CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa

Paulo Alexandre ROCHA
SINTEF Tele og Data
Postboks 124 Blindern
NO-0314 Oslo, Noruega
Paulo.A.Rocha@informatics.sintef.no

Diana SANTOS
SINTEF Tele og Data
Postboks 124 Blindern
NO-0314 Oslo, Noruega
Diana.Santos@informatics.sintef.no

Abstract

This paper reports on the creation of CETEMPúblico, the largest publicly available corpus of Portuguese to date, containing 180 million words, created to boost research in language engineering in Portuguese. After providing some background for creating it, we focus on the processing required, explaining in detail some options taken, namely: the division of articles in extracts; their random reordering and numbering in the final corpus; the marking of structural units such as sentence separation, titles and author identification; the use of a partial system for contents classification; and the distribution methods.

Resumo

Este artigo relata a criação do CETEMPúblico, um corpus de cerca de 180 milhões de palavras em português europeu, que esperamos expandirá consideravelmente a pesquisa na engenharia linguística do português. Descrevemos a motivação e o processamento envolvido na sua construção, explicando em pormenor algumas das opções tomadas: a divisão dos artigos em extractos; a sua reordenação aleatória; a marcação de unidades estruturais tais com a divisão em frases e a identificação de títulos e autores; o uso de um sistema parcial de classificação dos artigos do corpus; e os métodos de distribuição considerados. Finalizamos discutindo a questão das versões e referindo trabalho futuro.

1. Introdução

Apesar de existirem vários corpora de português de Portugal, não existe nenhum que se possa considerar de grandes dimensões. A título de exemplo, o maior corpus de português acessível é o Corpus NILC/São Carlos, contendo cerca de 36 milhões de palavras em português brasileiro; para o português europeu, a quantidade total de palavras em todos os corpora livremente disponíveis não excede os 10 milhões.[1]

Tentando contrariar esta falta de recursos, o Ministério da Ciência e da Tecnologia (MCT) português encomendou-nos a produção de um corpus de maior dimensão, distribuído numa forma apropriada para um leque de tarefas de processamento de linguagem natural (PLN) o mais vasto possível.

Assim, negociámos com um jornal português com alguma experiência computacional, o *Público*, um contrato para disseminar 100 milhões de palavras de texto corrido, retiradas das suas edições, em troca de uma pequena contrapartida financeira. O corpus por nós criado não poderia, contudo, permitir a reconstrução dos artigos completos e/ou das edições integrais do jornal. Portanto, uma das condições acordadas no protocolo assinado entre o MCT e o *Público*, garantindo ao primeiro o uso de material do segundo, impõe a impossibilidade da reconstrução automática dos artigos usando o corpus.

A versão 1.0 do CETEMPúblico (**C**orpus de **E**xtractos de **T**extos **E**lectrónicos **M**CT/**Público**), criada a 25 de Julho de 2000, contém cerca de 180 milhões de palavras distribuídas por 1.567.625 extractos, correspondentes a cerca de 1.500 edições diárias

[1] Veja-se o catálogo de recursos do nosso projecto no URL http://www.portugues.mct.pt

(algumas delas incompletas), quase inteiramente em português europeu. Para facilitar o seu uso, criámos a página http://cgi.portugues.mct.pt/cetempublico/ onde publicamos informação actualizada sobre o corpus. Este, além de distribuído gratuitamente em CD, organizado em vinte volumes, encontra-se incluído no nosso serviço de acesso a corpora através da Internet (AC/DC).

Neste artigo descrevemos o trabalho envolvido na criação deste recurso, documentando as opções tomadas e o resultado obtido.

2. Breve descrição do *Público*

Fundado em 1990, o *Público* é um jornal diário português, o primeiro a ser publicado simultaneamente em Lisboa e no Porto. Além dos suplementos locais de cada edição, o jornal inclui as seguintes secções: Destaque, Política, Internacional, Sociedade, Ciência, Educação, Desporto, Média, Cultura, Economia e Última Página. Semanalmente são publicados cinco suplementos (Computadores, Economia, Artes, Sons e Leituras) e uma revista dominical (a Pública).

A maioria dos artigos do *Público* é escrita por jornalistas portugueses. Apesar de existirem alguns textos de autores brasileiros e africanos, eles representam uma parte diminuta – menos de 0.2% – e seguem frequentemente a norma portuguesa. O jornal publica também artigos de agências noticiosas internacionais (Reuters e AFP) e de jornais do grupo *World Media*, ao qual pertence (como o espanhol *El País*, o britânico *The Guardian*, o francês *Libération* e o italiano *La Repubblica*).

