Capítulo 20

Disponibilizando a <OBRA>Colecção Dourada</OBRA> do <ACONTECIMENTO> HAREM </ACONTECIMENTO> através do projecto <LOCAL | ORGANIZACAO | ABSTRACCAO> AC/DC </LOCAL | ORGANIZACAO | ABSTRACCAO>

Paulo Rocha e Diana Santos

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 20, p. 307–326, 2007.

Ao concertar dois projectos caros à Linguateca (o HAREM e o AC/DC) num único recurso, este capítulo tem dois objectivos distintos:

1. Disponibilizar a colecção dourada do HAREM num formato mais amigável para a sua exploração por uma comunidade mais abrangente, e apresentar alguma informação quantitativa que permitirá avaliar a dificuldade subjacente ao Primeiro HAREM;

2. Produzir documentação mais actualizada sobre o projecto AC/DC, descrevendo como codificar (e consequentemente usar) outro tipo de informação (a que chamamos informação estrutural) a partir de uma colecção anotada, e cujo processo até agora nunca tinha sido descrito em pormenor.

Este capítulo começa por descrever brevemente o projecto AC/DC, explicando os motivos para disponibilizar a colecção dourada como um corpus. De seguida, é feita uma pequena introdução ao formalismo subjacente ao AC/DC, para explicar as opções tomadas na codificação da colecção dourada (ilustradas com exemplos de procuras não triviais no âmbito do corpus CDHAREM). O capítulo termina por uma descrição quantitativa da colecção dourada (e das colecções douradas parciais que foram usadas em 2005 e 2006), de forma a contribuir para uma caracterização e medição rigorosas do problema que os sistemas tentaram resolver no HAREM.

20.1 O projecto AC/DC

O projecto AC/DC, Acesso a Corpora/Disponibilização de Corpora (Santos e Bick, 2000; Santos e Sarmento, 2003) é um projecto que pretende facilitar o acesso a corpora em português, tanto para o utilizador casual, como para o investigador na área. O AC/DC disponibiliza todos os corpora que a Linguateca possui num ponto único de acesso , num formato pensado para ser usado por seres humanos.

Este projecto teve início em 1998, e o número de corpora disponibilizados tem crescido sustentadamente desde essa data; actualmente, é possível consultar no sítio do AC/DC (`http://www.linguateca.pt/ACDC/`) cerca de vinte corpora, através de uma interface simples e padronizada. Estes corpora, na sua maioria criados por entidades exteriores à Linguateca, abrangem vários géneros textuais e proveniências, e incluem alguns de grande dimensão, nomeadamente o CETEMPúblico (Rocha e Santos, 2000) com mais de 180 milhões de palavras de texto jornalístico em português europeu, e o Corpus NILC/São Carlos, com mais de 32 milhões de palavras em português do Brasil, bem como outros corpora de menor dimensão mas geralmente com mais informação linguística associada. Embora não fazendo estritamente parte do AC/DC, convém referir que também o COMPARA (Frankenberg-Garcia e Santos, 2002), um corpus paralelo de textos literários em português e inglês, e a Floresta Sintá(c)tica (Bick et al., 2007) se podem considerar continuadores do

AC/DC, no sentido de que resultam de uma estratégia de enriquecimento deste, mantendo a filosofia original.

Note-se que os corpora do AC/DC permitem também a criação de outros recursos, como é exemplo a própria Colecção HAREM, em cuja compilação vários corpora do AC/DC foram empregues, ou a colecção dourada usada nas Morfolimpíadas (Santos et al., 2003; Costa et al., 2007).

Cremos poder afirmar que o projecto AC/DC tem cumprido a sua missão, ao registar cerca de 6.000 acessos mensais em Abril de 2007, totalizando cerca de 250.000 acessos desde o seu início.

20.1.1 A criação de um corpus novo no AC/DC

Os corpora, como simples conjunto de textos, só permitem realizar consultas simples, como, por exemplo, verificar as concordâncias de uma determinada unidade no corpus e quantas vezes ocorre. Assim, de modo a permitir consultas mais elaboradas, os corpora do AC/DC são enriquecidos com informação adicional relevante.

Em primeiro lugar, os corpora são anotados gramaticalmente com o analisador sintáctico PALAVRAS (Bick, 2000), que adiciona informação complementar, tal como o lema ou a categoria gramatical de cada palavra existente nos corpora, o género ou o tempo verbal, ou a função sintáctica dos vários constituintes.

De igual modo, aplicam-se a todos os corpora procedimentos sistemáticos e rigorosos de atomização e separação de frases em português[1]. São também geradas listas de formas e lemas presentes nos corpora.

Além disso, alguns corpora são marcados com anotações adicionais, como por exemplo o período de tempo a que se referem, o país de origem ou a fonte dos textos, permitindo restringir as procuras a uma subsecção do corpus. As anotações utilizadas pelo corpus da CD do HAREM são descritas na secção 20.2.[2]

20.1.2 IMS-CWB, o sistema subjacente

Os corpora são compilados usando o IMS Corpus Workbench ou IMS-CWB[3](Christ et al., 1999; Evert, 2005), que se revelou robusto e eficiente para os nossos propósitos (Santos e Ranchhod, 1999). O IMS-CWB é detentor de uma linguagem poderosa de interrogação de corpora através do seu módulo *Corpus Query Processor* (CQP), permitindo codificar a informação associada a um corpus de duas formas complementares: atributos estruturais e atributos posicionais.

[1] No sítio do AC/DC pode ser encontrada informação mais detalhada sobre os critérios de separação em frases e sobre as ferramentas usadas para essa tarefa.

[2] Para mais informação sobre os outros corpora, consulte as páginas do AC/DC.

