Cálculo de frequências para entradas de dicionários através do uso conjunto de analisadores morfológicos, *taggers* e corpora.

Paulo Alexandre Rocha
Alberto Manuel Simões
José João Almeida
Departamento de Informática
Universidade do Minho

Apresentamos neste documento uma possível abordagem à extracção de frequências de palavras a partir de corpora, baseada numa utilização cooperativa de várias ferramentas de Processamento de Linguagem Natural (PLN).

Pretendemos, além da determinação das frequências concretas, disponibilizar um conjunto de ferramentas *open-source* que possibilitem o cálculo dessas frequências de forma automática, aplicáveis a qualquer corpus, produzindo resultados utilizáveis em vários contextos, como sejam estudos contrastivos.

Introdução

O problema do cálculo de frequências de palavras normalizadas (lemas) para entradas de dicionários é importante em diversos domínios, nomeadamente o do ensino e aprendizagem da língua. Este problema, embora já abordado noutras línguas (Kilgarriff 1996), tem sido no entanto por vezes negligenciado quanto ao português[1], por ser de difícil automatização, apresentando vários problemas conceptuais– nenhum dicionário de português do nosso conhecimento contem este tipo de informação. Uma parte substancial das palavras gráficas (tipos) ocorrentes em textos portugueses é ambigua: não pode ser atribuída automaticamente a um determinado lema. Por exemplo, *lava* pode ser um substantivo, ou uma ocorrência do verbo *lavar*. Embora intuitivamente se presuma que a maior parte das ocorrências da palavra sejam relativas à forma verbal, não se pode verificar tal hipótese simplesmente contabilizando o número de existências de palavras em corpora. Há necessidade de examinar caso a caso um número significativo de ocorrências de tal palavra.

[1] ver no entanto o trabalho realizado pelo grupo de Processamento de Linguagem Natural do INESC (Medeiros et al. 1992) e o projecto Léxico Multifuncional Computorizado do Português Contemporâneo, onde abordagens semelhantes foram usadas.

***Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 407-418**

O método mais preciso para levantar a ambiguidade é certamente a verificação manual de cada ocorrência da palavra. Tal processo é no entanto bastante moroso. Desta forma, torna-se importante obter estes valores de uma forma automática ou, por vezes, semi-automática.

Em resumo, neste documento descreve-se o resultado da utilização conjunta de várias ferramentas e recursos, nomeadamente de:

- um analisador morfológico (*jspell*, e módulo Perl associado);
- um etiquetador morfo-sintáctico (EMS);
- um processador de corpora (CQP – Corpus Query Processor);
- vários corpora (CETEMPúblico, Diário do Minho, etc.).

O nosso objectivo é construir um conjunto de ferramentas que permita obter várias estatísticas:

- taxa de ocorrências de cada entrada (valor total das ocorrências das suas flexionadas e derivadas directas).
- padrão de ocorrência dessas palavras (conjunto de variantes da palavra que realmente aparecem).

Para cada um destes valores deveremos calcular uma medida de confiança (tendo em conta as ambiguidades mórfo-sintácticas).

Para que haja cooperação efectiva entre ferramentas, foi necessária uma plataforma de desenvolvimento comum: no nosso caso, optamos pela plataforma Linux, uma variante do sistema operativo Unix.

Ferramentas e recursos utilizados

Nesta secção apresentamos de forma independente as várias ferramentas e recursos utilizados.

Corpora

É naturalmente indispensável para este trabalho, uma vez que não há outro modo científico de determinar frequências.

Fazemos notar que o uso de um único corpus não balanceado pode induzir em erro o utilizador não prevenido. Por exemplo, embora nos nossos testes tenhamos usado extensivamente o corpus CETEMPúblico (Rocha & Santos, 2000), por ser um dos maiores corpora de língua portuguesa livremente disponíveis, este corpus sendo um corpus jornalístico, não cobre convenientemente a totalidade dos aspectos da língua: há diferenças significativas, quando comparado com discurso falado,

texto literário ou científico. As formas da primeira pessoa singular que colidem com substantivos da mesma raiz (por ex.: abandono, gosto, uso) que são presumivelmente menos frequentes num texto jornalístico que num corpus que inclua, discurso falado. Recorremos também ao corpus "Museu da Pessoa", porque, embora pequeno corresponde à transcrição de entrevistas, apresentado diferenças significativas. Mas, de um modo geral, procurou-se construir ferramentas que permitissem determinar as estatísticas e que pudessem ser ensaiadas com vários corpora.