O *Público* foi o primeiro jornal português com uma edição completa *online* (http://www.publico.pt), lançada em 1995. Este serviço fornece as últimas sete edições do jornal gratuitamente e disponibiliza vários serviços de informação exclusivamente na rede. Foi igualmente o primeiro (e até agora único) jornal português a publicar um livro de estilo (LEP). Finalmente, desde a sua fundação o *Público* tem fornecido material textual a vários grupos de I&D interessados no processamento da língua portuguesa, como se pode ver nos nomes dos corpora portugueses existentes no nosso catálogo: Natura/Público e BD-Público. Além disso, fornece o material português para o sistema de concordâncias por correio electrónico GlossaNet, desenvolvido pelo LADL.

3. Material base

Numa primeira fase, recebemos seis CDs, a que chamaremos conjunto A, contendo cada um aproximadamente seis meses de artigos referentes aos anos de 1996 a 1998. Cada CD estava dividido em vários directórios, cada um contendo os artigos de uma edição (diária) do jornal. Não recebemos textos dos suplementos, excepto cerca de três semestres do suplemento Computadores. Na maior parte dos casos, cada ficheiro correspondia a uma notícia. No entanto, algumas notícias estavam divididas em secções menores (uma em cada ficheiro), como documentado no Livro de Estilo do *Público*; por outro lado, certos ficheiros continham várias notícias (as Breves).

Além dos artigos incluídos em cada edição do jornal, recebemos também muitos artigos que não chegaram a ser publicados por falta de espaço ou de oportunidade. Como esses artigos não diferiam estilisticamente dos outros, integrámo-los no corpus.

Foram-nos fornecidas versões dos artigos em texto simples e HTML. Uma vez que o conteúdo era basicamente o mesmo, decidimos usar os ficheiros de texto, pois os ficheiros HTML continham muita informação inútil para os nossos objectivos e o seu formato foi alterado algures no decorrer do período coberto pelos CDs, sendo o formato de texto mais consistente durante a totalidade do período que estávamos a tratar.

O material foi-nos entregue em formato Macintosh Roman, que convertemos para ISO-8859-1, por ser um padrão muito mais usado e, além disso, compatível com os programas que utilizamos para a disponibilização do corpus na rede.

Após ter processado o material enviado pelo jornal, descobrimos que apenas abrangia cerca de 57 milhões de palavras. O *Público* enviou-nos então um segundo conjunto de CDs, cobrindo os anos de 1991 a 1995, daqui em diante referido como conjunto B. Esse conjunto continha apenas ficheiros de texto, e cada ficheiro era relativo a uma edição. Não pudemos, portanto, utilizar os algoritmos de selecção de artigos baseados no nome dos ficheiros que usámos para o conjunto A. Não podemos, também, saber se os ficheiros do conjunto B incluem artigos rejeitados.

4. Processo de criação

Na Figura 1 apresentamos uma visão geral do processamento efectuado, passando a descrever cada passo em pormenor.