[3] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

Os **atributos estruturais** referem-se às etiquetas usadas no texto para marcar ou delimitar um subconjunto desse texto. No exemplo abaixo, as etiquetas PESSOA e OBRA são transformadas em atributos estruturais homónimos.

```
Entre as propostas mais ousadas, decidiu-se pedir ao <PESSOA TIPO="CARGO"
MORF="M,S"> Presidente da República </PESSOA> que proponha um referendo
sobre a <OBRA TIPO="PUBLICACAO" MORF="F,S"> Lei do Financiamento </OBRA>
```

Aos atributos estruturais podem ser associados valores, como por exemplo, <DOCID id=¨899¨>. Estes valores podem ser usados na restrição de uma consulta, mas não podem ser alvo de consultas de distribuição.[4]

Os **atributos posicionais** contêm valores que se atribuem a cada unidade no texto. Usando como exemplo o início da frase anterior e o atributo posicional pos (categoria gramatical, do inglês *part-of-speech*), obtemos a seguinte linha de texto:

```
Entre/PRP as/DET propostas/N mais/ADV ousadas/ADJ, ...
```

Uma descrição mais completa da sintaxe usada no IMS-CWB pode ser encontrada em http://acdc.linguateca.pt/acesso/anotacao.html. Recomendamos vivamente a sua leitura, visto que reúne informação considerável sobre o uso específico do PALAVRAS como anotador no AC/DC e sobre o pós-processamento a que a anotação deste é sujeita. Essa página também remete para vários outros locais de ensino do CQP. Mencione-se, a propósito, que a anotação do PALAVRAS também é a base da parte portuguesa do COMPARA (Santos e Inácio, 2006; Inácio e Santos, 2006) e da Floresta Sintá(c)tica (Afonso et al., 2002; Afonso, 2006), ambas revistas posteriormente.

20.2 Disponibilizando a CD do HAREM como corpus

Apesar de as CD estarem publicamente disponíveis do sítio do HAREM desde o fim das respectivas avaliações conjuntas[5], optámos por disponibilizá-las igualmente como um recurso no âmbito do AC/DC, facilitando assim o acesso à riqueza de informação associada à colecção e adicionando informação morfossintáctica. Tal permite um número de pesquisas na colecção que não seriam fáceis ou práticas de efectuar por um utilizador menos experimentado.

O corpus resultante, denominado CDHAREM, é então formado pelo texto das duas CD usadas nas duas avaliações conjuntas do HAREM, acrescido de toda a informação anexa a esse texto e da sua anotação gramatical.

[4] Ao contrário das concordâncias, onde se pede o texto, nas consultas de distribuição (ou consultas agregadas, em terminologia de bases de dados) pretende-se habitualmente saber a quantidade de vezes que um dado fenómeno ocorre, e qual a distribuição quantitativa dos elementos que satisfazem a procura em questão.

[5] Em http://poloxldb.linguateca.pt/harem.php?l=coleccaodourada

```
Procura:  "Lisboa".
Distribuicao de em
Corpus:  Corpus CD HAREM, 0.1

40 casos.
Distribuição
Houve 11 valores diferentes de em.
```

Lisboa	28
Universidade_de_Lisboa	3
Metropolitano_de_Lisboa	1
Universidade_Nova_de_Lisboa	1
Grande_Lisboa	1
Emissores_Associados_de_Lisboa	1
10h00_de_Lisboa	1
Hotel_Lisboa_Plaza	1
Governadora_Civil_do_Distrito_de_Lisboa	1
Instituto_Técnico_de_Lisboa	1
Departamento_de_Matemática_da_Universidade_de_Lisboa	1

Figura 20.1: Distribuição de uma palavra por EM.

20.2.1 Opções gerais de codificação

Na tabela 20.1 apresentamos, de forma condensada e para referência subsequente, a lista de conversões de atributos presente na CD para os formatos usados no AC/DC, com o objectivo de facilitar vários tipos de pesquisa, que nos parecerem especialmente relevantes neste contexto. Como norma geral, para o corpus CDHAREM, foram usadas letras maiúsculas para os atributos estruturais, e minúsculas para os atributos posicionais. A única excepção foram os atributos estruturais `<p>` e `<s>`, provenientes da separação de frases. Caso um atributo posicional não se encontre definido para uma determinada unidade, é-lhe atribuído o valor "0".

20.2.2 O atributo EM

O atributo estrutural EM, como o seu próprio nome indica, identifica uma EM, independentemente da sua classificação. A consulta seguinte encontra exclusivamente a EM *Porto*, excluindo assim os casos em que esta palavra faz parte de uma EM maior (por exemplo, *Porto Seguro*):

```
<EM> "Porto" </EM>
```

Na próxima consulta encontramos os casos em que a expressão *São Paulo* é parte de uma EM:

Tipo de atributo	Colecção dourada	Atributo estrutural	Valores	Atributo posicional	Valores
Delimitador de um documento	DOC	DOC	docid= genero= origem=	–	–
Identificação do documento da CD	DOCID...	-	-	-	HAREM-871-07800, etc.
Género de texto	GENERO...	-	-	-	Web, Técnico, etc.
País de origem do texto	ORIGEM...	-	-	-	PT, BR, etc.
Delimitador do texto de um documento	TEXTO	TEXTO	tam=	–	–
Entidade mencionada	LOCAL, PESSOA, etc...	EM	tam=	–	–
Categoria(s) a que pertence a palavra	**OBRA** TIPO="ARTE" MORF="M,S", etc.	LOCAL, PESSOA, etc.	–	categoria,	PESSOA, LOCAL, etc...
Tipo(s) a que pertence a palavra	OBRA TIPO="**ARTE**" MORF="M,S", etc.	–	–	tipo, local, pessoa, etc.	ADMNISTRATIVO, INSTITUICAO, etc.
Género e número da EM (revisto manualmente)	OBRA TIPO="ARTE" **MORF**="M,S", etc.	–	–	morf	M, S, F, P, etc.
Posição relativa na EM de uma palavra	–	–	–	prem	1,2,...,29
Delimitador de parágrafo	–	p	–	–	–
Delimitador de frase	–	s	–	–	–
Parte de uma anotação alternativa	<ALT> .... | ... </ALT>	ALT	num=	alt	P, M ou F, seguido da categoria da alternativa, ou de 0. POBRA, FPESSOA, M0, etc.

