Extracção de corpora paralelo a partir da web: construção e disponibilização

José João Almeida jj@di.uminho.pt
José Alves de Castro jac@natura.di.uminho.pt
Alberto M. Simões albie@alfarrabio.di.uminho.pt

Departamento de Informática
Universidade do Minho

Resumo

Ao longo deste documento descrever-se-á um conjunto de ferramentas construídas para extracção automática de recursos bilingues a partir da Web, a partir de um *site* específico, a partir de um sistema de ficheiros contendo alguns textos que sejam traduções de outros, ou ainda a partir de memórias de tradução.

Neste trabalho apresenta-se todo o processo de construção de corpora paralelo desde os algoritmos de minagem dos dados (data mining) até à construção de vários tipos de recursos bilingues incluindo a construção automática de corpus paralelos pesquisáveis na Internet.

1 Introdução

É universalmente reconhecido que os recursos multilingue são cruciais para várias áreas ligadas a estudos da língua, como sejam a área da tradução, a extracção de terminologia ou a criação de dicionários. No entanto estes recursos são normalmente difíceis de obter e organizar.

O projecto TerminUM, onde este artigo se enquadra, tem como objectivo a recolha, tratamento e disponibilização de recursos bilingues na Internet. Para isso foi definido um grafo que permita obter recursos paralelos a partir de documentos extraídos na Internet. A figura 1 mostra os vários processos intervenientes no TerminUM. Nesta imagem "*fich*" corresponde a ficheiros, "*seg*" a segmentos (ou frases) e "*p*" a palavras.

As secções seguintes resumem cada um dos processos apresentados no diagrama. Este artigo irá centrar-se especialmente no processo de "tratamento de corpora" e "alinhamento à frase", apresentados na secção 3. No entanto, para que o artigo seja auto-contido, iremos apresentar na secção 2 o processo de extracção e validação de corpora da Web.

1.1 Recolha

Para ser possível a recolha de textos paralelos na Internet torna-se necessário encontrá-lo. Existem vários métodos para realizar esta operação. Um dos métodos, apresentado em [10, 9], utiliza um motor de pesquisa na Internet para encontrar páginas em línguas diferentes que se relacionem de forma particular.

Um outro método, bastante mais leve consiste em analisar um conjunto de endereços de Internet (URL's) e usar heurísticas sobre os nomes dos ficheiros para descobrir relações entre eles. De facto, é habitual que na construção de *sites* em várias línguas se atribuam nomes elucidativos às pastas ou ficheiros em causa.

Finalmente, podemos obter textos paralelos de um *site* que conhecemos e que sabemos conter textos paralelos. Nesse caso, podemos ir buscar todo o *site* e utilizar métodos de comparação entre os ficheiros para encontrar de forma automática a relação entre ficheiros.

No final deste processo acabamos com pares de ficheiros candidatos que vamos validar e confirmar se são traduções das línguas que andamos à procura.