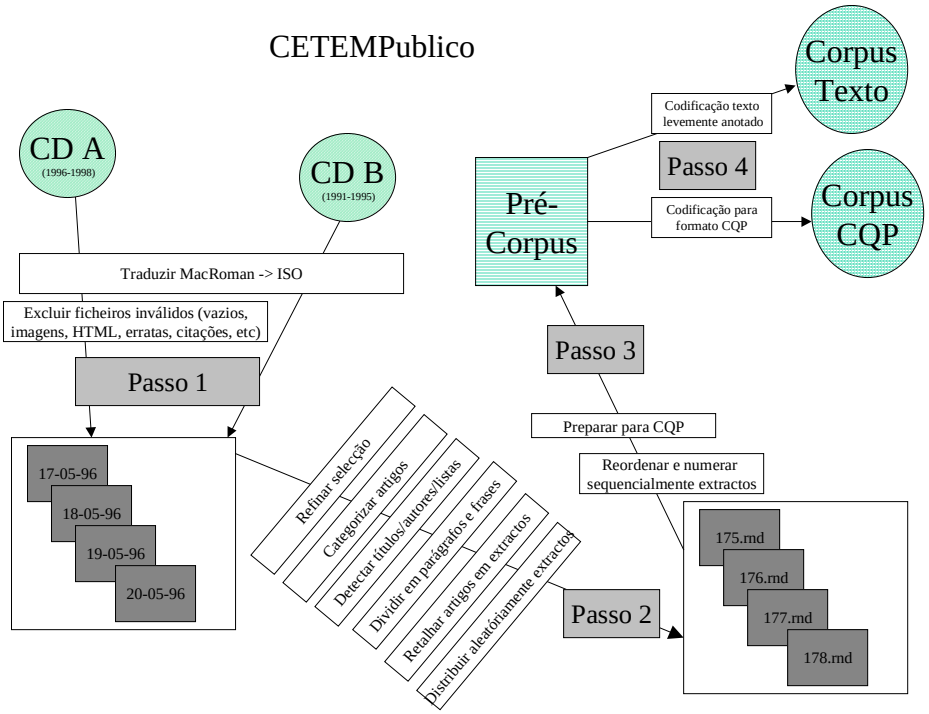

Figura 1: Passos para a construção do corpus

4.1 Passo 1

Este passo é iniciado por uma rotina que lê cada ficheiro do CD e cria uma versão modificada no disco. Para o conjunto A, essa rotina inclui também uma primeira filtragem do material aceite, excluindo à partida imagens, ficheiros em formato HTML, citações de outros jornais e erratas — ficheiros esses facilmente identificáveis pelo nome. Uma vez que o processo de leitura a partir do CD é demorado, todo o tratamento posterior é feito com base neste novo conjunto de ficheiros armazenados no disco duro. Ainda durante a execução desta rotina o texto é convertido do formato Macintosh Roman para o formato ISO-8859-1.

Para facilitar o manuseamento, foi criado um ficheiro para cada edição diária, num total de aproximadamente mil ficheiros, cada um com cerca de 120 artigos. Cada artigo foi

etiquetado com o nome do ficheiro original[2] e o semestre a que pertence (por exemplo, 96a identifica o primeiro semestre de 1996, 98b o segundo semestre de 1998).

Para o conjunto B, este passo consiste apenas na leitura, recodificação e gravação do texto. Uma vez que não havia ficheiros separados para cada artigo, a divisão em artigos e a sua selecção foi adiada para o passo seguinte.

4.2 Passo 2

No segundo passo, várias operações são efectuadas em sequência sobre cada edição diária:

- repartição em artigos
- selecção dos artigos (para o conjunto B) e seu refinamento (para ambos os conjuntos)
- classificação dos artigos
- identificação e anotação de títulos e autores
- separação de frases
- divisão de artigos em extractos

Estas tarefas decorrem de forma ligeiramente diferente para os dois grupos de CDs devido às diferenças no formato e conteúdo.

4.2.1 Selecção

Primeiro, repartimos os ficheiros correspondentes às edições diárias em artigos, excluindo algumas imagens que tenham passado pelo filtro anterior e, sempre que possível, artigos compostos exclusivamente por resultados desportivos, classificações e *rankings*. Todo este processo é automático, e é trivial para o conjunto A. Para o conjunto B, a divisão da edição em artigos é heurística, recorrendo a uma expressão regular que tenta detectar a linha inicial que identifica um novo artigo.

Alguns artigos existem em duplicado nos CDs que contêm o material base, seja porque aparecem simultaneamente em ambos os suplementos locais, seja porque uma notícia não publicada num dia o foi no dia seguinte. Apesar de inicialmente termos pensado criar uma rotina que evitasse a existência de artigos duplicados, ao verificar que por vezes estes mudam apenas de título, enquanto que um mesmo título pode corresponder a artigos diferentes, concluímos ser praticamente impossível levar a bom termo a tarefa sem verificar os artigos um a um. No entanto, para o conjunto B, conseguimos eliminar vários duplicados, nos casos em que a expressão regular referida acima emparelhava com a identificação de um artigo já incluído no corpus.

Em seguida, dentro de cada artigo, excluímos legendas de fotografias e algumas etiquetas de HTML existentes nalguns (poucos) ficheiros.

4.2.2 Reclassificando o material

Visto que, no conjunto A, era fácil identificar a secção do jornal à qual pertenciam os textos, resolvemos fornecer uma classificação do tipo de texto a que o extracto pertence.

O *Público* tem o seu próprio critério de classificação de artigos. Por exemplo, um ficheiro cujo nome começa por 'C' pertence à secção de Cultura, enquanto um ficheiro com um 'D' inicial pertence à secção de Desporto. A Figura 2 mostra a distribuição de palavras por categoria, de acordo com os critérios do *Público*, no conjunto A.

