac{184}{283} = 0.465$	0.635	---	---	---	$\frac{234}{235} = 0.996$
PTEN	$\frac{98}{98} = 1.000$	$\frac{98}{262} = 0.374$	0.544	---	---	---	$\frac{86}{89} = 0.966$

Tabela 8.22 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{160}{160} = 1.000$	$\frac{160}{271} = 0.590$	0.742	$\frac{26}{32} = 0.813$	$\frac{26}{63} = 0.413$	0.548	$\frac{140}{145} = 0.966$
ENSP	$\frac{150}{151} = 0.993$	$\frac{150}{266} = 0.571$	0.725	$\frac{63}{64} = 0.984$	$\frac{63}{87} = 0.724$	0.834	$\frac{160}{168} = 0.952$
PTFR	$\frac{189}{189} = 1.000$	$\frac{189}{282} = 0.670$	0.802	---	---	---	$\frac{183}{187} = 0.979$
PTSP	$\frac{253}{253} = 1.000$	$\frac{253}{283} = 0.894$	0.944	---	---	---	$\frac{269}{277} = 0.971$
PTEN	$\frac{153}{153} = 1.000$	$\frac{153}{262} = 0.584$	0.737	---	---	---	$\frac{122}{129} = 0.946$

Tabela 8.23 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{203}{203} = 1.000$	$\frac{203}{271} = 0.749$	0.856	$\frac{32}{38} = 0.842$	$\frac{32}{63} = 0.508$	0.634	$\frac{193}{198} = 0.975$
ENSP	$\frac{163}{164} = 0.994$	$\frac{163}{266} = 0.613$	0.758	$\frac{64}{66} = 0.970$	$\frac{64}{87} = 0.736$	0.837	$\frac{170}{173} = 0.983$
PTFR	$\frac{216}{216} = 1.000$	$\frac{216}{282} = 0.766$	0.867	---	---	---	$\frac{141}{143} = 0.986$
PTSP	$\frac{273}{273} = 1.000$	$\frac{273}{283} = 0.611$	0.759	---	---	---	$\frac{244}{247} = 0.988$
PTEN	$\frac{168}{168} = 1.000$	$\frac{168}{262} = 0.641$	0.781	---	---	---	$\frac{175}{181} = 0.967$

Tabela 8.24 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 20% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{219}{219} = 1.000$	$\frac{219}{271} = 0.808$	0.894	$\frac{34}{41} = 0.829$	$\frac{34}{63} = 0.540$	0.654	$\frac{215}{220} = 0.977$
ENSP	$\frac{176}{176} = 1.000$	$\frac{176}{266} = 0.662$	0.797	$\frac{68}{71} = 0.958$	$\frac{68}{87} = 0.782$	0.861	$\frac{175}{179} = 0.978$
PTFR	$\frac{222}{222} = 1.000$	$\frac{222}{282} = 0.787$	0.881	---	---	---	$\frac{145}{148} = 0.980$
PTSP	$\frac{274}{274} = 1.000$	$\frac{274}{283} = 0.968$	0.984	---	---	---	$\frac{245}{248} = 0.988$
PTEN	$\frac{178}{178} = 1.000$	$\frac{178}{262} = 0.679$	0.809	---	---	---	$\frac{177}{182} = 0.973$

Tabela 8.25 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{174}{174} = 1.000$	$\frac{174}{271} = 0.642$	0.782	$\frac{45}{56} = 0.804$	$\frac{45}{63} = 0.714$	0.756	$\frac{131}{138} = 0.949$
ENSP	$\frac{127}{128} = 0.992$	$\frac{127}{266} = 0.477$	0.644	$\frac{79}{81} = 0.975$	$\frac{79}{87} = 0.908$	0.940	$\frac{183}{191} = 0.958$
PTFR	$\frac{183}{183} = 1.000$	$\frac{183}{282} = 0.649$	0.787	---	---	---	$\frac{135}{138} = 0.978$
PTSP	$\frac{265}{265} = 1.000$	$\frac{265}{283} = 0.936$	0.967	---	---	---	$\frac{265}{269} = 0.985$
PTEN	$\frac{135}{135} = 1.000$	$\frac{135}{262} = 0.515$	0.680	---	---	---	$\frac{145}{153} = 0.948$

