I WORKSHOP DA APL

SOBRE

LINGUÍSTICA COMPUTACIONAL

Lisboa, Maio de 1998

DICIONÁRIOS ELECTRÓNICOS E ANÁLISE LEXICAL AUTOMÁTICA

Elisabete Ranchhod (FLUL/CAUTL)

DICIONÁRIOS ELECTRÓNICOS E ANÁLISE LEXICAL AUTOMÁTICA[*]

Elisabete Marques Ranchhod

FLUL/CAUTL[**]

Resumo

Como se sabe, uma das utilizações dos dicionários electrónicos é a análise lexical automática de textos escritos. Dependendo do conteúdo dos dicionários, a análise lexical pode ir desde uma simples verificação morfológica até à identificação das unidades lexicais complexas, muito frequentes em qualquer tipo de texto. Os resultados destas análises podem levar a uma etiquetagem do texto, cujo maior ou menor grau de adequação depende de vários factores, entre eles o número de etiquetas utilizado. Dada a enorme ambiguidade das línguas naturais, o analisador pode propor, para um mesmo texto, várias análises lexicais, embora só uma seja adequada. A eliminação das análises incorrectas passa por resolver a ambiguidade através de gramáticas para tal concebidas. Etiquetagem e resolução de ambiguidades são duas operações distintas. A primeira pode ser feita de forma independente, a segunda só pode ser feita em simultâneo ou após a etiquetagem do texto.

Neste artigo, referir-se-ão de forma breve as características dos dicionários electrónicos do português DIGRAF e DIGRACF, respectivamente para palavras simples e compostas, para falarmos com maior detenimento do tipo de análises lexicais que podem ser obtidas por aplicação do conteúdo desses dicionários a um dado texto.

Palavras-chave

Dicionários electrónicos; análise lexical; etiquetagem lexical; processamento das línguas naturais.

---

[*] Estudo parcialmente financiado pelos Programas PRAXIS XXI (Proj. 2/2.1/CSH/775/95) e Cooperação JNICT/Embaixada de França (Proj. 098 J4).

[**] Faculdade de Letras da Universidade de Lisboa e Centro de Automática da Universidade Técnica de Lisboa, Instituto Superior Técnico – Av. Rovisco Pais – P1096 LISBOA.

elisabet@label.ist.utl.pt , http://www.ist.utl.pt/pt/investigacao/

1. **Dicionários DIGRAMA de palavras simples e compostas**

Os dicionários de palavras simples e compostas do sistema DIGRAMA são instrumentos linguísticos especificamente concebidos para serem utilizados por programas informáticos em processamento de textos escritos em português (Eleutério e al., 1993). Isto faz com que tenham características completamente diferentes das dos dicionários de uso, sejam estes apresentados em papel, em suporte magnético, óptico ou magnético-óptico (Ranchhod e al., 1996). A informação linguística contida nos dicionários foi determinada, formalizada e codificada à mão por uma equipa de linguistas, a geração das formas flexionadas é feita automaticamente a partir das regras formuladas pelos linguistas. Como qualquer dicionário, são constituídos por *artigos*; cada artigo é composto por uma *entrada* (uma palavra) e um *conteúdo* (a descrição linguística formalizada da entrada). O sistema de dicionários é modular e está estruturado do seguinte modo:

- DIGRAS: módulo de *palavras simples*, entendendo-se por palavra simples qualquer sequência de caracteres alfabéticos delimitada por *separadores*. Um separador é um caracter não alfanumérico. As entradas do DIGRAS representam *formas canónicas*, correspondendo estas ao masculino singular para nomes e adjectivos que têm essa variação, feminino singular para os nomes e adjectivos que são exclusivamente femininos, infinitivo para os verbos, etc. São entradas do DIGRAS os exemplos: *ver*, *ontem*, *exposição*, *interessante*, etc.
- DIGRAC: módulo das *palavras compostas*, isto é, das unidades lexicais não composicionais que, pertencendo a várias categorias gramaticais, são formadas por *uma sequência de palavras simples e de separadores adequados*. As entradas do DIGRAC são nomes (*conferência de imprensa, campo magnético)*, adjectivos (*certo e sabido, mal-educado*), advérbios (*de facto, a par e passo*), preposições (*acerca de*), etc.