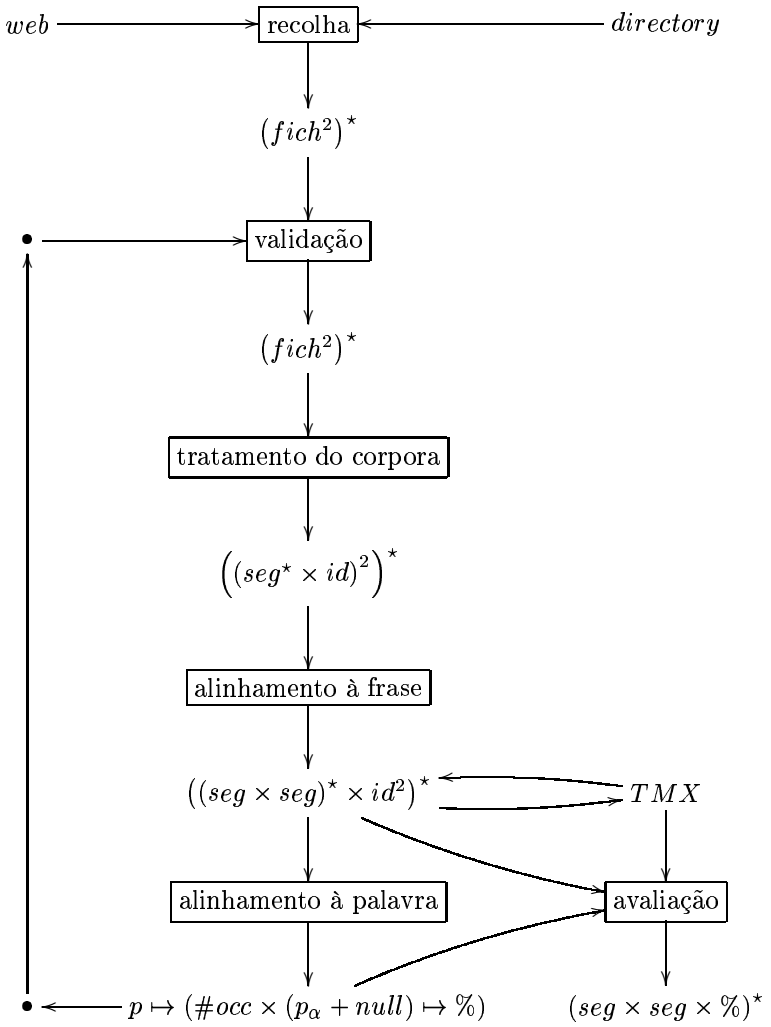


Figura 1: Ciclo de vida e diagrama de construção TerminUM

1.2 Validação

Aos pares candidatos é aplicado um processo de validação. Os métodos automáticos de recolha não nos garantem um conjunto de propriedades destes textos. O processo de validação verifica:

- se as línguas dos ficheiros em causa estão correctas;
- se os ficheiros têm ambos o mesmo tipo, e se o tipo é reconhecido;
- se nenhum dos ficheiros é demasiado grande em relação ao outro;
- se os nomes dos ficheiros são semelhantes;
- se o conteúdo não textual é parecido entre os dois ficheiros (como a pontuação, números, comandos);

Aos pares que passarem este processo de validação iremos chamar ficheiros ou textos paralelos.

1.3 Tratamento do corpora

Dos ficheiros recolhidos é necessário remover a marcação específica do formato de ficheiro, limpando tudo o que não seja interessante para a criação do corpus. Este processo é dependente do formato em que o ficheiro se encontra.

Depois de limpo, o texto é segmentado e etiquetado num formato XML que denominamos por PML, em que as etiquetas só assinalam os segmentos obtidos.

Posteriormente, cada ficheiro PML é inserido no *Corpus Workbench* (CWB)[7] para permitir eficiência em consultas subsequentes. Este processo pode ser complementado com a lematização dos corpora utilizando um analisador morfológico.

1.4 Alinhamento à frase

Depois de criados, estes corpora podem ser alinhados à frase. Para isso é utilizado o *easyalign*, ferramenta que faz parte do CWB. Estes corpora alinhados podem ser consultados como corpora independentes ou de forma síncrona.

Existe, também, a possibilidade de converter corpora paralelos em ficheiros Translation Memory Exchange (TMX) e vice-versa, para permitir o seu uso por terceiros, como sejam tradutores.

1.5 Alinhamento à palavra

Depois de alinhados à frase, o corpora pode ser alinhado à palavra. Para isso, estamos a utilizar um conjunto de ferramentas denominado NATools (Natura Alignment Tools). Estas ferramentas são baseadas no alinhador desenvolvido por Hiemstra [5, 4] no âmbito do projecto Agenda 21 da Comunidade Europeia.

Este alinhamento permite obter dicionários entre as duas línguas em questão, em que a cada palavra de uma das línguas é associado um conjunto de possíveis traduções e suas respectivas probabilidades de tradução.

1.6 Avaliação

Estes mesmos dicionários podem ser usados posteriormente para enriquecer o processo de validação de traduções. De facto, uma medida de avaliação de traduções pode ser obtida calculando, para cada palavra de uma frase se uma das suas possíveis traduções se encontra na outra frase. A média pesada destes valores dá um óptimo estimador da correcção da tradução.

Este estimador está a ser usado para filtrar memórias de tradução de forma a obter apenas as unidades que têm qualidade superior a determinado valor.

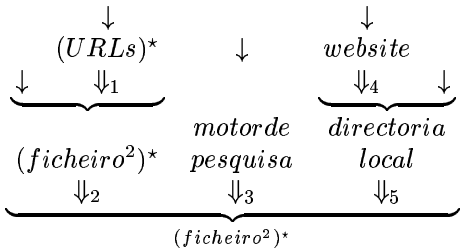
2 Recolha de corpora

Este processo consiste em procurar pares de ficheiros passíveis de serem traduções e em, posteriormente, analisar cada um desses pares com o intuito de obter apenas os que nos interessam.