Tabela 8.26 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{216}{216} = 1.000$	$\frac{216}{271} = 0.797$	0.887	$\frac{46}{60} = 0.767$	$\frac{46}{63} = 0.730$	0.748	$\frac{201}{207} = 0.971$
ENSP	$\frac{193}{194} = 0.995$	$\frac{193}{266} = 0.726$	0.839	$\frac{81}{84} = 0.964$	$\frac{81}{87} = 0.931$	0.947	$\frac{226}{235} = 0.961$
PTFR	$\frac{239}{239} = 1.000$	$\frac{239}{282} = 0.848$	0.918	---	---	---	$\frac{196}{199} = 0.985$
PTSP	$\frac{279}{279} = 1.000$	$\frac{279}{283} = 0.986$	0.992	---	---	---	$\frac{269}{275} = 0.978$
PTEN	$\frac{195}{195} = 1.000$	$\frac{195}{262} = 0.744$	0.853	---	---	---	$\frac{186}{196} = 0.949$

Tabela 8.27 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{235}{235} = 1.000$	$\frac{235}{271} = 0.867$	0.929	$\frac{49}{64} = 0.766$	$\frac{49}{63} = 0.778$	0.772	$\frac{241}{250} = 0.964$
ENSP	$\frac{199}{200} = 0.995$	$\frac{199}{266} = 0.748$	0.854	$\frac{81}{85} = 0.953$	$\frac{81}{87} = 0.931$	0.942	$\frac{238}{248} = 0.960$
PTFR	$\frac{246}{246} = 1.000$	$\frac{246}{282} = 0.872$	0.932	---	---	---	$\frac{233}{237} = 0.983$
PTSP	$\frac{281}{281} = 1.000$	$\frac{281}{283} = 0.993$	0.996	---	---	---	$\frac{270}{275} = 0.982$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{226}{236} = 0.958$

Tabela 8.28 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 70% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{243}{243} = 1.000$	$\frac{243}{271} = 0.897$	0.956	$\frac{48}{65} = 0.738$	$\frac{48}{63} = 0.762$	0.750	$\frac{252}{261} = 0.966$
ENSP	$\frac{204}{205} = 0.995$	$\frac{204}{266} = 0.767$	0.866	$\frac{82}{88} = 0.932$	$\frac{82}{87} = 0.943$	0.937	$\frac{235}{245} = 0.959$
PTFR	$\frac{251}{251} = 1.000$	$\frac{251}{282} = 0.890$	0.942	---	---	---	$\frac{241}{246} = 0.980$
PTSP	$\frac{282}{282} = 1.000$	$\frac{282}{283} = 0.996$	0.998	---	---	---	$\frac{270}{275} = 0.982$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{235}{243} = 0.967$

Tabela 8.29 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{131}{131} = 1.000$	$\frac{131}{271} = 0.483$	0.651	$\frac{29}{38} = 0.763$	$\frac{29}{63} = 0.460$	0.860	$\frac{74}{78} = 0.949$
ENSP	$\frac{102}{103} = 0.990$	$\frac{102}{266} = 0.383$	0.552	$\frac{63}{64} = 0.984$	$\frac{63}{87} = 0.724$	0.834	$\frac{104}{107} = 0.972$
PTFR	$\frac{138}{138} = 1.000$	$\frac{138}{282} = 0.489$	0.658	---	---	---	$\frac{83}{85} = 0.976$
PTSP	$\frac{222}{222} = 1.000$	$\frac{222}{283} = 0.784$	0.879	---	---	---	$\frac{247}{248} = 0.996$
PTEN	$\frac{98}{98} = 1.000$	$\frac{98}{262} = 0.374$	0.544	---	---	---	$\frac{97}{102} = 0.951$