As entradas do DIGRAS e do DIGRAC contêm informações linguísticas codificadas, que são utilizadas por um algoritmo para geração automática das formas flexionadas das palavras simples e compostas, dando origem, respectivamente, ao DIGRAF e ao DIGRACF.

O DIGRAMA está preparado para poder receber, de forma acumulativa, informações de natureza sintáctica e semântica (por agora, registadas em matrizes), o que significa que, à medida que essas informações forem sendo adicionadas, as gramáticas vão sendo simultaneamente construídas[1].

Os métodos de elaboração dos dicionários bem como o formato adoptado são idênticos aos dos seus correspondentes concebidos no LADL[2] para o francês, DELAS (Courtois, 1990) e DELAC (Silberztein, 1990). Assim, foram facilmente elaboradas versões dos dicionários do português com as notações do sistema INTEX (Silberztein, 1993). Os dicionários do português estão, pois, integrados no INTEX, tanto na versão para NextStep como na mais recente para Windows 95-NT. As pesquisas e análises lexicais de que iremos falar adiante foram feitas por utilização do DIGRAF e do DIGRACF integrados no INTEX.

1.1. *Codificação da informação linguística*

As entradas do DIGRAS e do DIGRAC têm, esquematizando, a seguinte forma:

$< \mathbf{P_i} \,.\, \mathbf{I_j} >$ ou $< \mathbf{P_i} \,.\, \mathbf{I_j} \,.\, \mathbf{I_k} >$, em que:

---

[1] O elevado número de entradas do DIGRAS (mais de 100.000) e do DIGRAC (por agora, apenas 20.000) impôs a necessidade de reduzir o léxico a analisar mais aprofundadamente. Constituiu-se, assim, um sub-dicionário (Ranchhod, no prelo) que contém cerca de 20.000 entradas, consideradas como vocabulário nuclear e de utilização mais frequente. Este léxico está a ser objecto de estudo sintáctico-semântico mais profundo. As restantes entradas (correspondentes a vocabulário menos utilizado ou caído em desuso) não disporão de informação gramatical adicional à já contida nos dicionários.

P$_i$ representa a *forma canónica* de uma unidade lexical (simples ou composta) e **I$_j$** corresponde à informação codificada, acima referida (categoria gramatical, regras de flexão, existência de sufixos diminutivos, aumentativos e superlativos, condições, no caso dos verbos, para a adjunção adequada de clíticos). O segundo formato representa, de forma simplificada, certos casos de homografia, isto é, certas situações em que a uma mesma palavra correspondem várias unidades lexicais. Por exemplo, *andar* e *forte* são palavras ambíguas, pois podem *a priori* ser analisadas, respectivamente, como (i): o masculino singular de um nome: *um andar com vistas panorâmicas*, ou a forma infinitiva de um verbo: *começaram a andar mais depressa*; (ii) o masculino singular de um nome: *um forte do século XVI,* ou de um adjectivo: *o café está forte*). Mais adiante referiremos outras situações de ambiguidade provocada pela homografia.

A título de exemplo, apresentam-se, com o formato do sistema Digrama, as entradas completas do DIGRAS e do DIGRAC de:

ontem,ADV1
ver,V2F16T
exposição,N308
forte,ADJ124S3.N200
conferência de imprensa,N,NDN300

O código alfabético colocado imediatamente após a vírgula indica a categoria gramatical da entrada (**N**, nome; **V**, verbo; **ADJ**, adjectivo; **ADV**, advérbio); o código numérico contém informações sobre a flexão da entrada. Em relação ao verbo, **V2** indica que se trata de um verbo da 2ª conjugação, que segue o modelo de flexão **F16** e se pode construir com clíticos acusativos e reflexos (**T)**. A codificação dos nomes compostos[3] inclui informações sobre o género do composto na sua globalidade, sobre a sua constituição interna, sobre os

---

[2] Laboratoire d'Automatique Documentaire et Linguistique, instituição com a qual mantemos há anos uma estreita colaboração.