Assumindo que a classificação de texto mais interessante para o processamento de linguagem natural (PLN) é a relativa ao assunto ou ao estilo e não à formatação ou à posição física no jornal, resolvemos construir os nossos próprios critérios. Apesar de uma secção de Desporto incluir quase exclusivamente artigos sobre desporto — e textos sobre desporto raramente aparecerem fora da sua secção, excepto muito excepcionalmente como

[2] Exclusivamente para uso interno. Esta informação é retirada posteriormente.

Destaque — casos como Última Página e Internacional revelam muito pouco acerca da natureza dos assuntos discutidos.

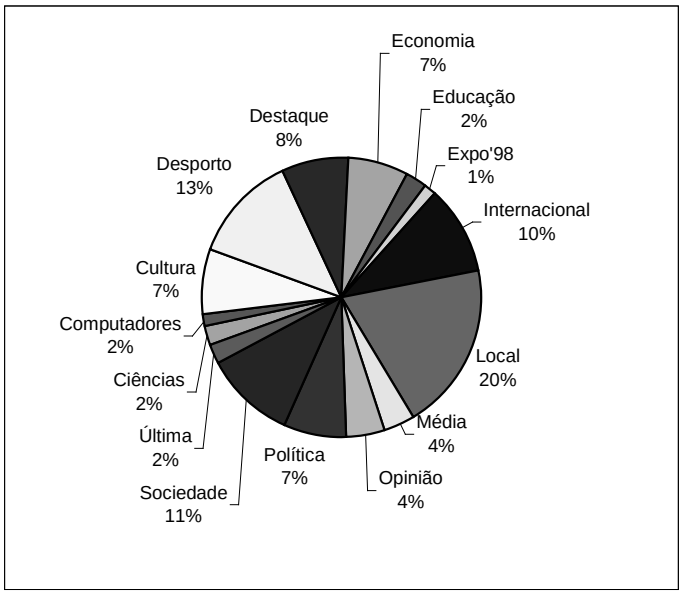

Figura 2: Distribuição de palavras, por categoria do Público, no conjunto A

Tentámos, portanto, criar automaticamente outro sistema de classificação: entre outras modificações, os artigos de opinião espalhados em diferentes secções foram agrupados, Educação foi incluída em `Sociedade`, e Média e Ciência em `Cultura`. Nalguns casos, optámos por atribuir uma classificação múltipla: por exemplo, notícias da secção Internacional são classificadas como `Política ou Sociedade`. Casos como Última Página e Destaque são classificados como `ND` (não determinável), visto que não considerámos obrigatório atribuir classificações.

Esta reclassificação é configurável, e tem o agradável efeito colateral de dificultar ainda mais a reconstrução de uma edição completa. A Figura 3 mostra a distribuição de palavras de acordo com as novas categorias, no conjunto A.

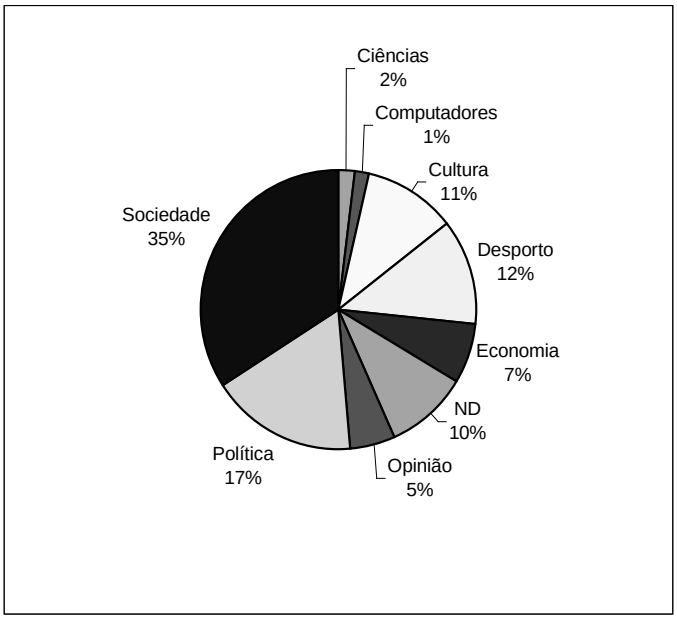

Figura 3: Redistribuição de palavras, por categoria do CETEMPúblico, no conjunto A