Jspell

O *jspell* é um analisador morfológico e corrector ortográfico para português europeu, desenvolvido na Universidade do Minho (Almeida & Pinto, 1994) com base num corrector ortográfico para o inglês: *ispell*. O vocabulário deste programa foi recentemente amplamente expandido e corrigido utilizando diversos métodos semi-automáticos (Almeida & Simões, 2001).

Em relação ao trabalho presente, a característica mais importante deste programa é a capacidade de indicar quais os possíveis lemas da forma flexionada de uma determinada palavra. Não pretendemos discutir aqui o significado de entrada de dicionário, cuja definição sai largamente do âmbito deste artigo. Assim, consideramos aqui as entradas cujas frequências queremos determinar como sendo as entradas separadas definidas no dicionário deste analisador morfológico.

No exemplo abaixo, verificamos qual a resposta do programa quando confrontado com uma palavra ambígua. Repara-se como as duas respostas possíveis indicam dois lemas diferentes.

```
$ jspell -d port -a
@(#) International Jspell Version 1.00b1, 11/07/2001

gosto
* gosto 0 :lex(gosto,  [CAT=nc,G=m,N=s], [], [], []),
          lex(gostar, [CAT=v,T=inf,TR=_], [], [P=1,N=s,T=p], [])
```

EMS – Etiquetador Mórfo-Sintáctico

O EMS (Reis & Almeida 1997) é um etiquetador mórfo-sintáctico, ou seja, um programa que recebendo um texto devolve esse mesmo texto com a respectiva anotação gramatical. O EMS é uma variante do etiquetador de Eric Brill (Brill 1992), alterado de modo a utilizar o *jspell* como modo de minimizar o léxico necessário e diminuir substancialmente o peso do processo de treino.

Aqui interessa-nos principalmente a categoria gramatical da palavra. Abaixo apresentamos um exemplo de frase etiquetada com o EMS:

```
Governo/NCMS impõe/VIH3S limites/NCMP durante/P seis/DNCNP meses/NCMP
```

Cada uma das etiquetas que aparece à direita da palavra caracteriza-a utilizando a seguinte convenção:

- **NC:** nome comum;
- **VI:** verbo intransitivo;
- **P:** preposição.;
- **DNC:** determinante numeral cardinal.

Para casos de palavras ambíguas, o EMS começa por atribuir a cada palavra uma categoria (baseada numa das respostas do *jspell*), e, em seguida, de acordo com o seu contexto (classificação gramatical das palavras vizinhas), tenta corrigir a classificação inicial.

Por exemplo, a frase "eu gosto do gosto da batata" começa por ser etiquetada incorrectamente:

```
Eu/PSN1S gosto/NCMS do/&MS gosto/NCMS da/&FS batata/NCFS ./.
```

Note-se como o primeiro "gosto" é classificado como nome comum. No entanto, esta etiquetação inicial é modificada através de uma regra de modificação existente num ficheiro de regras.

```
NCMS VIH1S PREVTAG PSN1S
```

Ou seja, nomes comuns masculinos singulares (NCMS) são transformados em primeiras pessoas do singular do presente do indicativo (VIH1S), no caso de a etiqueta anterior (PREVTAG) ser o pronome pessoal da primeira pessoa do singular (PSN1S). Estas regras são obtidas por um processo de aprendizagem automática baseada um texto anotadas.

O resultado final, neste caso, é uma frase correctamente etiquetada.

```
$echo 'eu gosto do gosto da batata.' | ems

eu/PSN1S gosto/VIH1S do/&MS gosto/NCMS da/&FS batata/NCFS ./.
```

Módulo CQP

O Corpus Query Processor (CQP) é parte do *Corpus Workbench* do *Institut für Maschinelle Sprachverarbeitung* (IMS), da Universidade de Estugarda (Christ et al. 1999). Este programa corre em ambiente Linux e tem bom comportamento mesmo quando usado com corpora de grandes dimensões. Apresentamos abaixo alguns exemplos do uso desta ferramenta para a extracção de concordâncias:

```
CETEMPUBLICO> "Putin";
 43588961: eira confiança, Vladimir <Putin>, que ocupava o cargo de
113251175: a substituição, Vladimir <Putin>, primeiro adjunto do ch
154049540: urança (FSB), Vladimir <Putin>, e com o chefe do servi
```

O CQP permite ainda a utilização de corpora anotado, como se pode ver no exemplo abaixo, e nada objecta ao uso directo de um corpora anotado morfo-sintacticamente. Na verdade, o projecto Processamento Computacional do Português disponibiliza a consulta na Rede de vários corpora gramaticalmente anotados. (`http://cgi.portugues.mct.pt/acesso/`).