elementos que podem flexionar e sobre os separadores possíveis. Como se disse, os dicionários de formas flexionadas são automaticamente gerados a partir destas informações. As formas do DIGRAF e DIGRACF correspondentes a estas entradas têm a seguinte informação :

ontem[ontem].ADV1
ver[ver].V2F16T.Indrfc(AQ).Infips(SAQ).Infpps_1s(SAQ)/2’s(SAQ)/3s(SAQ)
exposição[exposição].N308:fs
forte[forte].ADJ124S:mfs,forte[forte].N200:ms
conferência de imprensa[conferência de imprensa].N+NDN300:fs

Comentando apenas o caso do verbo, a forma **ver** é o lema da entrada (entre parênteses rectos), mas pode, além disso, corresponder à forma truncada do futuro e do condicional, sempre que há clíticos em posição medial (Indrfc(AQ)), ao infininivo impessoal sem clíticos (Infips(SAQ)) e à 1ª, 3ª e 2ª (você) pessoas do singular do infinitivo pessoal (Infpps).

A descrição linguística de cada entrada constitui o conteúdo de cada etiqueta associada à palavra aquando da análise de um dado texto. Esta breve amostragem dá uma ideia do conteúdo das etiquetas utilizadas: lema, classe e sub-classe gramatical, género, número, tempo, pessoa, grau, etc.[4]

1.2. *Homografia e ambiguidade lexical*

Numerosas entradas do dicionário de formas canónicas são ambíguas. Vimos antes o caso de *forte*, que, descontextualizado, tanto pode ser nome (*um forte do sec. XV*) como adjectivo (também ele ambíguo: *uma moeda forte, um café forte, um homem forte*). Nas

---

[3] Para uma análise pormenorizada dos problemas postos pelo tratamento dos nomes compostos ver BAPTISTA, J. (1994).

[4] Para uma discussão aprofundada dos diferentes métodos de etiquetagem, seus objectivos e possíveis avaliações remetemos para os trabalhos de LAPORTE, E. (1996, 1997).

línguas que, como o português, têm um sistema morfológico muito desenvolvido, a homografia das formas flexionadas é consideravelmente maior do que a que se observa nas formas canónicas. Apesar de os linguistas estarem conscientes desta fonte de ambiguidade lexical, não é sem alguma surpresa que se verifica que expressões que, para um humano, não desencadeiam senão uma interpretação têm, do ponto de vista da máquina, várias análises possíveis. A título ilustrativo, veja-se o seguinte texto:

Texto 1: *O livro foi revisto pelo amigo*

A forma *revisto* permite mais do que uma leitura, o que confere ao texto uma dupla interpretação. Para a máquina, contudo, todos os elementos lexicais são ambíguos, todos, sem excepção, podem à partida receber mais do que uma análise linguística possível, como se pode verificar pelo resultado apresentado, por ordem alfabética, pelo analisador do sistema INTEX (sistema que utilizaremos de agora em diante):

amigo, amigar.V:P1s[$SAQO]
amigo, amigo.A:ms
amigo, amigo.N:ms
foi, ir.V:J4s[$SAQO]:J3s[$SAQO]
foi, ser.V:J4s[$SAQO]:J3s[$SAQO]
livro, livrar.V:P1s[$SAQO]
livro, livro.N:ms
o, eu.PRO+Pes
o, o.DET+Art+Def
o, o.PRO+Dem
pelo, pelar.V:P1s[$SAQO]
pelo, pelo.PREPDET+Art+Def
revisto, rever.V:K[$S]
revisto, revestir.V:P1s[$SAQO]
revisto, revistar.V:P1s[$SAQO]
revisto, revisto.A:ms

Para a etiquetagem deste texto solicitou-se ao sistema que o indexasse, aplicando toda a informação contida nos dicionários. O texto não contém palavras compostas. Quando isso acontece, uma parte da ambiguidade das palavras simples fica resolvida, uma vez que o composto é logo identificado como uma unidade lexical, à qual é associada a etiqueta correspondente. Por exemplo, *andar* é, como se disse antes, uma palavra ambígua (nome e verbo) e *modelo* também o é (forma do verbo *modelar* e nome). Contudo, no interior do nome composto *andar modelo*, cada um desses elementos tem apenas um valor gramatical (ver Anexo I para outros exemplos). Isto não significa que não haja nomes compostos que não possam igualmente permitir uma análise como grupos nominais livres. Em:

(1) *Eles detestam mesas redondas,*
(2) *Eles compraram mesas redondas,*
(3) *Eles organizaram mesas redondas,*

mesa redonda, na posição de complemento de um verbo como *detestar*, em (1), confere à frase uma ambiguidade irredutível; em (2), constitui um grupo nominal livre, para, em (3), formar um nome composto.

