Introdução à sintaxe e à anotação no AC/DC

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

Primavera 2019

Algumas linguagens artificiais

- Expressões regulares: Uma linguagem para descrever cadeias de caracteres (sequências de caracteres), com os seguintes caracteres especiais: `.`, `*`, `+`, `?`, `[]`, `()`, `{}`, `|` e `\` .
- A sintaxe do OCWB: usa [ ] para as unidades, e elas próprias têm um conjunto de pares atributo valor
- Pode-se usar expressões regulares dentro dos valores das unidades, ou sobre as unidades elas próprias.
- Quando há mais do que uma unidade, tem de se indicar onde começa cada uma. Com aspas, ou com [].

Expressões regulares

- Formas de descrever um conjunto de unidades pertencentes a uma linguagem.
- Neste caso vamos aplicá-las ao português, que é uma língua (linguagem natural).
- Em primeiro lugar vamos aplicá-las a palavras (ou sinais de pontuação ou números), depois a sequências de palavras e/ou unidades.
- Vamos considerar que existe um conjunto de sinais gráficos (letras, dígitos e outros) que formam o nosso vocabulário.
- Além disso, existem outros sinais que têm um significado especial.

Expressões regulares: principais atores

*	zero ou mais vezes
+	uma ou mais vezes
?	zero ou uma vez
.	qualquer carater

Exemplos: cas+a, cald?o, .*inho, corr.+, .*lh.*, c.ma, dos*, das?

Expressões regulares: outros atores

E quando queremos não um carater específico, nem um qualquer, mas uma classe de carateres? Então incluímos dentro destes parênteses retos: []. Outra vezes indicamos qualquer carater exceto uma dada classe, indicando a negação por ^.

[ab] ou a ou b

[0-9] qualquer algarismo entre 0 e 9

[óôoõ] um o, acentuado ou não

[-?.Z] hífen, ponto de interrogação, ponto ou Z.

[^aeiou] tudo menos vogais

Exemplos: pr[éê]mios?, c[aeiou]rt[ao], almoç[áa]mos, A[0-9]+

Expressões regulares: mais complexidade

- Juntar disjunção de sequências, não só de unidades simples: caixa|saco, (ou|oi)ro,
- em vez de um número qualquer, poder escolher quantas vezes uma dada unidade é repetida: {n,m} mínimo `n` vezes, máximo `m` vezes

car{1,2}o, go{3,8}gle, (ca|co){1,3}

Expressões regulares no ACDC a outro nível

Os textos do AC/DC estão divididos por unidades. E a cada unidade, além da sua forma, está muito mais informação associada, em forma de pares atributo-valor.

- Por isso cada unidade é descrita por [], e podem escolher-se os atributos pelos quais se quer identificar uma unidade. A sua forma, o seu lema, a sua categoria gramatical, a variante a que pertence, o seu autor, etc.

```
[word="comida"], [lema="beber"], [autor="EQ"], [variante="
```

- ou conjunções desses critérios, usando & ou

```
[lema="beber" & autor="EQ"], [word="comida" & variante="BR
```

- E são usadas expressões regulares sobre esses pares

```
[pos="N.*"] [pos="ADJ.*"]{2,3} [word="que"]
```

Qual a informação associada a cada unidade?

Isso foi decidido por nós (AC/DC) e parcialmente pelo PALAVRAS. Grande parte do curso vai servir para compreender esse tipo de informação. Dois tipos diferentes:

- informação diferente para cada palavra, às vezes chamada também etiquetagem, do inglês "tagging"
- informação de agrupamento, ou por textos, às vezes chamada também marcação, do inglês "markup"
  - Apenas as fronteiras (os chamados atributos estruturais). (Quem sabe XML, ou HTML, percebe logo)
  - Em todos os elementos (também chamados metadados - simplificam as distribuições)

O que é uma palavra?

Não há definição padrão de uma palavra (forma gráfica), diferentes sistemas ou investigadores têm diferentes opiniões.

- Além disso, as palavras vão mudando, veja-se o Pepetela a escrever por exemplo *masé*.
- E veja-se o corpo ReLi...

Por isso a escolha foi nossa, e refere-se sobretudo a clíticos, contrações, locuções e nomes próprios.

O que é uma palavra?

A nossa solução foi em geral dar as duas possibilidades, tendo a separação ou a junção no lema.

```
word do -> lema de+o
word Maria -> lema Maria=da=Fonte
word da -> lema Maria=da=Fonte
word Fonte -> lema Maria=da=Fonte
word dei-mo -> lema dar+eu+ele
word dei-lha -> lema dar+lhe&lhes+ela
word esqueceu-se -> lema esquecer+se
```