O exemplo abaixo apresenta um extracto retirado de um destes corpora.

```
CETEMPANOT> "gosto";
     7126: PEC_rel me/PERS_refl dá/V <gosto/N> imaginar/V como/ADV_inte
    95008: essiva/ADJ de/PRP mau/ADJ <gosto/N> ./PU No/PRP+DET_artd pró
   135739: ão/N que/SPEC_rel não/ADV <gosto/V> de/PRP beringelas/N amei
   144462: U como/ADV_rel_ks eu/PERS <gosto/V> de/PRP dizer/V ,/PU um/D
```

Para facilitar o uso sistemático e repetitivo deste programa, foi criado um módulo Perl[2] (CQP.pm). Embora se pudesse utilizar directamente o CQP, a construção deste módulo permitiu um grande número de facilidades que de outra forma não seriam integráveis tão facilmente na programação em Perl. Em particular, uma das funções, que dada uma palavra e um valor inteiro *n* retorna uma lista com *n* frases onde ocorre essa palavra foi bastante útil.

Esquema utilizado

A parte mais simples é a extracção de todos os tipos (ou seja formas gráficas distintas) e respectivas frequências existentes num corpora ou conjunto de corpora. O exemplo abaixo mostra um extracto de uma destas listas de frequência:

```
 6100  gosta
11835  gosto
 3207  gostava
 2551  gostou
 1718  gostos
```

Algumas destas formas não são ambíguas (gostou, gostos), e as suas ocorrências podem ser automaticamente atribuídas à palavra base correspondente (respectivamente, o verbo *gostar* e o substantivo *gosto*).

No entanto, *gosto* é uma palavra ambígua na língua escrita (embora não o seja na língua falada), uma vez que além de um substantivo, pode ser igualmente uma forma de verbo *gostar*, e só a análise do contexto nos permitirá determinar em que proporção as ocorrências desta palavra gráfica se repartem pelas duas palavras base.

Desta forma, depois de calculadas as ocorrências de cada palavra e detectadas as ambiguidades podemos criar uma lista com a respectiva confiança. Segue-se um algoritmo para obter estes resultados:

[2] À data da escrita deste artigo, não tinhamos conhecimento que os autores do CQP também estavam a desenvolver um módulo Perl para este programa, e contemplando uma maior funcionalidade.

```
use jspell;
jspell_dict("port");

foreach (numoco, palavra) in stdin
       w = lemas(palavra)
       duvida = (comprimento(w) > 1);

    foreach (lema) in (w)
           if (duvida) { oco[p][duv] += numoco; }
           else        { oco[p][gar] += numoco; }

foreach (palavra) in (sort dom(oco))
      total = oco[palavra][duv] + oco[palavra][gar];
      conf = oco[palavra][gar]/total;
      print palavra, total, conf;
```

Este algoritmo e variantes deram origem à ferramenta *freqnormpt*, disponível nas páginas do projecto Natura.

Depois de executar este programa, obtemos uma lista como a que se segue:

```
abade        571 (conf=100%)
abadia       270 (conf=100%)
abafar      1103 (conf=98%)
...
gostar     43194 (conf=70%)
...
gosto      14400 (conf=12%)
```

Esta ferramenta permite obter resultados mais complexos bastando para isso alterar as opções de linha de comando. Uma das possibilidades é a geração de um ficheiro Perl que pode ser incluído directamente em qualquer *script* Perl. Isto permite que se poupe muito tempo no desenvolvimento das aplicações.

Outra opção, mais legível pelo utilizador comum, pode ser obtida com a opção *–complex*, que retorna uma lista no seguinte formato:

```
$ freqnormpt -complex f
     ajuda (:229)
            ajuda(176) ajuda/ajudar
            ajudas(53) ajuda/ajudar
     ajudar (175:229)
            ajuda(176) ajuda/ajudar
            ajudar(140)
            ajudas(53) ajuda/ajudar
            ajudam(15)
            ajudaram(10)
            ajude(10)
            ...
```

Estes resultados estão divididos em grupos (um por lema). Junto a cada lema aparece o número de ocorrências não ambiguas seguido do número de ocorrências ambiguas. Segue-se uma lista de palavras gráficas com o respectivo número de ocorrências. No caso de ambiguidade, são indicados os lemas possíveis.