Tabela 20.1: Conversão de atributos da CD do HAREM para o corpus CDHAREM do AC/DC.

```
"São" "Paulo" within EM
```

O atributo EM é codificado no corpus juntamente com o tamanho (em unidades) da EM, como é ilustrado no exemplo abaixo:

```
<EM TAM=3>
<PESSOA>
Presidente
da
República
</PESSOA>
</EM>
```

Para identificar a EM à qual um termo pertence, pode ser usado o atributo posicional `em`. Este atributo assume como valor o texto da EM, com sublinhados a separar as unidades; no exemplo acima, a cada uma das unidades *Presidente*, *da* e *República* é atribuído o valor *Presidente_da_República*. Pode-se assim mais facilmente descobrir a que EM um termo pertence e quantas vezes, tal como no exemplo da Figura 20.1.

20.2.3 Atributos relativos às categorias e tipos das EM

Todas as categorias existentes na CD equivalem a um atributo estrutural distinto. Estes atributos podem ser usados para facilitar a procura de uma determinada categoria de EM; por exemplo, para obter todas as EM de categoria `OBRA`:

```
<OBRA> []* </OBRA>
```

ou todas as EM de três palavras que sejam simultaneamente `ORGANIZACAO` e `LOCAL`:

```
<ORGANIZACAO> <LOCAL> [] [] [] </LOCAL> </ORGANIZACAO>
```

Para facilitar as consultas, usam-se também atributos posicionais para identificar as categorias e tipos, apropriadamente chamados `categoria` e `tipo` respectivamente. O exemplo seguinte mostra os valores do atributo `categoria` para um excerto particular.

```
<s> As/0 ilhas/0 de/0 Cabo/LOCAL Verde/LOCAL foram/0 descobertas/0 por/0
navegadores/0 portugueses/0 em/0 Maio/TEMPO de/TEMPO 1460/TEMPO ,/0 sem/0
indícios/0 de/0 presença/0 humana/0 anterior/0 ./0 </s>
```

No caso de uma EM pertencer a múltiplas categorias ou tipos, eles são listados por ordem alfabética, separados por sublinhados (ver secção 20.3.1).

Além disso, foi definido um atributo posicional para cada uma das categorias, que assumem o valor do tipo correspondente à EM. Os atributos posicionais têm o mesmo

nome dos estruturais, mas em minúsculas (`local`, `pessoa`, etc.). Assim, podemos procurar a palavra *Lisboa* como parte do nome de uma organização mas não parte do nome de um local (o valor "0" implica que o campo não tem um valor definido):

```
[word="Lisboa" & organizacao!="0" & local="0"]
```

Assim como podemos identificar os casos em que à categoria `PESSOA` corresponde o tipo `CARGO` (independentemente de outros):

```
<PESSOA> [pessoa=".*CARGO.*"]+ </PESSOA>
```

Se se quisesse apenas os casos em que `CARGO` é o único tipo, empregar-se-ia a seguinte expressão de consulta:

```
<PESSOA> [pessoa="CARGO"]+ </PESSOA>
```

20.2.4 O atributo `prem` para compatibilizar contagens por palavras e por EM

Um atributo posicional importante que foi inserido no corpus CDHAREM é o atributo `prem` (posição relativa na EM), que identifica o número de ordem de uma palavra dentro de uma EM. O atributo `prem` assume o valor "0" no caso de a palavra não pertencer a nenhuma EM.

Podemos usar este atributo também para identificar os casos em que *São Paulo* é a parte final de uma EM maior:

```
[word="São" & prem!="1" & prem!="0"] "Paulo"
```

Ou, pelo contrário, a parte inicial de uma EM maior:

```
"São" "Paulo" [prem="3"]
```

Assim como obter os casos de *Porto* que não fazem parte de uma EM.

```
[word="Porto" & prem="0"]
```

A maior utilidade deste atributo é permitir restringir as consultas de distribuição apenas às EM, e que devem ser feitas apenas sobre a primeira palavra de cada EM (ou seja, em que o valor de `prem` seja igual a 1), para que as outras palavras da EM não influenciem o resultado (senão, uma EM com cinco palavras contaria cinco vezes).

20.2.5 Atributos relativos ao texto

As etiquetas que delimitam documentos da CD (<DOC> e </DOC>) e os respectivos textos (<TEXTO> e </TEXTO>) foram convertidas no CDHAREM em atributos estruturais. à etiqueta <DOC> foi adicionada a informação constante das etiquetas <DOCID>, <GENERO> e <ORIGEM>, que não foram incluídas no corpus; à etiqueta <TEXTO> foi adicionado o tamanho do excerto, como se pode ver no exemplo abaixo.

```
<DOC docid=HAREM-871-07800 genero=Web origem=PT>
<TEXTO TAM=279>
```

Foram adicionados ainda outros três atributos posicionais com informação constante nas etiquetas removidas, e relativos ao documento propriamente dito:

- docid, a identificação do documento na colecção, no formato especificado no capítulo 19;
- genero, o tipo de texto, que pode ter um dos seguintes valores: Jornalístico, Web, CorreioElectrónico, Entrevista, Expositivo, Literário, Político, Técnico;
- origem, dado pelo código ISO do país de origem do texto: PT (Portugal), BR (Brasil), AO (Angola), MZ (Moçambique), CV (Cabo Verde), MO (Macau), IN (Índia) ou TL (Timor-Leste)[6].