2.1 Extracção

A recolha de corpora tem o seu início, como já foi dito, na identificação de pares de ficheiros que serão, possivelmente, traduções paralelas.

O diagrama que se segue ilustra os vários processos utilizados para a recolha destes pares.



As setas simples simbolizam os pontos de entrada. As setas duplas representam os processos descritos nas cinco subsecções que se seguem.

2.1.1 Lista de URLs

Dada uma lista de URLs, esta é processada de forma a constituir blocos de URLs semelhantes.

Começamos por remover os identificadores de língua em cada URL [1]. Depois, os URLs são associados entre si de acordo com o resultado desta operação.

Assim, URLs como

`http://www.ex.pt/index_pt.html` e `http://www.ex.pt/index_en.html`

ficam associados à sua parte comum:

`http://www.ex.pt/index_.html`

De seguida, cada bloco contendo mais do que uma entrada é processado para que dele sejam extraídos possíveis pares correspondentes às línguas procuradas. Os pares encontrados são descarregados e o resultado é uma lista de pares de ficheiros.

2.1.2 Lista de pares de ficheiros

À semelhança do processo descrito na secção anterior, uma lista de pares de ficheiros (ou URLs) pode ser criada manualmente. Esta lista consiste de linhas de pares onde os elementos se encontram separados por um [TAB].

2.1.3 Internet

O algoritmo utilizado para extracção de pares candidatos a partir da Internet é baseado no de Philip Resnik [9, 10]. Submetendo uma procura a um motor de pesquisa, procuramos ficheiros com ligações a duas páginas, onde cada uma dessas ligações menciona uma das línguas procuradas.

Cada uma das páginas retornadas pelo motor de pesquisa é descarregada e processada localmente, de modo a dela extrair todos os pares de URLs passíveis de serem traduções.

Deste processo resulta uma lista de pares de ficheiros.

2.1.4 Website

Dado o URL de um *website*, este é descarregado com a ferramenta GNU Wget [8]. A directoria resultante é processada como descrito na secção 2.1.5.

2.1.5 Directoria

Dada um directoria, a sua árvore de subdirectorias é criada e os seguintes casos para pares de ficheiros são analisados:

- estão na mesma directoria, até uma certa profundidade

[1] para inglês, por exemplo, estes identificadores seriam `english`, `eng` e `en`, conforme [6]

- estão à mesma profundidade, com uma directoria mãe comum, a uma certa distância;
- um está numa directoria mãe da do outro, a uma certa distância;
- estão a diferentes profundidades e têm uma directoria pai comum, a uma certa distância do mais afastado;

2.2 Validação

Após terem sido seleccionados, os pares candidatos passam por um processo de validação no qual podem ser eliminados se surgir motivo para tal. Os vários resultados encontrados vão fazer parte do cálculo final que nos vai indicar a probabilidade de que os ficheiros sejam traduções um do outro.

2.2.1 Identificação de língua

Foi implementado um módulo de identificação de língua; os pares cujos ficheiros não correspondem às línguas pretendidas são eliminados.

2.2.2 Validação de tipos

Foi implementado um módulo de identificação de tipos de ficheiros. Os pares cujos tipos sejam diferentes ou desconhecidos são automaticamente descartados.

2.2.3 Comparação de tamanhos

A comparação de tamanhos consiste em verificar que percentagem do maior ficheiro[2] é ocupada pelo menor. Caso este valor seja demasiado baixo, o par é descartado.

2.2.4 Semelhança de nomes e pontuação

Os nomes dos ficheiros são normalizados (i.e., desprovidos de caracteres não alfanuméricos e minuscularizados) e é depois utilizado um algoritmo de aproximação de texto para calcular a sua semelhança.

O mesmo algoritmo é aplicado aos segmentos de pontuação final existentes nos textos de ambos os ficheiros.

2.2.5 Conteúdos não textuais

Por conteúdos não textuais entendem-se imagens, ligações, etc. O método passa por identificar todos estes conteúdos e calcular a percentagem de conteúdos semelhantes sobre a sua totalidade.

2.2.6 Valor final

Com os resultados das comparações de tamanhos, nomes de ficheiros, pontuação e conteúdos não textuais é efectuado um cálculo final; este consiste numa média pesada cujos parâmetros são configuráveis e dependentes do tipo de ficheiro em questão.

Caso o resultado encontrado seja superior a um certo valor definido à partida (que uma vez mais é dependente do tipo de ficheiro), o par é considerado uma tradução paralela e armazenado para criação de corpora.