Usando métodos estatísticos seria possível determinar taxas de frequências mais precisas, recorrendo à informação das formas não ambiguas para determinar a redistribuição das formas ambíguas. Na secção seguinte vamos utilizar ferramentas sensíveis ao contexto para melhor analisar as situações de ambiguidade.

Depois do cálculo da taxa de ocorrências e da certeza respectiva podemos refinar alguns destes valores através da desambiguação das ocorrências de formas gráficas ambíguas extraindo frases exemplo da palavra em causa (usando o módulo CQP.pm) e classificando morfo-sintacticamente o extracto com base no contexto gramatical (usando o EMS).

```
A/DADFS Direcção/NCFS e/C o/DADMS treinador/JMS querem/VIH3P
que/QUE fique/VSH_S e/C eu/PSN1S gosto/VIH1S do/&MS Beira/NPMS
Mar/NPMS,/, por/P isso/PDNN .../...

No/&MS fundo/NCMS,/, transformando/VG Portugal/NPNN num/&MS
país/NCMS no/&MS qual/PRANS dê/VSH1S gosto/NCMS viver/VN ./.

À/ADV passagem/NCFS do/&MS último/JMS quarto/DNOMS de/P
hora/NCFS,/, de/P novo/JMS João/NPMS Paulo/NPMS voltou/VIP3S a/P
fazer/VN o/DADMS «/« gosto/NCMS ao/&MS pé/NCMS »/»,/, aproveitando/VG
um/DAIMS bom/JMS passe/NCMS de/P Carlitos/NPMS ./.

Quanto/ADV ao/&MS demais/JNS,/, tive/VIP1S o/DADMS gosto/NCMS
pessoal/JNS de/P saber/NCMS que/QUE,/, pelo/&MS menos/ADV,/,
quatro/DNCNP distritos/NCMP do/&MS país/NCMS,/, gostariam/VCH3P
que/QUE fosse/VSI3S deputado/JMS por/P esses/PDMP círculos/NCMP ./.

Mas/C,/, gosto/NCMS de/P lidar/VN com/P as/DADFP coisas/NCFP
na/&FS base/NCFS da/&FS verdade/NCFS ./.

Prova/NCFS disso/&NN mesmo/JMS é/VIH3S o/DADMS facto/NCMS de/P
todo/PFMS este/PDMS trabalho/NCMS ser/VN feito/JMS apenas/ADV por/P
gosto/NCMS ./.
```

A análise deste extracto etiquetado permite a contagem do número de ocorrências de determinada categoria nos extractos realizados. Do número obtido, podemos calcular uma percentagem da probabilidade de uma palavra pertencer a essa categoria. Desta forma, multiplica-se o número de ocorrências ambíguas por este valor obtendo um valor provável para o número de ocorrências. Este método permite ainda realizar uma estimativa que será tanto mais fiável quanto maior e mais equilibrado for o extracto. Além deste problema, podem ainda surgir palavras incorrectamente classificadas, como se pode ver no quinto exemplo anterior.

Usando o etiquetador PALAVRAS (Bick 2000), das 7.338 ocorrências do tipo *gosto*, 5.144 foram lematizadas com o substantivo, enquanto 2.187 forma lematizadas com o verbo *gostar*. No entanto, não há qualquer obrigatoriedade de analisar uma quantidade tão grande de ocorrências – amostras mais pequenas, desde que provenham de um corpus balanceado, podem entregar igualmente resultados válidos.

Realizado este processo sistematicamente para cada uma destas palavras ambíguas – o que pode ser um processo demorado, dada a grande quantidade de palavras a examinar— obtemos assim finalmente a lista ordenada de ocorrências de entradas de dicionários pretendida.