Tanto no caso das palavras simples como no das compostas, os resultados da análise lexical só serão adequados, ou aproximar-se-ão dos desejados, se o analisador utilizar informações sintácticas integradas em gramáticas de resolução de ambiguidades (mas desta questão não nos vamos ocupar aqui) .

2. **Utilização dos dicionários para análise lexical de texto**

A maioria dos métodos actuais de análise lexical automática baseiam-se fundamentalmente na utilização de: (i) dicionários e gramáticas formalizados, que são directamente aplicados ao texto; (ii) *corpora* previamente etiquetados ou não. No último caso, as etiquetas podem ser automaticamente apuradas mediante cálculos probabilísticos. Certos métodos utilizam simultaneamente um dicionário para determinar quais as etiquetas possíveis e um corpus para resolução de ambiguidades (E. Laporte 1997).

No caso dos sistemas que utilizam dicionários, método que adoptamos, dado um texto, entendido como uma sequência de palavras e de separadores, e um ou vários dicionários, concebidos como listas de artigos em que cada artigo é constituído por uma entrada (uma palavra, simples ou composta) e um conteúdo (a descrição formal da entrada), a etiquetagem do texto consiste em associar a cada uma das palavras que o constitui o conteúdo do artigo de que ela é a entrada. O exemplo apresentado em 1.2. é uma amostra de um texto (*O livro foi revisto pelo amigo*) etiquetado. O analisador retirou dos dicionários toda a informação que pode *a priori* ser associada a cada palavra e, dada a sua ambiguidade intrínseca, todas elas receberam pelo menos duas etiquetas (duas análises).

De facto, o sistema não teve em conta quaisquer dependências contextuais, que teriam levado, por exemplo, a que: (i) *livro* não pudesse ser considerado uma forma verbal, uma vez que precede uma outra forma que, essa sim, não pode ser senão um verbo numa forma finita; não podendo *livro* ser um verbo, terá de ser um nome, o que faz com que o elemento imediatamente à sua esquerda dificilmente seja um pronome, pelo que deverá ser um determinante (artigo definido). Se este tipo de restrições tivessem sido utilizadas (e o sistema INTEX permite fazê-lo) a etiquetagem do texto corresponderia melhor à análise que qualquer linguista dele faria.

No âmbito desta apresentação continuaremos, contudo, a não utilizar quaisquer gramáticas. Os resultados da análise de texto que se mencionarem serão tão só os que se obtêm pela exploração do conteúdo dos dicionários. Quanto às aplicações de uma análise deste tipo, elas são todas indirectas e têm que ver com as aplicações do processamento de texto em geral (detecção e correcção de erros ortográficos, translineação, exploração de bases de dados documentais, integração em sistemas de tradução assistida ou totalmente automática, etc.). O sistema INTEX está vocacionado para tratar *corpora* de grandes dimensões (várias dezenas de milhões de palavras), para, entre outras utilizações, serem usados pelos linguistas para verificar e pesquisar diferentes tipos de fenómenos léxico-sintácticos. Indicaremos alguns tipos de análises lexicais possíveis. Os resultados das pesquisas são, se desejado, apresentados sob forma de concordâncias, como se ilustra a seguir.

2.1. *Extracção de concordâncias*

O processo de extracção de concordâncias consiste em procurar num texto todas as ocorrências de uma palavra, expressão ou qualquer outra entidade linguística. Os elementos desejados serão apresentados individualmente numa linha, inseridos no seu contexto. As concordâncias têm sido sobretudo utilizadas em trabalhos de índole lexicográfica ou estilística, mas podem ter outras utilizações. Assim, dado o texto:

Texto 2

«A moeda única integrará uma larga maioria de países e um pacto de estabilidade que sob pena de pesadas sanções compromete os membros da zona euro a cumprirem a prazo as orientações para os défices públicos. O pacto vai assim conferir à moeda europeia uma dimensão de estabilidade e rigor.»

