Dicionários Electrónicos do Português
Características e Aplicações

In Actas del VIII Simposio Internacional

de

Comunicación Social

pp. 636-642

Santiago de Cuba, 2003

SAMUEL ELEUTÉRIO*
ELISABETE MARQUES RANCHHOD**
CRISTINA MOTA*
PAULA CARVALHO**
LabEL (CAUTL-IST)*
Faculdade de Letras de Lisboa & LabEL (CAUTL-IST)**
Lisboa - Portugal
email: samuel@label2.ist.utl.pt; elisabet@label.ist.utl.pt;
cristina@label.ist.utl.pt; paula@label2.ist.utl.pt

Dicionários Electrónicos do Português. Características e Aplicações[*]

Abstract. 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 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.

1. Introdução

Os dicionários electrónicos de palavras simples e compostas, elaborados pelo LabEL[i], são léxicos formalizados de ampla cobertura, especificamente concebidos para serem utilizados por programas informáticos em processamento automático de textos escritos em português. Isto faz com que estes léxicos tenham características completamente diferentes das dos dicionários de uso, sejam eles apresentados em papel ou em suporte informático (Ranchhod 2001). Inicialmente, foram elaborados para serem utilizados pelo sistema DIGRAMA, desenvolvido no LabEL (Eleutério *et al.,* 1995); mais tarde, foram convertidos para o formalismo INTEX[ii], desenvolvido por Silberztein, e integrados também neste sistema (Ranchhod *et al*., 1999; Ranchhod, 1999).

2. Dicionários electrónicos do Português

O sistema de dicionários do LabEL é constituído por vários módulos, que, em combinação com gramáticas ou isoladamente, são aplicados na análise automática de textos. O conjunto destes dados lexicais está organizado de acordo com a complexidade formal das unidades lexicais.

2.1. Dicionário de palavras simples: módulos DELAS e DELAF [iii]

O módulo *DELAS* constitui o elemento central do sistema de dicionários. Contém cerca de 120 000 *palavras simples* (lemas), cujos atributos lexicais estão sistematicamente representados por um código. Este módulo corresponde, pois, à forma mais simples de dicionário: uma lista de palavras, representadas pelo seu lema[iv] e de informações linguísticas codificadas sobre a sua categoria gramatical e sobre as regras de flexão que se lhes aplicam. Conforme a categoria gramatical da entrada, as informações dirão respeito a: variação em género, número, caso, tempo, modo e pessoa e às adjunções de sufixos diminutivos, aumentativos e superlativos. Em relação aos verbos, foi igualmente formalizada a possibilidade de poderem construir-se com pronomes clíticos, uma vez que, alguns deles, quando se encontram à direita das formas verbais, levam a uma alteração formal dos verbos, podendo eles próprios sofrer simultaneamente alterações de forma.

As formas flexionadas de um determinado lema são automaticamente geradas a partir dos códigos associados às entradas do DELAS. O dicionário assim obtido constitui o DELAF (cerca de 1 200 000 formas flexionadas).

As entradas dos dicionários DELAS e DELAF têm a forma geral[v] que os exemplos seguintes ilustram:

DELAS
campeão,N046
cedo,ADV1_dh1_dt1_ss1
central,A111_ss018
central,N311
comprar,V101t