Estes atributos posicionais, gerados a partir das etiquetas homónimas, podem ser usados, por exemplo, para identificar todas as pessoas assinaladas em texto jornalístico brasileiro:

```
<PESSOA> [origem="BR" & genero="Jorn.*"]* </PESSOA>
```

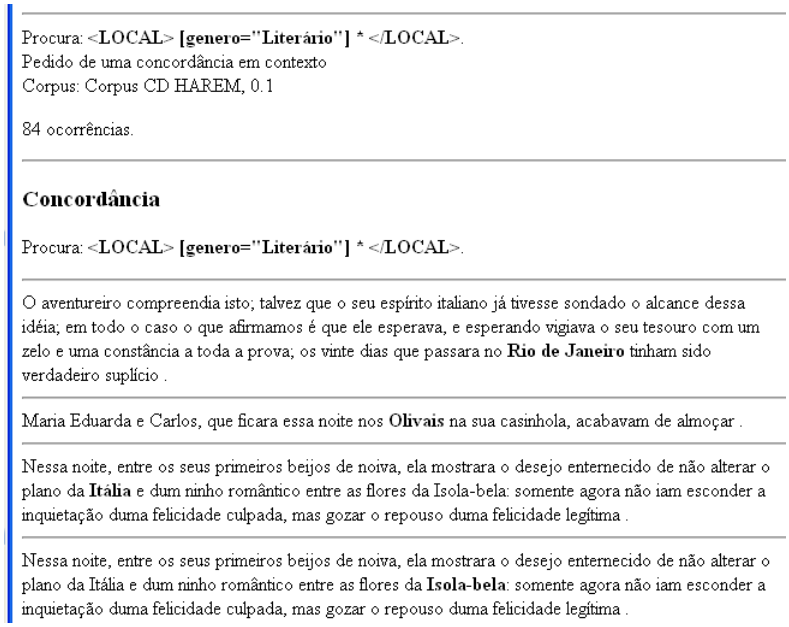
Escolhendo a distribuição das EM por categoria, podemos ver a distribuição das EM em texto técnico (note-se o uso de prem para que só uma palavra de cada EM seja contabilizada):

```
[genero="Técnico" & prem="1"]
```

Refinando ainda mais esta consulta, podemos seleccionar a distribuição por tipo apenas das EM da categoria COISA em texto técnico:

```
[genero="Técnico" & prem="1" & coisa!="0"]
```

[6] Embora existam textos de São Tomé e Príncipe (ST) e da Guiné-Bissau (GW) na colecção do HAREM, estes não aparecem nas colecções douradas.



Procura: <LOCAL> [genero="Literário"] * </LOCAL>.
Pedido de uma concordância em contexto
Corpus: Corpus CD HAREM, 0.1

84 ocorrências.

Concordância

Procura: <LOCAL> [genero="Literário"] * </LOCAL>.

O aventureiro compreendia isto; talvez que o seu espírito italiano já tivesse sondado o alcance dessa idéia; em todo o caso o que afirmamos é que ele esperava, e esperando vigiava o seu tesouro com um zelo e uma constância a toda a prova; os vinte dias que passara no **Rio de Janeiro** tinham sido verdadeiro suplício .

Maria Eduarda e Carlos, que ficara essa noite nos **Olivais** na sua casinhola, acabavam de almoçar .

Nessa noite, entre os seus primeiros beijos de noiva, ela mostrara o desejo enternecido de não alterar o plano da **Itália** e dum ninho romântico entre as flores da Isola-bela: somente agora não iam esconder a inquietação duma felicidade culpada, mas gozar o repouso duma felicidade legítima .

Nessa noite, entre os seus primeiros beijos de noiva, ela mostrara o desejo enternecido de não alterar o plano da Itália e dum ninho romântico entre as flores da **Isola-bela**: somente agora não iam esconder a inquietação duma felicidade culpada, mas gozar o repouso duma felicidade legítima .

Figura 20.2: Exemplo de concordância: locais referidos em texto literário (excerto)

20.2.6 Atributos relativos à classificação morfológica

A informação morfológica da CD do HAREM foi mantida no CDHAREM com a ajuda do atributo posicional `morf`. Desta forma, podemos por exemplo procurar todas as referências a pessoas do sexo feminino na CDHAREM:

```
<PESSOA> [tipo="INDIVIDUAL" & morf="F,S"]+ </PESSOA>
```

ou pedir a distribuição por género e número da categoria dos acontecimentos.

```
<ACONTECIMENTO> []
```

20.2.7 Atributos relativos à anotação sintáctica do AC/DC

Foram também adicionados atributos estruturais relativos aos parágrafos (`<p>`) e às frases (`<s>`). Podemos assim, por exemplo, pedir ao serviço AC/DC todas as frases contendo a palavra *Luanda*.

```
<s> []* "Luanda" []* </s>
```

Por fim, existe a informação gramatical acrescentada pelo analisador sintáctico PALAVRAS. Esta informação é gerada automaticamente e não foi, até agora, revista manualmente – para avaliações parciais do desempenho do PALAVRAS, veja-se Bick (2000), Santos e Gasperin (2002) ou Santos e Inácio (2006) – mas permite fazer consultas poderosas

desde que se tome esse facto em consideração. Um exemplo pode ser a distribuição das EM por função sintáctica:

```
[prem="1"]
```

ou das EM da categoria PESSOA como sujeito de um verbo de locução:

```
<PESSOA> [func="SUBJ"]* </PESSOA> [lema="dizer|afirmar|relatar"]
```

Pode-se também combinar numa consulta atributos de fontes diferentes, ou seja, atributos posicionais vindos do HAREM e da anotação gramatical automática, como o demonstra a seguinte procura de EM precedidas por um adjectivo:

```
[pos="ADJ" & prem="0"] [prem="1"]
```

20.3 Vagueza

Como várias vezes referido neste livro e noutras publicações (Santos et al., 2006), a codificação explícita da vagueza é um dos pontos fortes do HAREM.

20.3.1 Vagueza na classificação (categorias ou tipos com |)

Um total de 271 EM (2,9% do total das EM da CD) apresentavam anotações alternativas (66 entre tipos da mesma categoria, 202 entre duas categorias distintas e 3 entre três categorias), embora contendo exactamente as mesmas palavras. Nestes casos, as anotações foram mantidas e as EM foram assinaladas no CDHAREM com todas as suas categorias e tipos.