(*Jornal de Notícias*, 29 de Abril de 1998)

o utilizador pode querer verificar quais as palavras que começam por *pra* ou verificar os contextos em que ocorre a palavra *estabilidade*. O resultado da última pesquisa será:

uma larga maioria de países e um pacto de	estabilidade	que sob pena de pesadas sanções compro
onferir à moeda europeia uma dimensão de	estabilidade	e rigor.

Mas, para fazer pesquisas um pouco mais complicadas (por exemplo, a procura de todas as formas associadas a um dado lema), é necessário utilizar um texto previamente etiquetado[5] com informações adequadas ou consultar dicionários do tipo dos que antes mencionámos.

2.2. *Pesquisa de unidades lexicais complexas*

Apesar de ter ficado demonstrado (Gross, M. 1984, 1986) que uma boa parte do léxico de uma língua é constituído por palavras compostas, nem sempre se tem dado a esta realidade a devida importância. Porém, qualquer texto apresenta numerosos compostos. E, coisa não desprezável, uma parte do sentido de um texto pode estar ancorada nos nomes compostos que contém. Apresenta-se a indexação dos compostos contidos no *Texto 2*:

---

[5] No **Anexo I** apresenta-se a indexação da primeira linha do texto, a lematização e etiquetagem das palavras não ambíguas do texto completo.

a prazo,a prazo. ADV1+PC
défices públicos,défice público. N+NA:mp
moeda única,moeda única. N+NA:fs
sob pena de,sob pena de. ADV1+PCDN

e as respectivas concordâncias:

ete os membros da zona euro a cumprirem	a prazo	as orientações para os défices públicos. O
cumprirem a prazo as orientações para os	défices públicos	. O pacto vai assim conferir à moeda euro
A	moeda única	integrará uma larga maioria de países e u
ia de países e um pacto de estabilidade que	sob pena de	pesadas sanções compromete os membros

As expressões *a prazo* e *sob pena de* encontram-se no dicionário do texto classificadas como advérbios, pertencendo embora a sub-classes distintas, respectivamente PC e PCDN, por razões sintácticas, que não vamos aqui expor; *défices públicos* e *moeda única* são nomes compostos pertencente a uma classe muito produtiva (NA: nome adjectivo). Há ainda no texto a expressão *zona euro*, um nome composto (classe NN) ainda não integrado nos dicionários e que não foi, por isso, reconhecido como tal. A manutenção dos dicionários é uma questão que não abordaremos no âmbito deste trabalho, mas que merece discussão.

2.3. *Pesquisas definidas por outros critérios linguísticos*

Depois de indexado, o texto pode ser utilizado para pesquisar estruturas linguísticas (morfo-sintácticas, léxico-sintácticas), representadas sob forma de expressões racionais. Os resultados da pesquisa podem ser indexados, contextualizados na sua concordância (como acima) ou destacados a negrito no texto. Daremos alguns exemplos, escolhendo esta última opção. A expressão racional:

<N:fs>

reconhece todas as palavras simples que são potenciais nomes femininos no singular e todos os nomes compostos que, na globalidade, sejam femininos e estejam no singular (caso de *moeda única*):

A **moeda única** integrará uma **larga maioria** de países e um pacto de **estabilidade** que sob **pena** de pesadas sanções compromete os membros da **zona** euro a cumprirem a prazo as orientações para os défices públicos. O pacto vai assim conferir à **moeda europeia** uma **dimensão** de **estabilidade** e rigor.

Recorde-se que não foram aplicadas ao texto gramáticas de resolução de ambiguidades e, por isso, palavras como *europeia* estão descritas no dicionário como nomes e adjectivos. Pelas mesmas razões, a expressão racional <V:Y2s>, que reconhece a 2ª pessoa do singular do imperativo não negativo, identificaria no texto, erradamente, a forma *confere*, que aqui corresponde a uma forma homógrafa do presente do indicativo .

Se o linguista quiser verificar quais os verbos transitivos (directos e indirectos) existentes no texto, a expressão racional:

<V> (<PREP> + <E>) <N>

reconhecerá as estruturas lexicais correspondentes aos verbos que podem ter um complemento directo ou um complemento regido, explicitando, neste caso, a preposição.