[2] para este efeito, o ficheiro é primeiramente desprovido de conteúdo não textual

3 Criação e disponibilização de corpora

Nesta secção descreve-se algumas das sub-tarefas na construção de corpora bilingues consultáveis via *web* e na construção de versões dos corpora em formato TMX.

Mais uma vez, cada processo de transformação pode ser feito de diversos modos, sendo no entanto fornecido pelo projecto TerminUM um comando que o execute sem qualquer acção interactiva, de modo a permitir composição dos vários comandos num único, e de modo a permitir o tratamento de grandes quantidades de informação. Em vários dos processos recorreu-se às ferramentas do CWB normalmente integradas com alguns programas que fazem as necessárias adaptações de formatos.

Na sequência do anteriormente descrito, vamos descrever um conjunto de tarefas que:

- fazem a limpeza e normalização dos ficheiros;
- constroem informação adicional: juntar a cada palavra os possíveis lemas e `POS`;
- alinhamento à frase;
- tradução de e para memórias de tradução (TMX);
- disponibilização para consulta via *cgi-web*[3, 12, 11];

3.1 Conversão de ficheiros HTML em PML

No processo de normalização e limpeza dos ficheiros de entrada, há que retirar certa informação contida nos textos HTML, dividir os mesmos em frases (segmentação) e convertê-los para um formato comum, usado nos passos seguinte: PML.

O formato PML não é mais que simples texto com marcas XML para separar diferentes ficheiros — `<f id="num" name="nome">...</f>` — e para separar períodos — `<p>...</p>`. Embora estejamos a usar uma etiqueta existente em HTML, não incluímos só parágrafos mas um conjunto de unidade mais vasto como sejam títulos, legendas, partes de tabelas e outros sub-elementos por vezes designados de unidades de tradução.

O atributo "`id`" do elemento "`f`" é usado para sincronização no processo de alinhamento à frase.

Esta separação está a ser baseada numa tabela em que o conjunto das etiquetas HTML está a ser dividido em:

- etiquetas a remover, preservando o seu conteúdo (Ex. as etiquetas `html, em, i, u, body`);
- etiquetas a remover bem como o respectivo conteúdo (Ex. as etiquetas `frameset, head, meta, script`);
- etiquetas separadoras de unidade de tradução (Ex. as etiquetas `h1, li, p, blockquote`);
- etiquetas a conservar.

O conjunto de etiquetas em cada classe pode ser alterado usando uma série de opções disponíveis.

Após feita a divisão em unidades de tradução de acordo com a tabela anterior, é ainda feita a segmentação de cada unidade, usando um segmentador tradicional (incluído no módulo Lingua::PT::pln).

3.2 Lematização de ficheiros PML e sua indexação com CWB

Durante esta fase do processo, é feito o cálculo dos possíveis lemas e *part-of-speech* (POS) de cada palavra e conversão para o formato aceite pelo CWB.

O cálculo de lemas e POS está a usar a biblioteca perl do Jspell [1, 13] e produz lemas e POS ambíguos: cada palavra pode dar origem a vários lemas e vários POS. As várias hipóteses

encontradas estão a ser formatadas de acordo com o formato esperado pelos atributos ambíguos em CWB.

Seguidamente é determinado qual o conjunto de etiquetas usadas neste corpora e é construído um corpus CWB, usando os seus indexadores com um conjunto de opções determinadas automaticamente.

3.3 Alinhamento à frase

O alinhamento à frase é feito usando o *easyalign* — um alinhador à frase que faz parte do CWB. Este alinhador funciona sem qualquer tipo de interactividade — o que é crucial neste processo, devido aos tamanhos envolvidos.

Havendo (pelo menos) duas línguas em análise são feitos alinhamentos em ambos os sentidos.

3.4 Conversão de e para memórias de tradução (TMX)

A conversão para TMX foi ditada pela vontade de realizar intercâmbios com outras comunidades como seja a comunidade de tradução. Deste modo permitimos que:

- outras pessoas possam produzir recursos a serem utilizados no TerminUM;
- certas ferramentas ligadas à tradução possam ser usadas sobre os corpora criados no projecto;
- um conjunto mais vasto de pessoas possa validar implicitamente o trabalho realizado.

A conversão para TMX está a ser realizada com uma mistura de navegação no corpus paralelo CWB com uma geração do texto XML ligado à TMX.

A conversão TMX para CWB está a usar um reconhecedor de XML e a fazer uma transformação estrutural para PML com atributos de alinhamento, sendo depois reutilizados os conversores descritos anteriormente.