No caso do conjunto B, ocorreram duas situações distintas: Enquanto que não foi possível extrair a informação correspondente aos artigos de 1991, os artigos posteriores a 1992 indicavam a secção numa das suas linhas iniciais, embora com uma classificação algo distinta.[3]

No CETEMPúblico temos, pois, cerca de 15% dos extractos com a etiqueta `ND` (não disponível), incluindo todos os artigos de 1991. Mesmo que muitos extractos não estejam classificados, considerámos que a informação que conseguimos fornecer poderia ser útil para os utilizadores. As Figuras 4 e 5 apresentam a distribuição das categorias no corpus completo, por palavras e por extractos, respectivamente.

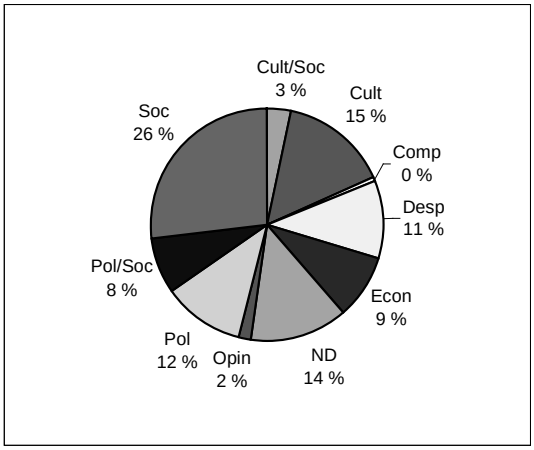

Figura 4: Distribuição de palavras por categoria no CETEMPúblico

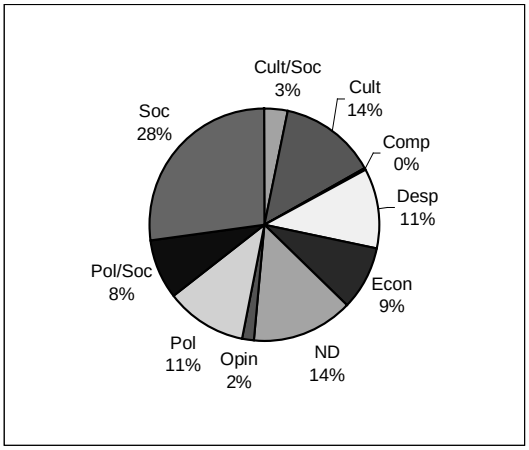

Figura 5: Distribuição de extractos por categoria no CETEMPúblico

As classificações são evidentemente arbitrárias, tal como afinal as próprias classificações do *Público*. Por exemplo, numa análise a diferentes portais, o SAPO (http://www.sapo.pt/) classifica desporto sob a categoria `Entretenimento`, enquanto o AEIOU (http://www.aeiou.pt/) tem uma categoria própria `Desporto` e o CUSCO (http://www.cusco.pt/) agrupa numa só categoria `Desporto e Lazer`.

4.2.3 Anotação do texto

Cada artigo foi anotado com o semestre de publicação e com a nova classificação que lhe foi atribuída. Além dessa informação externa, também identificámos título e subtítulos e autores. Uma vez que todos os sistemas de PLN que lidem com texto jornalístico terão que se confrontar com a existência destas categorias de uma ou outra forma, pareceu-nos desnecessário criar um corpus mais difícil do que a realidade.

Em primeiro lugar tratamos da identificação dos autores. Note-se que a identificação correcta de nomes próprios em textos é um processo complexo (Mikheev et al., 1999). Nos artigos do *Público*, os autores podem aparecer no princípio ou no fim de um artigo, e podem ser jornalistas ou colunistas ocasionais (como o presidente *Mário Soares* ou a primeira-ministra *Gro Harlem Brundtland*). Os pequenos artigos são habitualmente assinados apenas com as iniciais do(s) jornalista(s). Como se pode verificar na Tabela 1, uma elevada percentagem dos artigos são, de facto, assinados. A percentagem de artigos nos quais conseguimos identificar um autor varia de 68%, nos artigos de Opinião, até 19%, nos artigos de Última Página.

O processo de identificação de autores contempla dois casos distintos:

[3] Note-se que as secções do *Público* não permaneceram inalteradas desde a criação do jornal, tendo inclusive ocorrido alterações durante a redacção do presente artigo.