O estudo pode apenas dizer respeito à construção específica de um verbo, por exemplo, do auxiliar verbal *ir*. O texto-amostra não contém nenhuma ocorrência deste tipo, mas a expressão racional que reconheceria as estruturas em que *ir* é auxiliar de um verbo no infinitivo ou no gerundivo (*vai comprometer*; *vai comprometendo*) seria:

<ir> (<V:W> + <V: G>)

Porém, não seriam reconhecidas as situações em que entre *ir* e o verbo auxiliado existisse um advérbio, simples ou composto: *vai (brevemente + em breve) comprometer*. Isso pode ser obviado, indicando a possibilidade de uma tal inserção:

<ir> (<ADV> + <E>) (<V:W> + <V: G>)

A pesquisa pode ser refinada se o utilizador aplicar ao texto um conjunto adequado de gramáticas locais representadas por grafos (Gross, M., 1997).

3. **Conclusão**

Apresentámos alguns exemplos de análise automática de texto baseada na utilização de dicionários de palavras simples e compostas, especificamente elaborados.

Uma das utilizações possíveis dos sistemas fundamentados em bases de dados linguísticos deste tipo (como é o caso dos sistemas DIGRAMA e INTEX) é permitir aos linguistas uma exploração de *corpora* para investigar fenómenos morfo-sintácticos e léxico-sintácticos variados. Assim,

- os lexicógrafos poderão extrair dos textos exemplos de atestação do uso das entradas dos seus dicionários. Os que se dedicam à elaboração de dicionários de palavras compostas, pertencentes ou não a léxicos terminológicos, poderão recensear numerosas entradas, procurando estruturas léxico-sintácticas características dos compostos;
- os linguistas que estudam estruturas sintácticas ou morfo-sintácticas específicas poderão encontrar nos textos ilustrações do seu uso. Como as línguas têm a particularidade de surpreender mesmo os linguistas mais experientes, a exploração orientada de textos põe em evidência dados linguísticos interessantes, que de outro modo seriam dificilmente detectáveis.

Anexo I

Indexação da primeira linha do *Texto 2*

a,a.PREP
a,eu.PRO+Pes
a,o.DET+Art+Def
a,o.PRO+Dem
de,de.PREP
e,e.CONJ
estabilidade,estabilidade.N:fs
integrará,integrar.V:F4s[$S]:F3s[$S]
larga,larga.N:fs
larga,largar.V:P2s[$L]:P4s[$SAQO]:P3s[$SAQO]:Y2s[$SAQOZ]
larga,largo.A:fs
maioria,maioria.N:fs
moeda,moeda.N:fs
pacto,pacto.N:ms
países,país.N:mp
pena,pena.N:fs
pena,penar.V:P2s[$L]:P4s[$SAQO]:P3s[$SAQO]:Y2s[$SAQOZ]
pena,peno.A:fs
pena,peno.N:fs
que,que.CONJ
que,que.PRO+Exc
que,que.PRO+Int
que,que.PRO+Rel
sob,sob.PREP
um,um.DET+Art+Indef
um,um.DET+Num
uma,um.DET+Art+Indef
uma,um.DET+Num
uma,umar.V:P4s[$SAQ]:P3s[$SAQ]:Y2s[$SAQZ]
única,único.A:fs
única,único.N:fs

Lematização das palavras não ambíguas

A {moeda/única} {integrar} uma larga {maioria} {de} {país} {e} {um} {pacto} {de} {estabilidade} {que} {sob/pena/de} pesadas {sanção} {comprometer} os {membro} {do} zona {euro} a {cumprir} {a/prazo} as {orientação} para os {défice/público}. {S}
O {pacto} {ir} {assim} {conferir} {ao} {moeda} {europeu} uma {dimensão} {de} {estabilidade} {e} {rigor}.{S}