Uso de corpora anotado

O uso de corpora anotado permite evitar muitos problemas, partindo do princípio que a anotação do corpus está (maioritariamente) correcta. Por exemplo, estando cada palavra anotada com o devido lema, podemos usar uma ferramenta incluída no IMS-CWB, o lexdecode, para obter uma lista da quantidade de ocorrências de cada lema, como no exemplo abaixo, em que usamos o corpus etiquetado com o PALAVRAS.

```
$ lexdecode -f -P lema -p abandon.* cetempanot
  17507 abandonar
   4180 abandono
     68 abandonar+se
   ....
```

Note-se que, neste caso, a maior parte das ocorrências do lema abandonar são em casos onde não há ambiguidade quanto ao lema (abandone, abandonassemos, abandonando, etc.).

Este método, embora os resultados finais sejam rápidos de obter, assume que o etiquetador usado no corpus é estável e fiável, uma vez que o processo de anotação do corpus, com os etiquetadores actualmente disponíveis, é demorado. Além disso, não permite a distinção entre palavras com o mesmo lema. Alguns lemas ambíguos podem ser no entanto distinguidos com recurso a uma análise morfológica. Por exemplo, existem pelo menos três significados possíveis não figurativos do lema "lama". Um desses significados (lodo) pode ser frequentemente identificado automaticamente devido às suas características morfológicas (nomeadamente, é um substantivo feminino); no entanto, a distinção entre os dois lemas que são substantivos masculinos (sacerdote budista e mamífero sul-americano) exige um detalhe e provavelmente um trabalho manual que estão para além dos nossos objectivos.

Patamarização de frequências

Para que se consiga uma fácil leitura das frequências de cada lema, pode-se dividi-los em patamares (ou degraus) que podem ser definidos de várias formas, dada a lista de lemas ordenadas por ocorrências:

- dividi-los em grupos contendo igual número de lemas (por exemplo, os mil lemas mais frequentes, os mil lemas seguintes, etc.);
- dividir de acordo com o número de ocorrências usando uma escala logarítmica
- dividi-los em k grupos, cada qual contendo os lemas correspondentes a $1/k$ do corpus

No próximo exemplo usamos a primeira opção, em que os números indicados correspondem à identificação do patamar em que a palavra se encontra, usando como base o CETEMPúblico anotado.

```
$ lexdecode -f CETEMPANOT > ocorrencias
$ freqnormpt -oco -steps=1000:2000:3000 ocorrencias
...
de        3
a         3
o         3
que       3
...
desportivo 2
ferir      2
curto      2
...
colaborador 1
erguer      1
mercadoria  1
...
mobiliário  0
acender     0
suspeitar   0
...
```

Abaixo apresentamos uma tabela numa escala logarítmica de tamanho 6 em relação à frequência da palavra mais comum (*de*, que ocorre cerca de 5 milhões de vezes), em que apresentamos a última palavra de cada um desses patamares.

Ocorr.	Lema	Qt. Lemas no patamar	Patamar
500000+	seu	14	6
50000+	social	152	5
5000+	travar	1819	4
500+	debilitar	6595	3
50+	adorador	18907	2
5+	abananar	55069	1

Finalmente, apresentamos uma tabela onde a cada um dos cinco patamares corresponde um número aproximadamente igual de ocorrências no corpus (i.e., os lemas do primeiro patamar correspondem a um quinto do corpus, e assim sucessivamente). Cada patamar é representado na tabela pelo seu primeiro e último lema.

Lemas	Qt. Lemas no patamar	Patamar
de ... ser	6	4
um ... também	30	3
sobre .... tipo	280	2
lançar ... praça	1231	1
adesão ...	212939	0

Problemas

Um problema já referido é referente aos casos de corpora não balanceados. Por exemplo, na tabela seguinte, mostramos as ocorrências por milhão de palavras de alguns lemas em dois corpora que, embora ambos baseados em texto jornalístico, apresentam um uso de vocabulário extremamente diverso.

	Avante	Diário do Minho
camarada(s)	170,6	4,3
arcebispo(s)	2,4	78,0

Assim sendo, para o bom desenrolar deste projecto é essencial usar um conjunto de corpora de textos diversificado e bem balanceado. Reunindo corpora de diferentes fontes, pode-se inclusive excluir das listas palavras cuja elevada ocorrência no corpus se deve a uma única fonte (Kilgarrif 1996), ou indexar uma palavra a um domínio específico (religião, política, etc.).

Em relação a este problema foi desenvolvida uma ferramenta que permite a comparação entre a frequências de corpora distintos, e que, embora trivial, pensamos ser útil para o tratamento constrastivo de frequências em corpora. Assim, é possível obter grupos de palavras, mais frequentes num dos corpora que nos restantes. Este comparador pode ser para vários fins como seja o desenvolvimento de ferramentas de classificação automática de assuntos baseando-se na anterior aprendizagem sobre textos pré-classificados.