Casos como:

```
<PESSOA|ORGANIZACAO TIPO="GRUPOCARGO|SUB" MORF="F,S">
Convenção
</PESSOA|ORGANIZACAO>
```

foram, em termos de atributos estruturais, codificados como

```
<PESSOA TIPO="GRUPOCARGO" MORF="F,S">
<ORGANIZACAO TIPO="SUB" MORF="F,S">
Convenção
</ORGANIZACAO>
</PESSOA>
```

Assim sendo, apenas as dez categorias simples de EM estão codificadas directamente em atributos posicionais e estruturais. Para encontrar EM classificadas como pertencendo a múltiplas categorias, há várias maneiras possíveis de efectuar a consulta:

```
<PESSOA> <ORGANIZACAO> []
<ORGANIZACAO> <PESSOA> []
[pessoa!="0" & organizacao="0"]
[categoria="ORGANIZACAO-PESSOA"]
```

20.3.2 Vagueza na identificação: as etiquetas <ALT>

Uma vez que o formato usado para as etiquetas `<ALT>` leva à repetição dos textos das EM na CD, tivemos de proceder a algum processamento adicional de forma a codificar as anotações alternativas assinaladas com estas etiquetas.

Há um total de 122 etiquetas `<ALT>` na CD que foram identificadas na CDHAREM com o atributo posicional `alt`, contendo um valor diferente de 0.

De momento, codificámos a primeira alternativa, indicando o número de alternativas como valor do atributo estrutural `<ALT>`, bem como o valor de `alt` à categoria ou categorias das alternativas, iniciada por P M ou F (princípio, meio e fim). Quando a alternativa fosse nula (não pertencesse a EM), considerámos 0 como nome da categoria. Quando o princípio, meio e fim coincidissem, marcámos sempre primeiro o princípio, seguido de meio e só em último lugar do fim.

Seguem alguns exemplos ilustrativos:

```
... no jogo <ALT> <ACONTECIMENTO TIPO="EVENTO" MORF="M,S"> Académica-Benfica
</ACONTECIMENTO> | <PESSOA TIPO="GRUPOMEMBRO"> Académica </PESSOA> - <PESSOA
TIPO="GRUPOMEMBRO"> Benfica </PESSOA> </ALT>.

<ALT num=2>
<ACONTECIMENTO>
Académica PPESSOA
-         P0
Benfica   PPESSOA
</ACONTECIMENTO>
</ALT>

<ALT> Governo de <PESSOA TIPO=INDIVIDUAL>Cavaco Silva</PESSOA> | <ORGANIZACAO|
PESSOA TIPO=ADMINISTRACAO|GRUPOIND> Governo de Cavaco Silva</ALT>

<ALT num=2>
Governo   PORGANIZACAO
de        MORGANIZACAO
<PESSOA>
Cavaco    MORGANIZACAO
```

```
Silva     FORGANIZACAO
</PESSOA>
</ALT>

Um pouco de <ALT> HISTÓRIA | <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">
HISTÓRIA </ABSTRACCAO> </ALT>

<ALT num=2>
HISTÓRIA PABSTRACCAO
</ALT>
```

Em termos de procuras possibilitadas pelo AC/DC, além de podermos observar que sequências alternativas foram consideradas <ALT> nas CD:

```
<ALT> []+ </ALT>
```

podemos também localizar os casos em que a palavra *Governo* faz parte, na CD, de uma EM alternativa à assinalada no corpus:

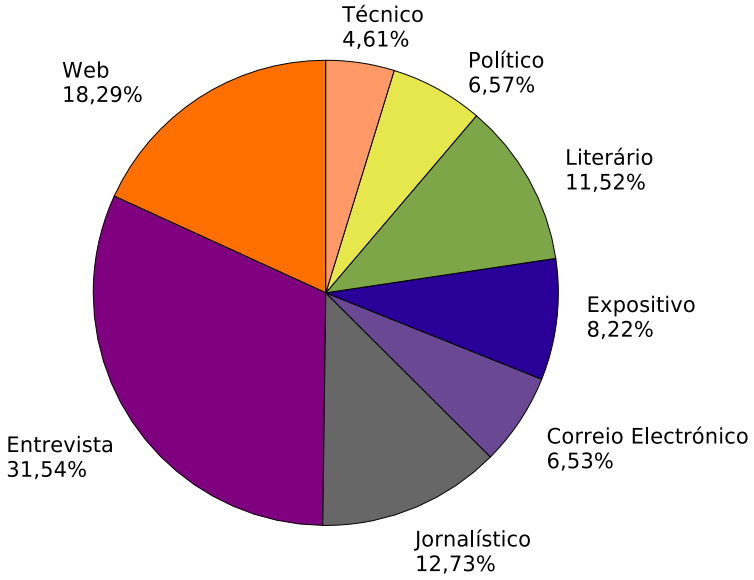
```
[word="Governo" & alt!="0"]
```

	Documento	Parágrafo	Frase
Média de número de entidades por	33,65	1,71	1,05
Mediana do número de entidades por	30	1	0
Número máximo de EM num	205	9	9
Número mínimo de EM num	2	0	0
Unidades textuais com 0 EM	0%	40,5%	50,3%
Unidades textuais com 1 EM	0%	25,4%	24,6%
Unidades textuais com 2 EM	1,9%	13,8%	12,4%
Unidades textuais com 3 EM	0%	7,0%	5,6%
Unidades textuais com 4 EM	2,3%	4,1%	3,1%
Unidades textuais com 5 a 9 EM	7,8%	6,9%	3,7%
Unidades textuais com 10 a 19 EM	24,0%	2,0%	0,3%
Unidades textuais com 20+EM	63,6%	0,4%	0,1%