Tabela 8.30 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{161}{161} = 1.000$	$\frac{161}{271} = 0.594$	0.745	$\frac{28}{37} = 0.757$	$\frac{28}{63} = 0.444$	0.560	$\frac{136}{141} = 0.965$
ENSP	$\frac{151}{152} = 0.993$	$\frac{151}{266} = 0.568$	0.723	$\frac{71}{73} = 0.973$	$\frac{71}{87} = 0.816$	0.888	$\frac{177}{185} = 0.957$
PTFR	$\frac{198}{198} = 1.000$	$\frac{198}{282} = 0.702$	0.825	---	---	---	$\frac{168}{171} = 0.982$
PTSP	$\frac{253}{253} = 1.000$	$\frac{253}{283} = 0.894$	0.944	---	---	---	$\frac{258}{263} = 0.981$
PTEN	$\frac{153}{153} = 1.000$	$\frac{153}{262} = 0.584$	0.737	---	---	---	$\frac{137}{143} = 0.958$

Tabela 8.31 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{204}{204} = 1.000$	$\frac{204}{271} = 0.859$	0.924	$\frac{35}{47} = 0.745$	$\frac{35}{63} = 0.556$	0.637	$\frac{202}{210} = 0.962$
ENSP	$\frac{164}{165} = 0.994$	$\frac{164}{266} = 0.617$	0.761	$\frac{73}{76} = 0.961$	$\frac{73}{87} = 0.839$	0.896	$\frac{203}{211} = 0.962$
PTFR	$\frac{224}{224} = 1.000$	$\frac{224}{282} = 0.794$	0.885	---	---	---	$\frac{216}{219} = 0.986$
PTSP	$\frac{273}{273} = 1.000$	$\frac{273}{283} = 0.965$	0.982	---	---	---	$\frac{264}{269} = 0.981$
PTEN	$\frac{168}{168} = 1.000$	$\frac{168}{262} = 0.641$	0.781	---	---	---	$\frac{194}{203} = 0.956$

Tabela 8.32 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 80% entre os vectores de frequências dos textos. No método da proporcionalidade dos tamanhos utilizou-se uma taxa de tolerância de 40% para todos os corpus.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{220}{220} = 1.000$	$\frac{220}{271} = 0.812$	0.896	$\frac{38}{51} = 0.745$	$\frac{38}{63} = 0.603$	0.667	$\frac{226}{233} = 0.970$
ENSP	$\frac{177}{177} = 1.000$	$\frac{177}{266} = 0.665$	0.799	$\frac{77}{81} = 0.951$	$\frac{77}{87} = 0.885$	0.917	$\frac{203}{212} = 0.958$
PTFR	$\frac{230}{230} = 1.000$	$\frac{230}{282} = 0.816$	0.899	---	---	---	$\frac{220}{224} = 0.982$
PTSP	$\frac{274}{274} = 1.000$	$\frac{274}{283} = 0.968$	0.983	---	---	---	$\frac{265}{270} = 0.981$
PTEN	$\frac{178}{178} = 1.000$	$\frac{178}{262} = 0.679$	0.809	---	---	---	$\frac{201}{208} = 0.956$

Aplicação dos métodos baseados nos nomes dos ficheiros e em cognatos.

Tabela 8.33 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 70% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{167}{167} = 1.000$	$\frac{167}{271} = 0.616$	0.762	$\frac{49}{69} = 0.710$	$\frac{49}{63} = 0.778$	0.742	$\frac{138}{144} = 0.958$
ENSP	$\frac{185}{186} = 0.995$	$\frac{185}{266} = 0.695$	0.818	$\frac{82}{85} = 0.965$	$\frac{82}{87} = 0.943$	0.954	$\frac{178}{186} = 0.957$
PTFR	$\frac{197}{197} = 1.000$	$\frac{197}{282} = 0.699$	0.823	---	---	---	$\frac{153}{157} = 0.975$
PTSP	$\frac{271}{271} = 1.000$	$\frac{271}{283} = 0.958$	0.979	---	---	---	$\frac{273}{283} = 0.965$
PTEN	$\frac{168}{168} = 1.000$	$\frac{168}{262} = 0.641$	0.781	---	---	---	$\frac{133}{138} = 0.964$