- o primeiro, em que a assinatura é simplesmente as iniciais, no qual assinalámos os grupos de duas ou mais maiúsculas seguidas de ponto no final de um artigo (por exemplo, *J.V.M.*).
- o segundo, em que pelo menos um nome é apresentado por extenso, para o qual seleccionámos, de entre as linhas sem pontos finais, de interrogação ou exclamação, aquelas em que todas as palavras tinham uma maiúscula inicial (*Mário Soares*), com ou sem a preposição *de* ou uma das suas contracções entre essas palavras (*Tolentino da Nóbrega*). Levámos em atenção os casos em que há uma ou mais iniciais intermédias (*José V. Malheiros*), em que o autor é seguido da sua localização (*Karim El-Gawhry, no Cairo*), em que há mais do que um autor (*Margarida Gomes e Vítor de Sousa*), e também aqueles em que os autores são identificados como '*(d)o nosso enviado*'.

Este processo é, evidentemente, imperfeito e identifica por exemplo como autores (sub)títulos como *Vinho do Porto* ou *Brinquedos da Idade da Pedra*. Chamamos também a atenção para a diferença entre as expressões *Paulo Moura, em Washington* e *Paulo Moura em Washington*: a primeira é identificada como assinalando o autor, enquanto a segunda é considerada um título, visto que uma expressão do tipo *Jorge Sampaio em Bucareste* refere-se provavelmente a uma visita presidencial à Roménia e não a uma reportagem de um jornalista chamado Jorge Sampaio. Assim sendo, o algoritmo não está imune a erros existentes no texto original.

Após uma primeira execução deste processo, detectámos vários subtítulos que eram frequente e erradamente identificados como autores e excluímos esses casos (*Médio Oriente*). Excluímos, também, os casos em que o título começava com certas preposições ou artigos (*Na Universidade de Coimbra)*, continha certas palavras institucionais (*República, Sociedade, Clube*) ou era exclusivamente composto de maiúsculas (*PERGUNTA DO DIA*). Impusemos, ainda, que certos tipos de artigos nunca contêm autores, nomeadamente os barómetros[4] políticos e desportivos. Em seguida, reexecutámos o processo.

Para títulos e subtítulos, utilizámos uma estratégia muito simples: as linhas sem sinais de pontuação que não tinham sido identificadas como autores foram consideradas títulos, desde que não fossem a última linha do artigo. Além disso, considerámos como títulos linhas iniciadas por reticências e sem outra pontuação (*... e na Bulgária também* – este caso é um subtítulo, sendo o título principal *Greves na Polónia*). De igual modo, foram aceites títulos onde exista uma inicial a meio da linha (*Porto festeja S. João*).

Avaliámos estes algoritmos manualmente, usando como amostra cinco edições do jornal de outros tantos anos distintos, concluindo que cerca de 4% dos autores assinalados não o eram (15 em 340), enquanto 12 autores não eram identificados. A correcção é mais elevada no conjunto A que no B, devido à maior facilidade de identificar os artigos individuais (nomeadamente os barómetros). Igualmente, cerca de 5% dos autores foram erradamente classificados como títulos. Quanto aos títulos, os resultados revelaram uma substancial sobre-identificação. Longas listas sem pontuação tinham sido identificadas como sequências de títulos, tais como a programação televisiva ou os percursos dos candidatos presidenciais na campanha eleitoral de 1996 em Portugal. Decidimos, por isso etiquetar tais sequências como listas e não como títulos, mudando a anotação de título para lista num passo posterior.

Na Tabela 1 apresentamos a distribuição destas anotações no corpus. O número de títulos identificados excede largamente o número de artigos, pois a grande maioria dos

[4] Um barómetro é um artigo de opinião onde se avalia positiva ou negativamente a prestação de diversos intervenientes nas cenas política e desportiva portuguesas na semana anterior.

artigos contêm subtítulos, seja seguindo ou precedendo o título principal, seja algures a meio do texto.

4.2.4 Separação em frases

Em seguida, procedeu-se à separação em frases utilizando a biblioteca de programas desenvolvidos no projecto AC/DC para lidar com corpora de português. Para cada parágrafo que não tivesse sido identificado como autor ou título foi invocado o programa separador de frases, que entra em conta com abreviaturas, codificação de numerais e outras particularidades da língua portuguesa envolvendo sinais de pontuação.

O resultado deste processo (que não é 100% fiável, mas cujo desempenho tem vindo a ser testado e melhorado no conjunto dos outros corpora processados) é a introdução de marcas de `<s>` e `</s>` (princípio e fim de frase) no texto.