3.5 Criação de corpora consultáveis via *web*

Uma das questões a que se atribuiu grande importância foi à construção de mecanismos de consulta remota. Deste modo, estamos a construir recursos que são úteis à comunidade e ao mesmo tempo que sejam implicitamente testados.

Os corpora ficam consultáveis via *web* à custa de um programa (CGI) que:

- começa por determinar quais os corpora disponíveis no sistema;
- determina quais deles são paralelos;
- constrói uma página HTML que permite escolher o corpus a usar, bem como qual o tamanho do contexto a apresentar.

3.6 Sistema geral em funcionamento

Se dispusermos de um ficheiro `F.pairs` contendo em cada linha um par de nomes de ficheiros que sejam a tradução um do outro:

```
f1_pt.html    f1_eng.html
f2_pt.html    f2_eng.html
...
```

e executarmos o comando

```
mkterminum F.pairs
```

todos os recursos referidos nesta secção ficam calculados, incluindo o facto de que o corpus fica consultável na Internet.

Para além deste comando geral, todas as etapas anteriores estão acessíveis com comandos individuais, permitindo realizar as respectivas tarefas de modo mais controlado.

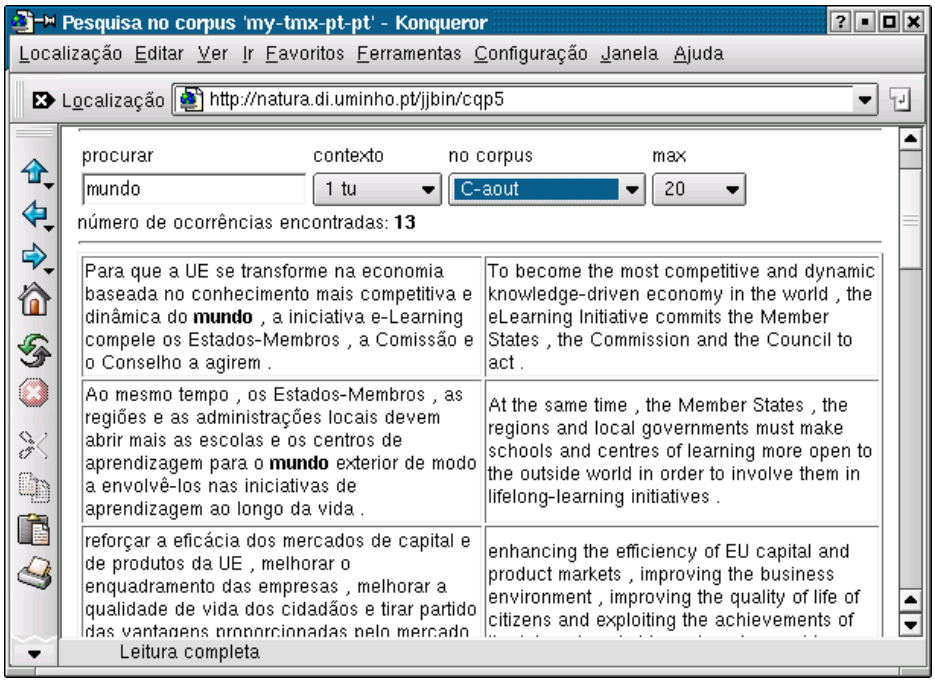


Figura 2: Pesquisa em corpora paralelos usando a CGI de pesquisa

4 Avaliação de resultados

Foram realizados dois tipos de testes preliminares, cujos resultados se apresentam a seguir.

4.1 Primeira experiência - Web Site

Como primeira experiência, descarregamos uma pequena secção de um *site*, composta por 34 ficheiros. Os mesmos foram analisados manualmente para concluirmos que entre eles existiam 12 pares que eram traduções.

A identificação de língua eliminou cerca de 65% das comparações necessárias (de notar que os ficheiros identificados erróneamente se tratavam de páginas de *frames*, que devido ao seu tamanho seriam também descartadas nas primeiras fases de validação).

O programa identificou 13 pares como sendo traduções, dos quais 11 estavam correctos (uma precisão de 85%). 11 dos 12 pares originais foram identificados (uma cobertura de 92%).

ficheiros	34
ficheiros em Português	15
ficheiros em Inglês	19
pares tradução	12
identificados como Português	12 (80%)
identificados como Inglês	16 (84%)
comparações efectuadas	192
comparações possíveis	561
pares encontrados	13
pares correctos	11
precisão	11/13 = 85%
cobertura	11/12 = 92%
tempo	3.17 segundos
tempo por comparação	0.02 segundos

O teste completo demorou 3.17 segundos[3] (incluindo a identificação de língua), o que dá 0.2 segundos por comparação.

