Processamento de Linguagens
LCC (2º ano)

Trabalho Prático nº 1 — 3ª semana
(Lex)

Ano lectivo 08/09

1 Objectivos e Organização

Este trabalho prático tem como principais **objectivos**:

- aumentar a experiência de uso do ambiente linux, da linguagem imperativa C (para codificação das estruturas de dados e respectivos algoritmos de manipulação), e de algumas ferramentas de apoio à programação;
- aumentar a capacidade de escrever *Expressões Regulares (ER)* para descrição de *padrões de frases*;
- desenvolver, a partir de ERs, sistemática e automaticamente *Processadores de Linguagens Regulares*, que filtrem ou transformem textos;
- utilizar *geradores de filtros/processadores de texto*, como o Flex

Para o efeito, esta folha contém 2 enunciados, dos quais deverá resolver pelo menos um. O programa desenvolvido será apresentado aos membros da equipa docente, totalmente pronto e a funcionar (acompanhado do respectivo relatório de desenvolvimento) e será defendido por todos os elementos do grupo (3 alunos), em data a marcar.
O **relatório** a elaborar, deve ser claro e, além do respectivo enunciado, da descrição do problema, das decisões que lideraram o desenho e a implementação, deverá conter exemplos de utilização (textos fontes diversos e respectivo resultado produzido). Como é de tradição, o relatório será escrito em LaTeX ou NuWeb (Literate Programming).

2 Enunciados

Para sistematizar o trabalho que se lhe pede em cada uma das propostas seguintes, considere que deve, em qualquer um dos casos, realizar a seguinte lista de tarefas:

1. Especificar os padrões de frases que quer encontrar no texto-fonte, através de ERs.

2. Identificar as acções semânticas a realizar como reacção ao reconhecimento de cada um desses padrões.

3. Identificar as Estruturas de Dados globais que possa eventualmente precisar para armazenar temporariamente a informação que vai extraindo do texto-fonte ou que vai construindo à medida que o processamento avança.

4. Desenvolver um Processador de Texto para fazer o reconhecimento dos padrões identificados e proceder à transformação pretendida, com recurso ao Gerador Flex.

2.1 WikictionaryPT — Um dicionário de sinónimos e antónimos em português

O Wikcionário é um projecto colaborativo para produzir um dicionário poliglota em português, com significados, etimologias e pronúncia. Wikcionário é um complemento léxico para todo o conteúdo aberto da enciclopédia Wikipédia. Uma das vantagens deste Wikcionário é o formato em que se encontra disponível — XML (eXtensible Markup Language), facilitando o seu processamento e transformação.

Neste trabalho pretende-se que analise este formato particular do Wikcionário de forma a familiarizar-se com o seu conteúdo e poder executar as tarefas abaixo descritas.

Listing 1: Fragmento do WikictionaryPT

```
<page>
    <title>atordoar</title>
    <id>42546</id>
    <revision>
      <id>293478</id>
      <timestamp>2007-08-25T21:42:18Z</timestamp>
      <contributor>
        <username>Marcot</username>
        <id>3079</id>
      </contributor>
      <comment>Pron</comment>
      <text xml:space="preserve">{{-pt-}}
==Pronúncia==
{{Áudio|Pt-br-atordoar.ogg|Português (BR)}}
==Verbo==
'''a.tor.do.ar''' ''transitivo direto''
#(''e pronominal'') [[confundir]] os sentidos ou o raciocínio,
geralmente por efeito de [[som]], possivelmente forte ou
surpreendente, ou apenas por alguma [[pancada]],
[[surpresa]] ou forte [[emoção]], ou ainda por [[embriaguez]]
# [[perturbar]] os ouvidos
# causar grande [[surpresa]]

===Sinonímia===
#[[abalar]], [[aturdir]], [[confundir]], [[perturbar]]
#[[atroar]], [[azucrinar]], [[ensurdecer]]
#[[abalar]], [[surpreender]]
```

```
28
29 ===Etimologia===
30 *''controversa''
31 **do Espanhol ''[[atolondrar]]'', relacionado ao Latim
32 ''[[tonitrus]]'' ([[trovão]]) pela suposta forma ''atordonare''
33 **do Espanhol ''[[aturdir]]''
34
35 ===Conjugação===
36 *''regular''
37 {{conj.pt.oar|atord}}
38
39 [[Categoria:Verbo_(Português)]]
40
41 [[fr:atordoar]]
42 [[hy:atordoar]]</text>
43     </revision>
44  </page>
```