4.2.5 Criação de unidades de texto mais pequenas

Em seguida, para respeitar as condições do acordo que impunham a impossibilidade de reconstrução dos artigos, fragmentámos esses artigos em unidades menores, delimitadas por `<ext>` e `</ext>` para serem posteriormente misturadas aleatoriamente.

Basicamente, um extracto é composto de dois parágrafos. No entanto, parágrafos com menos de 15 palavras, assim como títulos e autores, são agrupados com o parágrafo anterior ou, no caso de iniciarem um artigo, com o seguinte. Após repartir os artigos de cada edição em extractos segundo estes critérios, cada extracto é gravado num ficheiro escolhido aleatoriamente, numerado de 000 a 999, para limitar o número de ficheiros existentes. Em média, cada artigo dá origem a 4,35 extractos, variando entre 6,54 no Destaque e 2,31 nas notícias de Última Página.

	Número	Por artigo	Por extracto
Artigos	112.509	-	-
Extractos	489.033	4,35	-
Títulos identificados	269.848	2,39	0,55
Autores identificados	47.763	0,42	0,10

Tabela 1: Valores dos marcadores estruturais, relativos ao conjunto A de CDs

4.3 Passo 3

Em primeiro lugar, é criado um ficheiro compacto, lendo sequencialmente cada um dos ficheiros criados no passo anterior. Após eliminar as anotações que identificam o ficheiro original, os extractos contidos nesse ficheiro são reordenados aleatoriamente, numerados, e gravados no formato de uma unidade de codificação ("token") por linha. Esse processo é repetido para todos os ficheiros, criando um ficheiro único para os 180 milhões de palavras.

4.4 Passo 4

Decidimos disponibilizar o corpus em dois formatos: Uma versão texto com codificação tipo SGML destina-se aos investigadores que quiserem usar o corpus como entrada para os seus sistemas, enquanto uma versão CQP[5] destina-se aos investigadores que queiram sobretudo consultar o corpus, interrogando-o, mas não o alterando, versão essa usada no serviço AC/DC. Assim, permitimos quer um uso simples do CETEMPúblico (em que o utilizador, tipicamente um linguista, não precisa de instalar outros programas, apenas

[5] O CQP, Corpus Query Processor (Christ et al., 1999) é uma das ferramentas do IMS CWB, o Corpus Workbench do Institut für Maschinelle Sprachverarbeitung da Universidade de Estugarda. Para uma apresentação das capacidades deste sistema para o português, veja-se também Santos & Ranchhod (1999).

procurar no corpus) quer um uso complexo (em que um utilizador, tipicamente um informático, pode manipular livremente o corpus sem restrições).

4.4.1 Criação do corpus em formato de texto

Usamos para o efeito um programa que, dado um ficheiro com uma unidade por linha, obtém uma versão em texto corrido, normalizado. Ou seja, sempre com o mesmo espaçamento entre sinais de pontuação, parágrafos, frases, etc. Esse programa, aliás trivial, foi desenvolvido no âmbito da cooperação com Eckhard Bick para a análise sintáctica dos corpora (Santos & Bick, 2000), mas demonstrou ser de grande utilidade para garantir a compatibilidade entre as versões nos dois formatos. A principal razão do seu uso no caso do CETEMPúblico é a de garantir que não haja qualquer diferença entre as versões texto e CQP do corpus, evitando que uma das versões contenha menos informação, ou inclua erros de codificação próprios.

Devido às dificuldades de manipulação de tão grande ficheiro, dividimo-lo em 20 partes, contendo cada uma 80.000 extractos (excepto a última, que contém apenas cerca de 37.500).

4.4.2 Criação do corpus em formato CQP

Para criar o CETEMPúblico no formato CQP, basta usar os programas associados a esta caixa de ferramentas, o que não significa que, dada a grande dimensão do corpus, essa utilização tenha sido isenta de problemas. O IMS CWB é distribuído ao público em geral com uma licença padrão para efeitos de investigação, havendo também licenças especiais para uso comercial. Para o CETEMPúblico, como aliás para o British National Corpus (BNC) antes dele, os investigadores de Estugarda compilam um *sampler*, ou seja, uma versão restrita dos programas que não carece de licença e que é apenas utilizável com um único corpus, distribuída livremente em CD.