Com estes dados, podemos estimar que o programa consiga, em cerca de um minuto, efectuar 3000 comparações.

4.2 Segunda Experiência - URLs

Para a segunda experiência, um robô recolheu uma lista de 18 217 452 URLs. Destes, foram conservadas para utilização futura 10 226 564, que se encontravam num pequeno conjunto de servidores[4].

Os restantes 7 990 888 de URLs foram processados de modo a construir blocos. O resultado foi de 66 116. Destes blocos foi possível extrair um conjunto de 4 326 pares candidatos, operação que demorou 112 minutos[5].

. A razão deste aparente decréscimo prende-se com o facto de estarmos a trabalhar com as línguas Português e Inglês; no entanto, os blocos construidos contêm muito mais informação que pode ser utilizada para encontrar novos pares noutras línguas.

URLs	7 990 888
blocos criados	66 116
pares candidatos (PT-EN)	4 326
tempo necessário para criar os blocos	112 minutos

Dos 4 326 pares encontrados, seleccionamos uma pequena parte, composta por 756 deles, para processar. Destes, 496 foram descarregados com sucesso, de entre os quais 253 foram identificados como sendo traduções.

pares examinados	756
pares descarregados com sucesso	496
pares identificados como traduções	253

5 Conclusões e trabalho futuro

Os métodos descritos neste documento permitem obter de forma eficiente corpora bilingue bastando para isso decidir de que forma se pretendem extrair da Internet. O seu tratamento, embora de forma automática, está a produzir corpora paralelo de qualidade razoável e a disponibilizá-lo para pesquisa.

Actualmente estamos a melhorar o alinhador à palavra para aumentar a sua eficiência temporal e para reconhecimento de termos compostos. Posteriormente será desenvolvida uma aplicação para tradução de conjuntos de palavras.

As ferramentas referidas neste documento estão disponíveis a partir das páginas do projecto Natura `http://natura.di.uminho.pt/terminum`.

Referências

[1] J.J. Almeida and Ulisses Pinto. Jspell – um módulo para análise léxica genérica de linguagem natural. In *Actas do Congresso da Associação Portuguesa de Linguística, Évora*, 1994.

[2] José João Almeida, Alberto Manuel Simões, and José Alves Castro. Grabbing parallel corpora from the web. Number 29, pages 13–20. Sociedade Española para el Procesamiento del Lenguaje Natural, Sep. 2002.

[3] este teste foi realizado numa máquina Pentium IV 1.5 Ghz com 256 Mb de RAM, em Linux

[4] estes URLs não considerados para este teste por várias razões, mas estima-se que apresentem resultados muito superiores aos dos restantes

[5] este teste foi realizado numa máquina AMD K6 333Mhz com 256 Mb de RAM, em Linux

[3] Ana Frankenberg-Garcia and Diana Santos. Apresentando o compara, um corpus português-inglês na web. In *Cadernos de Tradução*.

[4] Djoerd Hiemstra. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional lexicon from a parallel corpus. Technical report, University of Twente, Parlevink Group, 1998.

[5] Djoerd Hiemstra. Using statistical methods to create a bilingual dictionary. Master's thesis, Department of Computer Science, University of Twente, August 1996.

[6] ISO 639. *Language Codes*. International Organization for Standardization, 1992.

[7] Oliver Christ & Bruno M. Schulze & Anja Hofmann & Esther König. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stutgart, March 8, 1999 (CQP V2.2).

[8] Hrvoje Niksic. *GNU Wget Manual*. 2001.

[9] Philip Resnik. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *D. Farwell, L. Gerber, and E. Hovy (eds.), Machine Translation and the Information Soup (AMTA-98)*, 1998. Lecture Notes in Artificial Intelligence 1529, Springer.

[10] Philip Resnik. Mining the web for bilingual text. In *37th Annual Meeting of the ACL'99*, 1999. College Park, Maryland.

[11] Paulo Alexandre Rocha and Diana Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, pages 131–140, November 2000.

[12] Diana Santos and Eckhard Bick. Providing internet access to portuguese corpora: the AC/DC project. In Maria Gavrilidou et al, editor, *Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 205–210, June 2000.

[13] Alberto Manuel Simões and José João Almeida. `jspell.pm` — um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas da Associaçáo Portuguêsa de Linguística*, 2001.