As tarefas que deverá executar neste trabalho prático serão faseados por semanas da seguinte forma:

1ª semana Analise o documento XML correspondente ao Wictionary português[1] e faça a contagem das categorias (verbos, advérbios, substantivos, adjectivos, pronomes — pessoais, possessivos, demonstartivos, etc) que ocorrem no documento. No final, deverá produzir um documento em formato HTML com as categorias e respectivas contagens.

2ª semana Complete o seu processador da 1ª semana de modo a filtrar o significado, bem todos os sinónimos e antónimos (caso existam), dos verbos que aparecem no documento. Deverá ainda produzir um documento HTML com tal informação.

3ª semama Finalize o seu processador acrescentando a funcionalidade de pesquisa. Sempre que o utilizador introduza um verbo a pesquisar deverá mostrar o significado e respectivos sinónimos/antónimos. Como facilidade extra, poderá mostrar os significados/antónimos na forma de um grafo, recorrendo para tal à linguagem Dot do GraphViz[2]

2.2 BibTeXPro — Um processador de BibTeX

BibTeX é uma ferramenta de formatação de referências (ou citações) bibliográficas feitas em documentos LaTeX. BibTeX foi criada por Oren Patashnik e Leslie Lamport em 1985 com o objectivo de facilitar a separação da base de dados bibliográfica do texto principal do documento.

Listing 2: Exemplo de entrada em BibTex

```
1 @InProceedings{CPBFH07e,
2   author =    {Daniela da Cruz and Maria João Varanda Pereira
```

[1] Disponível em `http://download.wikimedia.org/ptwiktionary/20080304/ptwiktionary-20080304-pages-articles.xml.bz2`.

[2] Disponível em `http://www.graphviz.org`. Poderá ainda usar uma das ferramentas que processam Dot disponíveis em `http://www.graphviz.org/Resources.php`.

```
3                 and Mário Béron and Rúben Fonseca and
4                 Pedro Rangel Henriques},
5   title =     {Comparing Generators for Language-based Tools},
6   booktitle = {Proceedings of the 1.st Conference on Compiler
7                Related Technologies and Applications, CoRTA'07
8                --- Universidade da Beira Interior, Portugal},
9   year =      {2007},
10  editor =    {},
11  month =     {Jul},
12 }
```

De modo a familiarizar-se com o formato do BibTeX poderá consultar o ficheiro `lp.bib` disponível em `http://www.di.uminho.pt/~prh/lp.bib` e ainda a página oficial do formato referido (`http://www.bibtex.org/`).

As tarefas que deverá executar neste trabalho prático serão faseadas por semanas da seguinte forma:

1ª semana Analise o documento BibTeX referido acima e faça a contagem das categorias (phDThesis, misc, InProceeding, etc), que ocorrem no documento. No final, deverá produzir um documento em formato HTML com as categorias e respectivas contagens.

2ª semana Complete o seu processador da 1ª semana de modo a filtrar, para cada entrada, a respectiva chave, autores, título, ano e mês. Deverá ainda produzir um documento HTML com tal informação.

3ª semama Finalize o seu processador acrescentando a funcionalidade de pesquisa. Sempre que o utilizador introduza um palavra (nome de autor ou ano da publicação) a pesquisar deverá mostrar a respectiva chave, autores e título da publicação. Como facilidade extra, poderá mostrar os autores que normalmente com o autor em causa na forma de um grafo, recorrendo para tal à linguagem Dot do GraphViz[3]

[3]Disponível em `http://www.graphviz.org`. Poderá ainda usar uma das ferramentas que processam Dot disponíveis em `http://www.graphviz.org/Resources.php`.