Tabela 8.34 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 70% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{213}{213} = 1.000$	$\frac{213}{271} = 0.786$	0.880	$\frac{55}{72} = 0.764$	$\frac{55}{63} = 0.873$	0.814	$\frac{199}{207} = 0.961$
ENSP	$\frac{204}{205} = 0.995$	$\frac{204}{266} = 0.767$	0.866	$\frac{85}{93} = 0.914$	$\frac{85}{87} = 0.977$	0.944	$\frac{228}{238} = 0.958$
PTFR	$\frac{226}{226} = 1.000$	$\frac{226}{282} = 0.801$	0.890	---	---	---	$\frac{217}{223} = 0.973$
PTSP	$\frac{273}{273} = 1.000$	$\frac{273}{283} = 0.965$	0.982	---	---	---	$\frac{275}{285} = 0.965$
PTEN	$\frac{190}{190} = 1.000$	$\frac{190}{262} = 0.725$	0.841	---	---	---	$\frac{171}{179} = 0.955$

Tabela 8.35 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 70% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{240}{240} = 1.000$	$\frac{240}{271} = 0.886$	0.940	$\frac{52}{73} = 0.712$	$\frac{52}{63} = 0.825$	0.764	$\frac{248}{258} = 0.961$
ENSP	$\frac{206}{207} = 0.995$	$\frac{206}{266} = 0.774$	0.871	$\frac{85}{91} = 0.934$	$\frac{85}{87} = 0.977$	0.955	$\frac{242}{253} = 0.957$
PTFR	$\frac{243}{243} = 1.000$	$\frac{243}{282} = 0.862$	0.926	---	---	---	$\frac{250}{255} = 0.980$
PTSP	$\frac{278}{278} = 1.000$	$\frac{278}{283} = 0.982$	0.991	---	---	---	$\frac{277}{286} = 0.966$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{239}{250} = 0.944$

Tabela 8.36 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 70% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{244}{244} = 1.000$	$\frac{244}{271} = 0.900$	0.947	$\frac{48}{68} = 0.706$	$\frac{48}{63} = 0.762$	0.733	$\frac{257}{269} = 0.955$
ENSP	$\frac{204}{205} = 0.995$	$\frac{204}{266} = 0.767$	0.866	$\frac{82}{90} = 0.911$	$\frac{82}{87} = 0.943$	0.923	$\frac{238}{249} = 0.956$
PTFR	$\frac{253}{253} = 1.000$	$\frac{253}{282} = 0.897$	0.946	---	---	---	$\frac{255}{261} = 0.977$
PTSP	$\frac{282}{282} = 1.000$	$\frac{282}{283} = 0.996$	0.998	---	---	---	$\frac{277}{286} = 0.969$
PTEN	$\frac{203}{203} = 1.000$	$\frac{203}{262} = 0.775$	0.873	---	---	---	$\frac{239}{249} = 0.960$

Tabela 8.37 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 60% entre palavras e de 80% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{141}{141} = 1.000$	$\frac{141}{271} = 0.520$	0.684	$\frac{37}{47} = 0.787$	$\frac{37}{63} = 0.587$	0.672	$\frac{71}{75} = 0.947$
ENSP	$\frac{139}{140} = 0.993$	$\frac{139}{266} = 0.523$	0.685	$\frac{64}{64} = 1.000$	$\frac{64}{87} = 0.736$	0.848	$\frac{99}{103} = 0.961$
PTFR	$\frac{157}{157} = 1.000$	$\frac{157}{282} = 0.557$	0.715	---	---	---	$\frac{93}{95} = 0.979$
PTSP	$\frac{229}{229} = 1.000$	$\frac{229}{283} = 0.809$	0.894	---	---	---	$\frac{254}{262} = 0.969$
PTEN	$\frac{123}{123} = 1.000$	$\frac{123}{262} = 0.469$	0.639	---	---	---	$\frac{86}{88} = 0.977$

