A sintaxe do AC/DC: apresentação do CWB e das opções tomadas

Diana Santos

UiO, POR2102, Outono de 2012, v1.2

Para usar os corpos do AC/DC para estudar a gramática da língua portuguesa, há vários assuntos que têm de ser aprendidos primeiro...

Em primeiro lugar, a sintaxe geral do sistema de pesquisa subjacente, o IMS-CWB, agora Open CWB, e a forma como o "simplificámos" para ser usado na rede (Internete).

Depois, o uso específico que nós fizemos do IMS-CWB adicionando mais informação que considerámos apropriada para codificar a língua portuguesa, analisada pelo PALAVRAS. É por isso necessário também explicar algumas das opções deste sistema que usamos como um pilar fundamental.

Como finalmente também adicionámos informação semântica aos corpos, o que implicou ainda mais opções e nomenclatura, esse assunto terá de ser também explicado.

1 Primeiros passos: A sintaxe do CWB

A primeira coisa que um utilizador vê, é uma caixa de procura. O mais simples que pode colocar lá é uma palavra. Mas devido ao poder das expressões regulares (um tipo de linguagens formais muito usado na informática), pode-se procurar ao mesmo tempo muito mais do que uma palavra, desde que se use "." (que significa) qualquer carater, e se indique quantas vezes esse qualquer carater pode aparecer. Para essas ordens, existem três operadores:

? que significa 0 ou 1 vez

* que significa 0 ou mais vezes

\+ que significa 1 ou mais vezes

Experimente procurar

```
casas?
casa.*
casa.+
.*casament.*
```

Para especificar um número determinado, é ainda possível indicar, através de chavetas, o número mínimo e máximo de vezes que se quer um dado carater, por exemplo:

```
cas{0,2}a
```

Além disso, podemos indicar que queremos uma entre várias opções, usando o carater — entre as opções

```
religioso|laico
religios.*|laic.*
António|Antônio
```

No caso em que a diferença é só num único carater, pode usar-se a noção de classe de caracteres, identificada entre parênteses retos:

```
Ant[ôó]nio
c[aeiou]r.
```

Quando os carateres estão numa ordem (alfabética) ininterrupta, pode indicar-se apenas o primeiro e o último, tal como em

```
19[3-7].
[A-Z][a-z]
```

E ainda se pode representar um conjunto através de exclusão, ou seja, descrevendo como "todos os carateres menos uma certa lista". Para isso usa-se o caracter de acento circumflexo como primeiro carater dentro dos parênteses retos:

```
ch[^u]rr.*
[A-Z][^a-z][a-z]
```

Assim, vemos que os carateres ., *, ?, +, , , |, [e] são especiais, assim como os carateres ^ e - dentro do contexto []. Se por acaso os quisermos procurar, temos de indicar que os queremos literalmente, e não no sentido em que os usamos no CWB. Para isso usa-se o carater de barra inclinada para a esquerda, que portanto também é especial.

```
\?
\.\.\.
```

(Repare que codificámos as reticências em português no AC/DC como uma "palavra" com três pontos seguidos, e já agora também os travessões como dois hífenes seguidos.)

Finalmente, as aspas também são especiais, porque delimitam as palavras (ou unidades) que procuramos. Assim, todas as procuras acima correspondem a uma facilitação do AC/DC em relação ao CWB, em que teriam de estar sempre envolvidas dentro de aspas. (Repita as procuras com as aspas para se compenetrar disso).

As aspas são necessárias quando quisermos procurar uma sequência de palavras, em que portanto temos de indicar onde começa uma e acaba outra

```
"por" "exemplo"
"com" "a" "sua" "autorização"
"fim" "\."
```

Donde, se quiser procurar aspas também é preciso usar Ċontudo, no AC/DC nós codificámos as aspas iniciais como « e finais como », e para estas não precisa de colocar

Naturalmente, pode-se aplicar toda a sintaxe já aprendida a sequências de palavras ou outros sinais gráficos...

```
"d[eoa]s*" "car.*"
"como" "se"{2,2}
"como" "se"+
"com" "a"{0,1} "sua" "mãe"
"sim" ".*" "não"
"mas" "\."
```

De facto, ainda existe mais uma forma, aparentemente mais complicada, de fazer as mesmas procuras, e que consiste em integrá-las (para cada palavra ou unidade) dentro de [**word**=¨¨].

```
[word="irrequieto"]
[word="irrequiet.*"]
[word="crianças*"] [word="irrequiet.*"]
[word="crianças*"] [word="muito|muitíssimo"]* [word="irrequiet.*"]
```

Embora isto pareça um trabalho completamente escusado, a verdade é que ainda mostrámos apenas a ponta do icebergue da capacidades do CWB. E de facto o atributo **word** é apenas um dos muitos possíveis para guardar

informação sobre cada palavra num corpo de texto, como veremos na próxima secção.