As razões que nos levaram a utilizar este ambiente específico de processamento de corpora são múltiplas:

- O IMS-CWB é um sistema robusto e poderoso, do qual temos bastante experiência (Santos, 1998).
- Corre no sistema operativo Linux, um sistema distribuído livremente e não propriedade de uma companhia.
- Vários corpora de grande dimensões, como o BNC ou o CNC (Czech National Corpus), são distribuídos neste formato, tornando-o assim um standard de facto.
- É o sistema usado pelo nosso projecto para dar acesso aos corpora através da rede, no âmbito do projecto AC/DC.
- É particularmente apropriado para codificar corpora analisados (Santos & Bick, 2000), como pretendemos que o CETEMPúblico seja num futuro próximo.

Não pretendemos, contudo, forçar ninguém a utilizá-lo. A distribuição paralela em formato de texto permite a integração do corpus em qualquer ambiente.

5. Observações finais

Contamos com uma resposta maciça da comunidade interessada no processamento computacional do português que nos permita criar novas versões mais informadas após o uso e disseminação de uma primeira versão. Manteremos, pois, um controle rigoroso das versões do corpus, assim como uma documentação aturada das alterações. Da forma como o corpus é criado (usando um processo aleatório), não será possível manter a numeração de extractos de versão para versão, embora não seja previsível que o conteúdo sofra alterações significativas. Por isso, é importante chamar a atenção dos utilizadores para este

facto, e pedir que refiram, em todas as publicações e estudos que façam com base neste recurso, a versão que utilizaram.

Como referido, planeamos, em colaboração com Eckhard Bick, anotar o CETEMPúblico muito em breve.

Agradecimentos

Agradecemos a José Vítor Malheiros, responsável pela edição *online* do *Público*, sem cuja inestimável ajuda o corpus não existiria. Estamos também gratos a Stefan Evert e a Arne Fischer, da Universidade de Estugarda, pelo apoio em relação ao IMS-CWB, e a Miguel Andrade pelo apoio jurídico prestado.

Referências

AC/DC. Acesso a Corpora / Disponibilização de Corpora. http://cgi.portugues.mct.pt/acesso/

BNC. British National Corpus. http://info.ox.ac.uk/bnc/

Oliver Christ, Bruno M. Schulze, Anja Hofmann & Esther Koenig (1999). 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/

CNC. Czech National Corpus. http://ucnk.ff.cuni.cz/english/index.html

GlossaNet. http://glossa.ladl.jussieu.fr/

LEP (1998). *Livro de Estilo do Público*, Lisboa, 1998. http://bill.publico.pt/nos/livroestilo.html

Andrei Mikheev, Marc Moens & Claire Grover (1999). Named Entity Recognition with Gazetteers. *Proceeding of EACL'99* (Bergen, 8-12 June 1999), ACL, pp. 1-8.

Diana Santos (1998). Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. Antonio Rubio, Natividad Gallardo, Rosa Castro & Antonio Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 1, pp. 475-481.

Diana Santos & Eckhard Bick (2000). Providing Internet access to Portuguese corpora: the AC/DC project. *Proceedings of the Second International Conference on Language Resources and Evaluation* (Athens, 31 May-2 June 2000), vol. 1 pp. 205-210.

Diana Santos & Elisabete Ranchhod (1999). Ambientes de processamento de corpora em português: Comparação entre dois sistemas. *Actas do IV Encontro sobre o Processamento Computacional em Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro de 1999), pp. 257-268.

Anexo: Exemplo de um extracto

```
<ext n=1360003 sec=pol sem=94b>
<p><s>Para os comunistas, a questão de princípio é manter a Constituição
tal como está, recusando liminarmente qualquer alteração ao texto.</s>
<s>E é também nesta linha que rejeitam a tentativa de Cavaco Silva em
retirar do articulado a regionalização como um objectivo da construção
do Estado democrático.</s></p>
<p><s>Também neste aspecto se advinha que os dois partidos parlamentares
mais pequenos vão ter estratégias diversas.</s> <s>Enquanto para o PCP a
regionalização é mais um pretexto para atacar os adversários, o CDS
desvaloriza-a, sendo claro na aposta em mecanismos de auscultação da
opinião dos cidadãos.</s> <s>Nesta medida se compreende que os
centristas prefiram agitar um debate em torno do referendo sobre a
regionalização, em vez de esgrimir argumentos sobre a sua manutenção ou
não no texto constitucional.</s> <s>Uma posição que não é estranha à
vertente populista da actual direcção.</s></p>
<a>J.P.</a></ext>
```