Tabela 8.38 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 70% entre palavras e de 80% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{166}{166} = 1.000$	$\frac{166}{271} = 0.613$	0.760	$\frac{44}{55} = 0.800$	$\frac{44}{63} = 0.698$	0.746	$\frac{141}{146} = 0.966$
ENSP	$\frac{163}{164} = 0.994$	$\frac{163}{266} = 0.613$	0.758	$\frac{76}{79} = 0.962$	$\frac{76}{87} = 0.874$	0.916	$\frac{167}{175} = 0.954$
PTFR	$\frac{197}{197} = 1.000$	$\frac{297}{282} = 0.699$	0.822	---	---	---	$\frac{183}{187} = 0.979$
PTSP	$\frac{254}{254} = 1.000$	$\frac{254}{283} = 0.898$	0.946	---	---	---	$\frac{269}{277} = 0.971$
PTEN	$\frac{148}{148} = 1.000$	$\frac{148}{262} = 0.565$	0.722	---	---	---	$\frac{122}{129} = 0.946$

Tabela 8.39 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 80% entre palavras e de 80% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{166}{166} = 1.000$	$\frac{166}{271} = 0.613$	0.760	$\frac{39}{53} = 0.736$	$\frac{39}{63} = 0.619$	0.672	$\frac{214}{222} = 0.964$
ENSP	$\frac{163}{164} = 0.994$	$\frac{163}{266} = 0.613$	0.758	$\frac{78}{81} = 0.963$	$\frac{78}{87} = 0.897$	0.929	$\frac{209}{217} = 0.963$
PTFR	$\frac{197}{197} = 1.000$	$\frac{297}{282} = 0.699$	0.822	---	---	---	$\frac{222}{226} = 0.982$
PTSP	$\frac{268}{268} = 1.000$	$\frac{268}{283} = 0.947$	0.973	---	---	---	$\frac{273}{281} = 0.972$
PTEN	$\frac{148}{148} = 1.000$	$\frac{148}{262} = 0.565$	0.722	---	---	---	$\frac{197}{208} = 0.947$

Tabela 8.40 - Tabela de resultados da aplicação do método baseado em nomes de ficheiros e na proporcionalidade dos tamanhos. Para os corpus Euro-Lex e do Vaticano utilizou-se no máximo 2 edições e no corpus das UN usou-se no máximo 6 edições. No método baseado em cognatos usou-se semelhança de 100% entre palavras e de 80% entre os vectores de frequências dos textos.

Línguas usadas	Corpus						
	Euro-Lex			UN			Vaticano
	Precisão	Recall	F-measure	Precisão	Recall	F-measure	Precisão
ENFR	$\frac{220}{220} = 1.000$	$\frac{220}{271} = 0.812$	0.896	$\frac{38}{53} = 0.717$	$\frac{38}{63} = 0.603$	0.655	$\frac{230}{240} = 0.958$
ENSP	$\frac{177}{177} = 1.000$	$\frac{177}{266} = 0.665$	0.799	$\frac{77}{83} = 0.928$	$\frac{77}{87} = 0.885$	0.906	$\frac{205}{215} = 0.953$
PTFR	$\frac{231}{231} = 1.000$	$\frac{231}{282} = 0.819$	0.900	---	---	---	$\frac{232}{237} = 0.979$
PTSP	$\frac{274}{274} = 1.000$	$\frac{274}{283} = 0.968$	0.984	---	---	---	$\frac{272}{281} = 0.968$
PTEN	$\frac{178}{178} = 1.000$	$\frac{178}{262} = 0.679$	0.809	---	---	---	$\frac{205}{214} = 0.958$