Mas uma vantagem imediata devemos indicar já: é que é possível exprimir também casos negativos, que uma palavra não seja igual a uma dada palavra ou expressão. Senão veja-se

```
[word="última"] [word!="vez"]
[word="família"] [word="com"] [word!="muitos|poucos|alguns"] [word="filhos"]
```

2 Informação codificada nos corpos do AC/DC

De facto, para cada palavra (ou unidade) dos corpos preenchemos informação relativa a muitas formas diferentes de analisar, e que se chama geralmente "anotação".

A cada um desses tipos de informação se dá um nome, e se declara como "atributo posicional". Os corpos do AC/DC, por serem anotados com o PALAVRAS, têm todos os seguintes atributos correspondentes a traços morfossintáticos: **lema**, **pos**, **temcagr**, **pessnum**, **gen** e **func**.

Embora não vá explicar em detalhe o que significa cada uma destas "posições", tenho que indicar, por um lado, que duas delas, **temcagr** e **pessnum**, têm informação empilhada referente a assuntos diferentes – por uma questão de poupar espaço de armazenamento dos corpos, e por outro como tratámos os casos em que uma palavra (separada por espaços) não era equivalente a uma unidade gramatical.

2.1 Posições que contêm informação referente a mais de um campo

Temcagr acondiciona tempo (verbal), caso (pronominal) e grau (adjetival). Nos casos em que não tem nenhuma destas propriedades, o seu valor é zero.

Pessnum incorpora tanto a pessoa, que pode ser primeira, segunda ou terceira, do singular ou do plural, para verbos e pronomes, como o número para nomes e adjetivos. Nas palavras de classes a que não pertence nem número nem pessoa, como os advérbios ou as palavras gramaticais, o seu valor é zero.

2.2 Palavras que correspondem a mais de uma unidade sintática: contrações e clíticos

A escolha feita no AC/DC foi manter a palavra ortográfica, mas analisar em várias palavras nos atributos gramaticais, e separá-las pelo sinal +. Note

bem, esta é uma codificação própria do AC/DC, não tem nada a ver com o CWB, podíamos ter escolhido qualquer outro carater para separar.

Para se aperceber da forma como está feita a análise, o mais simples é procurar palavras desse tipo e investigar o **lema** e a **pos**.

```
amá-la-ia
deu-se
delas
comigo
deu-lhos
numa
```

Por outro lado, a sintaxe que aprendemos acima e que indica que tipo de posição estamos interessados, permite procurar coisas em vários níveis (reparem que quando estamos à procura de + como indicadore de separador de palavras, temos de usar .

```
[lema="fazer\+.*"]
[lema="em\+.*"]
[lema="em(\+.*)"] "casa"
[pos="V.*\+.*\"]
[lema="andar|correr"] [pos="DET.*"]* "mundos*"
```

2.3 Palavras que fazem parte de unidades maiores: nomes próprios e expressões com várias palavras

Finalmente, no caso de nomes próprios com mais de uma palavra, escolhemos manter cada palavra individualmente, mas juntar no lema todas as palavras que pertencem ao mesmo nome próprio, separadas pelo sinal de igual =. Isso explicará os casos de distribuição de lema que terão obtido na procura anterior

```
Ant[ôó]nio
```

Por outro lado, podem procurar nomes próprios específicos

```
[lema="Casa=Branca"]
[lema="Lula=da=Silva"]
[lema="Universidade=.*"]
```

3 A divisão dos corpos em unidades

Outro tipo de atributos que se podem especificar no CWB são os chamados atributos "estruturais", que indicam as divisões em que se podem classificar os textos. A mais natural é a frase ou sentença, **s**, seguida de outras como **p** parágrafo, **t** título, etc. Para cada corpo do AC/DC está indicado na sua página o conjunto dos seus atributos estruturais.

Além disso tanto a análise sintática como a semântica fazem uso do atributo **mwe** (expressão com várias palavras).