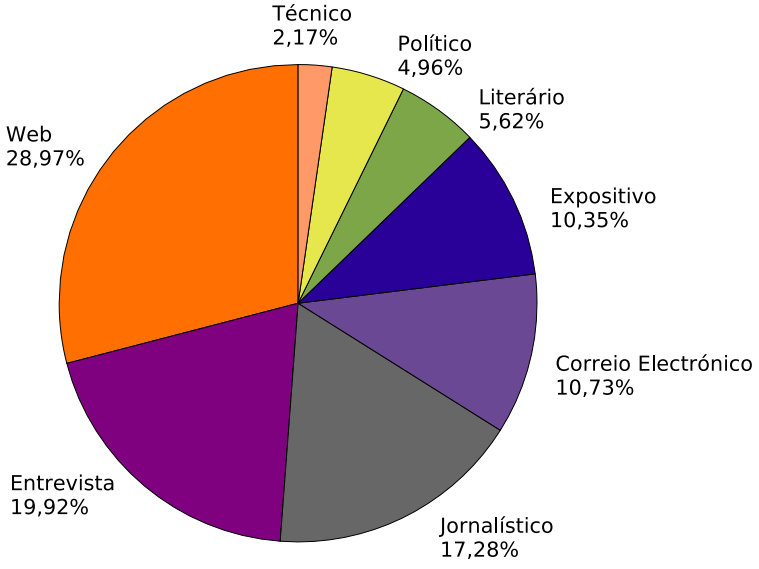
Tabela 20.2: Distribuição da quantidade de EM por unidades de texto.

20.4 Dados quantitativos

Segundo as normas de atomização do projecto AC/DC, o CDHAREM contém 154.863 unidades (133.569 das quais palavras, 86,3%), incluindo 8.976 EM que abrangem 17.206 unidades (16.821 das quais palavras, 97,2%).



(a) Em função do número de unidades.



(b) Em função do número de EM.

Figura 20.3: Distribuição por género dos termos existentes nas CD.

Quanto a EM, o CDHAREM apresenta um total de 8.967 EM (menos 463 que as CD originais, devido à nossa escolha relativa aos ALT), distribuídas por 8.184 frases (incluindo 990 fragmentos), agrupadas em 5.062 parágrafos e oriundas de 257 documentos distintos. Na Tabela 20.2, encontra-se uma distribuição quantitativa das EM por texto, por parágrafo e por frase.

Como mencionado acima, os documentos da CD foram classificados como pertencentes a oito géneros distintos de texto. A Figura 20.3(a) mostra a repartição dos textos da CD em função do número de unidades, enquanto que a Figura 20.3(b) mostra a repartição em função do número de EM, elucidando as diferenças em termos de densidade de EM em função do género literário: certos géneros são mais ricos (ou mais pobres) em EM do que outros.

Como se pode ver na Figura 20.4, as categorias de EM mais frequentes são `LOCAL` e `PESSOA`, que entre si cobrem quase metade das EM.

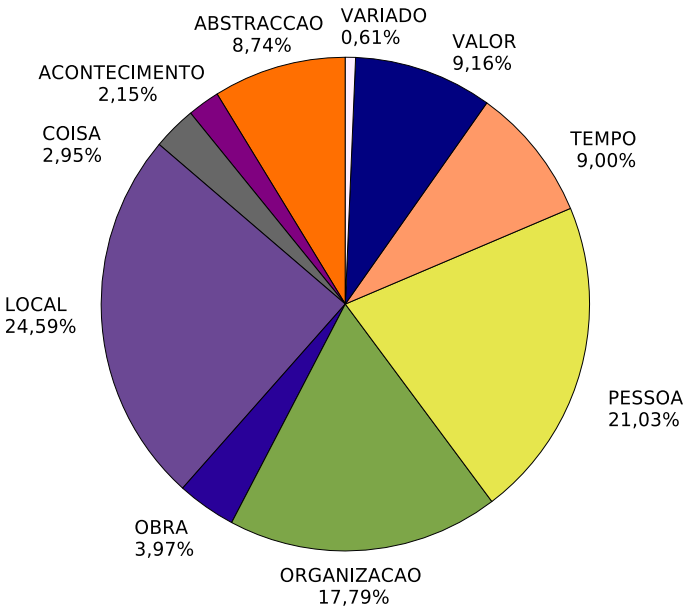


Figura 20.4: Distribuição das categorias semânticas de EM nas CD (sem peso).

As Figuras 20.5 e 20.6 mostram a relação entre as diferentes categorias de EM e os diversos géneros de texto.

Uma análise semelhante é feita em termos de variante, mas dado que a contribuição de textos em português não oriundos nem de Portugal nem do Brasil foi ínfima, considerámos apenas estas duas variantes na análise apresentada nas Tabelas 20.3 e 20.4 (correspondente assim a 251 textos, 150.041 unidades e 8.339 EM).

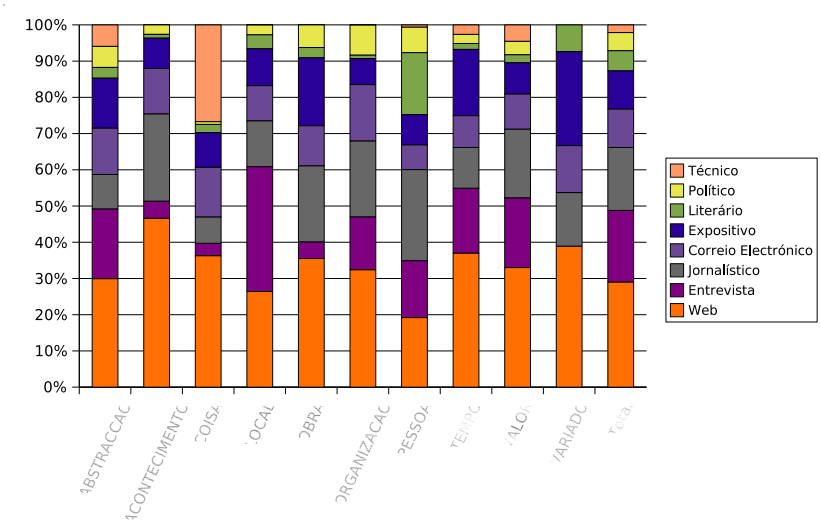


Figura 20.5: Distribuição das categorias semânticas de EM por género textual nas CD (sem peso).