Etiquetagem das palavras não ambíguas

A {moeda/única,.N+NA:fs} {integrará,integrar.V:F2's[$]:F3s[$]} uma larga {maioria,.N:fs} {de,.PREP} {países,país.N:mp} e um {pacto,.N:ms} {de,.PREP} {estabilidade,.N:fs} que {sob/pena/de,.ADV1+PCDN} pesadas {sanções,sanção.N:fp} {compromete,comprometer.V:P2s[L]:P2's[$AQO]:P3s[$AQO]:Y2s[$AQOZ]} os {membros,membro.N:mp} da zona {euro,.N:ms} a {cumprirem,cumprir.V:U2'p[$]:U3p[$]:V2'p[$AQN]:V3p[$AQN]} {a/prazo,.ADV1+PC} as {orientações,orientação.N:fp} para os {défices/públicos,défice/público.N+NA:mp}. {S}
O {pacto,.N:ms} {vai,ir.V:P2s[L]:P2's[$AQO]:P3s[$AQO]:Y2s[$AQOZ]} assim {conferir,.V:R[AQ]:U1s[$]:U2's[$]:U3s[$]:W[$AQ]:V1s[$AQ]:V2's[$AQ]:V3s[$AQ]} à {moeda,.N:fs} europeia uma {dimensão,.N:fs} {de,.PREP} {estabilidade,.N:fs} e {rigor,.N:ms}. {S}

Referências

BAPTISTA, Jorge (1994), *Estabelecimento e formalização de classes de nomes compostos,* Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa (242p.).

COURTOIS, Blandine (1990), Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française*, 87, «Dictionnaires électroniques du français», Paris: Larousse (pp. 11-22).

ELEUTÉRIO S.; E. RANCHHOD; H. FREIRE; J. BAPTISTA (1995), A System of Electronic Dictionaries of Portuguese. *Lingvisticae Investigationes*, XIX:1, Amsterdam/Philadelphia: John Benjamins Publishing Company (pp. 57-82).

GROSS, Maurice (1986), Une classification des phrases «figées» du français, *Lingvisticae Investigationes Supplementa*, Vol.8, *De la syntaxe à la pragmatique*, P. Attal & Cl. Muller (eds), Amsterdam/Philadelphia: John Benjamins Publishing Company (pp. 141-180).

GROSS, Maurice (1986), Lexicon-Grammar. The Representation of Compound Words, COLING-86, Bona (pp.1-6).

GROSS, Maurice (1997), The Construction of Local Grammars, *Finite-State Language Processing*, Cambridge, Mass./London: MIT Press (pp.329-354).

LAPORTE, Éric (1996), L'évaluation des résultats de la levée d'ambiguïtés lexicales, *Linx*, 34/35, Paris: Université Paris X, Centre de Recherches Linguistiques (pp.291-305).

LAPORTE, Éric (1997), Les mots. Un demi-siècle de traitements, *t.a.l.*, vol. 38, nº 2, Paris: Association pour le Traitement Automatique des Langues (pp. 47-68).

RANCHHOD, E.; S. ELEUTÉRIO (1996), Construção de Dicionários Electrónicos do Português. Problemas Teóricos e Metodológicos. In *Actas do Congresso Internacional sobre o Português*, Lisboa: Colibri (pp. 265-282).

RANCHHOD, E.; C. MOTA (1998), Elaboração de dicionários termiológicos. Seguros. In *Actas do Workshop sobre Linguística Computacional da APL* (no prelo)

SILBERZTEIN, Max (1990), Le dictionnaire électronique des mots composés, *Langue Française,* 87, «Dictionnaires électroniques du français», Paris: Larousse, pp. 71-83.

SILBERZTEIN, Max (1993), *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris: Masson (233 p.).

ELISABETE MARQUES RANCHHOD*

SUMMARY

The majority of the present methods for lexical parsing rely on the use of: (i) formalized dictionaries and grammars that are directly applied to texts; (ii) *corpora* previously tagged, by hand or by using statistical methods.

For Portuguese, we have built large-coverage dictionaries and grammars that are used by INTEX to parse real texts (several million words) in real time.

Within the scope of this note, we first refer to the main characteristics of Portuguese electronic dictionaries for simple and compound words; we afterwards give some examples of lexical parsing using the linguistic information contained in those dictionaries.

Keywords: electronic dictionaries, taggers, lexical parsing, corpus processing computational linguistics.

*) Faculdade de Letras da Universidade de Lisboa e Centro de Automática da Universidade Técnica de Lisboa (IST).

elisabet@label.ist.utl.pt

http://label2.ist.utl.pt