No CWB a especificação do princípio e do fim destas partes do corpo é feita usando os sinais de maior e menor, e podem ser procurados

```
<mwe> "caixa"
<mwe> "caixa" []+ </mwe>
<t> "Finalmente"
```

Um dos comandos importante do CWB é a possibilidade de restringir uma dada procura dentro de uma unidade estrutural, por exemplo, o caso mais comum sendo **s**. Se estamos por exemplo a procurar adjetivos predicativos após o nome, faz sentido que restinjamos essa procura dentro de uma frase só. Veja a diferença entre as duas primeiras procuras seguintes:

```
[pos="N.*"] []* [pos="ADJ.*"]
[pos="N.*"] []* [pos="ADJ.*"] within s
[lema="cidadão"] [word!=","]* [lema="feliz"] within entrevista
```

Repare que, se quiser selecionar casos que apenas ocorram dentro de um dado atributo estrutural, isso não se faz com a indicação **within t**, por exemplo. O within só se refere a "não saltar entre casos do atributo estrutural diferentes". Para procurar os casos dentro de um atributo estrutural, sem indicar a posição, pode usar a seguinte notação (avançada):

```
[lema="feliz" & _.t]
[lema="braço" & _.mwe]
```

4 A anotação semântica

Depois de termos tratado cabalmente a anotação sintática, resolvemos – no projeto AC/DC – adicionar mais atributos posicionais aos corpos, para conter informação doutra índole.

Neste momento existem os atributos **sema** e **grupo** referentes a três campos semânticos, nomeadamente cor, roupa e emoções, que passo a explicar.

Existem também três atributos **sinonimo**, **antonimo** e **hiperonimo** que têm uma lista de, respetivamente, sinónimos, antónimos e hiperónimos da palavra em questão (sem entrar em conta com o contexto).

4.1 O campo da cor

É preciso indicar que, ao incluir este nível, adicionámos uma conjunto de informações a que demos significado, em particular o uso de : para descrever tipos diferentes dentro dum mesmo campo, e o carater de sublinhado para aceitar vagueza de interpretação.

Assim, no atributo posicional **sema** os valores relativos ao campo da cor podem ter os seguintes valores, que podem ser procurados

```
[sema="cor"]
[sema="cor:original"]
[sema="cor:equipa"]
[sema="cor:raça"]
[sema="cor:vinho"]
[sema="cor:humana"]
[sema="cor:humana2"]
```

Se se quer apenas obter casos de cor independentemente do seu tipo, ou apenas os que não são cor pura, pode naturalmente usar-se respetivamente as seguintes expressões de pesquisa

```
[sema="cor.*" & sema!="coragem"]
[sema="cor:.*"]
```

No caso da cor pura, dividimos em grupos (com a primeira letra maiúscula) referentes às propriedades visuais, e colocámos no atributo posicional **grupo**. Quando não é cor pura, o grupo é 0. É pois possível fazer procuras como

```
[grupo="Vermelho"]
[grupo="Vermelho" & lema="cor"]
[grupo="Vermelho" & lema!="cor"]
[lema="vermelho" & sema!="cor"]
[lema="vermelho" & sema!="cor.*"]
[grupo="Vermelho" & _.mwe]
```

4.2 O campo das emoções

Neste momento ainda temos apenas as indicações de medo ou coragem, no atributo posicional **sema**.

```
[sema="coragem"]
[sema="medo"]
```

5 Outras informações

Uma questão avulsa que pode ser relevante conhecer, é o facto de se poder pedir para ver, para cada elemento da concordância, os valores de outros atributos posicionais que não apenas o atributo posicional **word**.

Para isso, tem de se dar o seguinte comando do CWB, e depois no fim colocar a expressão de pesquisa (que, se for simples, tem de estar rodeada por aspas):

```
show +pos; [sema="medo"]
show +pos; show +temcagr; "foi"
```

Outra coisa que até agora não mencionámos, mas que é extremamente relevante na interação com o AC/DC, é o uso de pedidos de distribuição e não apenas de concordância. Nesses casos, quando aquilo de que se pretende a distribuição não é o primeiro elemento da expressão de pesquisa, tem de se marcar o elemento escolhido através do símbolo @ (arroba), como os seguintes exemplos ilustram:

```
"de" @[sema="medo"]
[pos="N.*"] @[pos="ADJ.*"]
[lema="casa"] @"de|dos" "meus" "pais"
```

Finalmente, existem outras formas de interagir com o CWB, que pode ter interesse conhecer mesmo que não usem:

O comando **cut X** permite obter os primeiros X resultados

```
[sema="medo"] cut 10
```

E é possível procurar co-ocorrência de dois termos num dado atributo estrutural (nos exemplos seguintes a frase), sem indicar qual a ordem:

```
MU (meet "homem" "mulher" s)
MU (meet [lema="novo" & pos="ADJ.*"] [lema="velho" & pos="ADJ.*"] s)
```

6 Comentários finais

O caminho de cada um através desta floresta de documentação, não o pretendo predeterminar. Bom passeio, e boa colheita... de ideias, métodos, e programas!

Agradecimentos

Agradeço à equipa do AC/DC e aos meus alunos Anne Lise Holta, Heidi Jansen e Nicholas Jorgensen o ensejo que me deram de escrever estas instruções.