Este método não dá naturalmente nenhuns resultados quanto a palavras homógrafas pertencentes à mesma categoria gramatical (por exemplo, lama), embora o princípio do uso de corpora continue a ser válido, é indispensável para estes casos uma cuidadosa verificação manual, caso se pretenda efectivamente incluir no dicionário tal distinção; uma análise automática é possível (Stevenson & Wilks, 2001), mas é bastante trabalhosa e o grau de fiabilidade é baixo.

Conclusões

Neste artigo, em vez de um recurso final fechado (as frequências do português) propõe-se um conjunto de ferramentas que, dado um corpus, permita calcular essas mesmas frequências.

Existindo corpora anotados, estes podem ser aproveitados para calcular directamente as frequências.

No caso da não existencia destes, podemos calcular as frequências com base na capacidade lematizadora de um analisador morfológico e opcionalmente utilizar um etiquetador morfo-sintáctico para etiquetação parcial de alguns exemplos para desambiguação dos casos críticos.

Parece-nos importante disponibilizar ferramentas de domínio público que permitam o cálculo de frequências afectadas por um factor de confiança e que possam ser usadas em estudos sobre segmentos específicos da linguagem aplicados a quaisquer corpora.

Os resultados calculados dependem somente da qualidade do dicionário do analisador morfológico e do etiquetador morfo-sintáctico.

Estes mesmos resultados são imprescindíveis para serem usados por outras ferramentas (contrastivas, como o *freqcomp*, para serem sujeitas a patamarização ou genericamente incorporadas noutras ferramentas) pelo que estes devem ser produzidos num formato reutilizável.

Trabalho futuro

Pretendemos analisar o módulo Perl do CQP, disponibilizado pelos autores dessa ferramenta, para tentar tirar o máximo partido das funcionalidades oferecidas.

Também pretendemos testar outros etiquetadores e realizar contagens sobre corpora já anotados, realizadas extensivamente para que se pudessem comparar os nossos resultados com contagens realizadas por outros grupos (p.ex. o Léxico Multifuncional Computorizado do Português Contemporâneo).

Referências

IMS Corpus Workbench http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Léxico Multifuncional Computorizado do Português Contemporâneo
http://www.clul.ul.pt/sectores/projecto_lmcpc.html

Museu da Pessoa http://portugal.museudapessoa.net/

Projecto Natura http://natura.di.uminho.pt

Projecto Processamento Computacional do Português http://www.portugues.mct.pt/

Almeida, J.J. e Ulisses Pinto, "Jspell – um módulo para análise léxica genérica de linguagem natural", *Actas do Congresso da Associação Portuguesa de Linguística*, Évora, 1994. http://www.di.uminho.pt/~jj/pln/jspell1.ps.gz

Bick, Eckhard. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press,2000.

Brill, Eric. "A simple rule-based part of speech tagger". *Third Conference on Applied Natural Language Processing*. ACL, Trento, Itália, 1992.

Christ, Oliver, 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), http://www.ims.uni_stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

Kilgarriff, Adam. "Putting Frequencies in the Dictionary" In International Journal of Lexicography 10 (2), 1997, pp. 135-155 ftp://ftp.itri.bton.ac.uk/reports/ITRI-96-10.ps.gz

Medeiros, José Carlos, Rui Marques & Diana Santos. "Português Quantitativo", Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada) – EPLP'93, (Lisboa, 25-26 de Fevereiro de 1993), pp.33-8.

Reis, Ricardo & José João Almeida. "Etiquetador morfo-sintático para o Português", *Actas do XIII Congresso da Associação Portuguesa de Linguística*. Lisboa 1997. http://www.di.uminho.pt/~jj/bib/etiquetador2.ps.gz

Rocha, Paulo Alexandre & Diana Santos. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) (Atibaia,

São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp.131-140. http://www.portugues.mct.pt/Diana/download/RochaSantosPROPOR2000.pdf

Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavrilidou et al. (eds.), Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000 (Athens, 31 May-2 June 2000) http://www.portugues.mct.pt/Diana/download/SantosBickLREC2000.rtf

Stevenson, Mark & Yorick Wilks. 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. In *Computational Linguistics,* v.27, n.3, pp. 321-350.