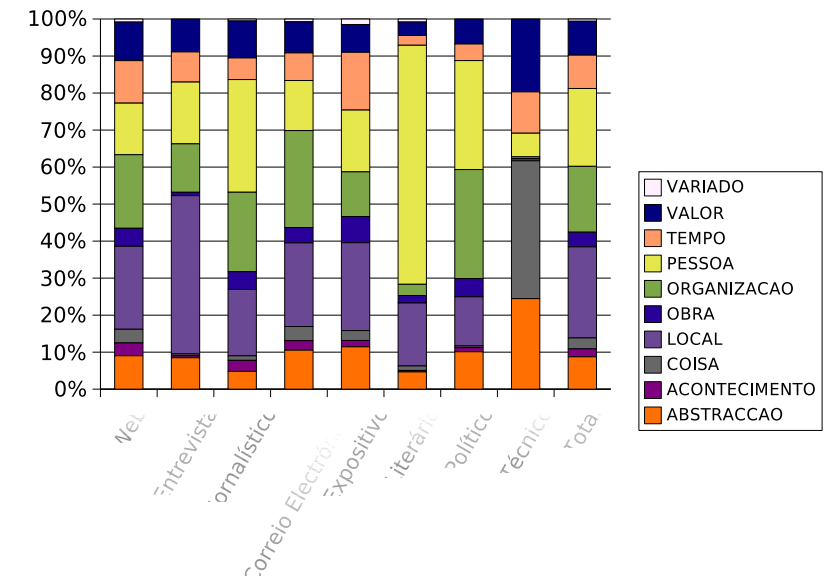


Figura 20.6: Distribuição do género textual das categorias semânticas de EM nas CD (sem peso).

A Tabela 20.5 apresenta a distribuição das categorias de entidades mencionadas na colecção dourada, repetindo em forma tabular a informação da figura 20.3.

A Tabela 20.6 apresenta o tamanho das entidades mencionadas em número de pala-

Categoria	Brasil	%	Portugal	%
ABSTRACCAO	364	49%	372	51%
ACONTECIMENTO	90	48%	96	52%
COISA	156	60%	104	40%
LOCAL	1.099	53%	987	47%
OBRA	147	46%	174	54%
ORGANIZACAO	785	51%	747	49%
PESSOA	920	51%	898	49%
TEMPO	349	45%	423	55%
VALOR	446	56%	354	44%
VARIADO	41	77%	12	23%
Total	4397	51%	4167	49%

Tabela 20.3: Distribuição das categorias semânticas por variante na CD (contando independentemente todas as classificações: EM pertencentes a múltiplas categorias são contadas para cada categoria).

Categoria	EM sem peso		EM com peso	
	Brasil	%	Portugal	%
ABSTRACCAO	7,6%	9,2%	7,9%	8,7%
ACONTECIMENTO	1,9%	1,8%	1,9%	2,3%
COISA	3,6%	2,7%	3,6%	2,5%
LOCAL	24,9%	23,3%	25,0%	23,9%
OBRA	3,2%	3,9%	3,2%	4,1%
ORGANIZACAO	17,4%	18,1%	17,7%	17,8%
PESSOA	21,5%	21,8%	21,4%	21,3%
TEMPO	8,2%	10,2%	8,0%	10,3%
VALOR	10,6%	8,5%	10,4%	8,7%
VARIADO	1,0%	0,3%	0,9%	0,3%

Tabela 20.4: Distribuição por variante das categorias semânticas na CD ; "EM sem peso" contam cada EM por cada categoria a que pertence; "EM com peso" contabilizam cada EM uma única vez atribuindo uma fracção a cada uma das suas categorias.

vras. Como se pode ver, mais de metade das EM contêm uma única palavra. A EM mais comprida (o título de uma palestra) contém 29 palavras.

Na Tabela 20.7 apresenta-se o tamanho médio das EM em número de palavras e a percentagem de EM simples (i.e., contendo uma única palavra), por categoria, e por cada variante. Todas as categorias de EM têm uma moda de 1 palavra, com excepção da categoria ACONTECIMENTO, onde a moda é de 3 palavras.

A Tabela 20.8 mostra a distribuição morfológica das EM em geral, e a Tabela 20.9 a mesma por categoria semântica. É interessante constatar a maioria esmagadora de entidades singulares.

A Tabela 20.10 mostra diferentes vertentes, permitindo uma primeira quantificação da

Categoria	CD 2005	CD 2006	Total	%
ABSTRACCAO	449	326	775	8,7%
ACONTECIMENTO	128	63	191	2,2%
COISA	82	180	262	3,0%
LOCAL	1.286	895	2.181	24,6%
OBRA	222	130	352	4,0%
ORGANIZACAO	956	622	1.578	17,8%
PESSOA	1.029	836	1.865	21,0%
TEMPO	434	364	798	9,0%
VALOR	484	328	812	9,2%
VARIADO	40	14	54	0,6%
Total	5.110	3.758	8.868	100,0%

Tabela 20.5: Distribuição das categorias de EM na CD.

Nº palavras	CD 2005	CD 2006	Total	%	Exemplo
1	2.769	2.052	4.821	54,3%	*Brasil*
2	1.049	888	1.937	21,8%	*São Paulo*
3	706	421	1.127	12,7%	*Universidade do Minho*
4	255	178	433	4,9%	*Rua 25 de Março*
5	165	94	259	2,9%	*25 de Abril de 1974*
6	48	36	84	0,9%	*Governador do Rio Grande do Norte*
7	46	22	68	0,8%	*26ª jornada da II Divisão de Honra*
8	20	12	32	0,4%	*Lei Antitruste (nº 8.884 / 94)*
9	19	12	31	0,3%	*Band of Gypsies: Live at the Fillmore East*
10+	38	43	81	0,9%	
Total	5.115	3.758	8.873	100,0%	

Tabela 20.6: Tamanho em número de palavras das EM.

Categoria	Texto completo			Textos brasileiros			Textos portugueses		
	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras
ABSTRACCAO	2,2	51%	5%	2,7	46%	8%	1,3	56%	1%
ACONTECIMENTO	3,7	20%	16%	4,0	26%	20%	3,4	17%	11%
COISA	1,4	72%	<1%	1,5	71%	1%	1,3	73%	0%
LOCAL	1,7	68%	2%	1,8	61%	2%	1,5	74%	1%
OBRA	3,4	26%	13%	3,8	22%	17%	3,1	33%	11%
ORGANIZACAO	2,2	57%	6%	2,0	61%	4%	2,4	54%	9%
PESSOA	2,0	41%	2%	1,9	44%	1%	2,0	38%	2%
TEMPO	1,7	69%	<1%	1,7	67%	<1%	1,7	71%	<1%
VALOR	1,7	46%	<1%	1,8	43%	1%	1,7	50%	0%
VARIADO	1,9	69%	6%	1,9	71%	7%	2,2	58%	0%
Todas as categorias	2,0	54%	3%	2,0	53%	3%	1,9	55%	3%

Tabela 20.7: Informação sobre o tamanho das EM em número de palavras por categoria

	S	P	?	Total
M	3713	214	0	3.927
F	2565	83	0	2.648
?	543	1	94	638
Total	6.821	298	94	7.213
	Sem classificação			1.655

Tabela 20.8: Informação morfológica sobre as EM em geral

Categoria	M	F	?	S	P	?	s/class.
ABSTRACCAO	292 (38%)	418 54%)	54 (7%)	686 (89%)	59 (8%)	19 (2%)	11 (1%)
ACONTECIMENTO	102 (53%)	76 (40%)	13 (7%)	174 (91%)	16 (8%)	1 (<1%)	0 (0%)
COISA	183 (70%)	41 (16%)	33 (13%)	198 (75%)	38 (15%)	21 (8%)	5 (2%)
LOCAL	978 (45%)	750 (34%)	352 (16%)	2022 (93%)	46 (2%)	12 (1%)	101 (5%)
OBRA	188 (53%)	98 (26%)	58 (16%)	301 (85%)	20 (6%)	18 (5%)	13 (4%)
ORGANIZACAO	695 (44)	819 (52%)	58 (4%)	1524 (97%)	44 (3%)	4 (<1%)	6 (<1%)
PESSOA	1384 (74%)	431 (23%)	48 (3%)	1798 (96%)	61 (3%)	4 (<1%)	2 (<1%)
TEMPO	75 (9%)	13 (2%)	2 (<1%)	83 (10%)	7 (1%)	0 (0%)	708 (89%)
VARIADO	23 (43%)	5 (9%)	20 (37%)	30 (56%)	3 (6%)	15 (28%)	6 (11%)
Todas as categorias	44,5%	29,9%	7,1%	76,9%	3,6%	1,0%	18,5%

Tabela 20.9: Informação morfológica sobre as EM por categoria semântica.

dificuldade associada à tarefa descrita pela colecção dourada do HAREM, em particular:

- o número de palavras em maiúscula na colecção e quantas faziam parte de uma EM;
- o número de unidades pertencentes a EM que fazem parte de EM distintas (excluindo números e sinais de pontuação);
- o número de EM que tiveram diferentes classificações em contexto (dentre as EM que aparecem mais do que uma vez);
- o número de palavras (independentemente de estarem em maiúsculas ou minúsculas) que aparecem na colecção tanto fora como dentro de EM (excluindo números e sinais de pontuação);
- quantas palavras pertencentes a EM têm categorias distintas (excluindo números e sinais de pontuação).

20.5 Observações finais

A conversão da CD do HAREM para o corpus CDHAREM do AC/DC teve como principais objectivos produzir um recurso de maior qualidade e de mais fácil acesso, e disponibilizar uma ferramenta que permita preparar, com mais conhecimento empírico do problema,

Questão	Valores absolutos	%
Palavras em maiúscula	5.191 em 14.705	35,3%
Palavras distintas pertencentes a várias EM	1.655 em 5.453	30,6%
EM que ocorrem mais do que uma vez e com várias interpretações	360 em 4996	7,2%
Palavras distintas dentro de EM que também aparecem fora	1.337 em 4.455	30,0%
Palavras pertencentes a EM de categorias distintas	862 em 4.455	20,8%

Tabela 20.10: Dificuldade da tarefa reflectida na CD do HAREM

próximas edições do HAREM, permitindo medições mais rigorosas do(s) problema(s) que se pretende(m) resolver.

Ao converter a CD num formato mais acessível a linguistas, esperamos também provocar um maior interesse na comunidade linguística sobre o problema de reconhecimento de entidades mencionadas, assim como aproximar projectos como a Floresta Sintá(c)tica e o COMPARA de iniciativas como o HAREM ou as Morfolimpíadas.

Por outro lado, ao desenvolver um esquema que, de certa forma, combina as escolhas da tarefa partilhada do CoNLL (Sang, 2002; Sang e Meulder, 2003), baseadas em palavras – donde, atributos posicionais em formato CQP, e do MUC/HAREM, baseadas em atributos estruturais, mais uma vez usando terminologia do CQP – esperamos poder congregar uma comunidade alargada em redor de uma representação combinada do problema de REM, permitindo comparações finas e informadas entre diferentes abordagens de REM.

Agradecimentos

Estamos muito gratos ao Nuno Cardoso por nos ter facultado as figuras e tabelas constantes do presente capítulo, reproduzidas ou recalculadas da sua apresentação no PROPOR 2006 e na sua tese.

Este capítulo foi escrito no âmbito da Linguateca, integralmente financiado pela Fundação para a Ciência e Tecnologia através do projecto projecto POSC 339/1.3/C/NAC.