DELAF

campeã,campeão.N:fs
campeão,campeão.N:ms
campeãs,campeão.N:fp
campeões,campeão.N:mp
cedinho,cedo.ADV:D
cedíssimo,cedo.ADV:S
cedo,cedo.ADV
central,central.A:fs:ms
centralíssimas,central.A:Sfp
centralíssimos,central.A:Sm
p
centralíssimo,central.A:Sms
centralíssima,central.A:Sfs
centrais,central.A:fp:mp
central,central.N:fs
centrais,central.N:fp
compramos,comprar.V:P1p
compraríamos,comprar.V:C1p
comprareis,comprar.V:F2p
compráramos,comprar.V:M1p
compra,comprar.V:P2s:P3s:Y2
s
compre,comprar.V:S1s:S3s
compravam,comprar.V:I3p
comprasse,comprar.V:T1s:T3s
comprem,comprar.V:S3p
comprassem,comprar.VT3p
comprámo,comprar.V:J1p
comprarias,comprar.V:C2s
compraríam,comprar.V:C3p
compro,comprar.V:P1s
compráramo,comprar.V:M1p
compraste,comprar.V:J2s:J2p
comprando,comprar.V:G
compram,comprar.V:P3p
comprara,comprar.V:M1s:M2s:M3s
compras,comprar.V:P2s
comprasses,comprar.V:T2s
compraria,comprar.V:C1s:C3s
comprássemos,comprar.V:T1p
comprava,comprar.V:I1s:I2s:I3s
comprarmos,comprar.V:U1p:V1p
comprámos,comprar.V:J1p
compres,comprar.V:S2s:Y2s
comprou,comprar.V:J3s
comprar,comprar.V:R:U1s:U3s:W:V1s:V
3s
comprado,comprar.V:K

Nas entradas do DELAS, a vírgula separa o lema do seu código de flexão. A primeira informação codificada diz respeito à categoria gramatical a que a palavra pertence. Nos exemplos: nome (N), advérbio (ADV), adjectivo (A) e verbo (V), respectivamente. O código numérico que se lhes segue representa um dado modelo de flexão. Por exemplo, o código 046, associado ao nome *campeão*, indica que este pode flexionar em número e género (0), segundo a regra de flexão 46. Pelo contrário, o outro nome da amostragem, *central*, tem associado o código 311, o que significa que se trata de uma forma exclusivamente feminina (3) que pode variar em número, neste caso, de acordo com a regra de flexão 11. Esta forma tem a particularidade de também poder ser um adjectivo. A entrada adjectival, igualmente ilustrada na amostragem, tem, contudo, informações morfológicas diferentes das do nome: flexiona em género e número (111) e pode receber, como acontece com determinados adjectivos, o sufixo do superlativo *-íssimo*, representado no DELAS por *ss018*. O advérbio *cedo* pode igualmente ser quantificado por meio de um sufixo superlativo (ss1), e aceitar também sufixos de diminutivo (dh1_dt1): *cedinho*, *cedito*. No que diz respeito às entradas verbais, o código numérico refere-se ao modelo de conjugação do verbo. Por exemplo, 101, associado a *comprar*, indica que se trata de um verbo regular da primeira conjugação. A transitividade do verbo, e, portanto, a possibilidade de se construir com clíticos, está representada pelo código *t*.

Como se referiu antes, as entradas do DELAF são formas flexionadas. Neste dicionário a informação está estruturada do seguinte modo: as formas flexionadas estão associadas aos respectivos lemas (*campeã,campeão*); em seguida especifica-se a categoria gramatical e a informação morfológica da entrada. Por exemplo, *campeã,campeão.N:fs* indica que *campeã* é a forma feminina do singular do nome *campeão* (N:fs). No caso de um determinado lema dar origem a formas com diferentes atributos morfológicos, essas informações são delimitadas por ':' e representadas sequencialmente. Esta é uma situação que se observa com bastante frequência, sobretudo nos verbos. Por exemplo, a entrada *comprar* (V:R:U1s:U3s:W:V1s:V3s) pode ser uma forma de infinitivo impessoal, de infinitivo flexionado (1ª e 3ª pessoas do singular) e de futuro do conjuntivo (1ª e 3ª pessoas do singular).

2.2. Dicionário de palavras compostas: Módulos DELAC e DELACF

As palavras compostas, isto é, as unidades lexicais formadas por uma sequência de palavras simples e de separadores adequados (espaço, hífen e apóstrofo, no caso do português) estão formalizadas no DELAC. Este dicionário contém cerca de 50 000 entradas, na maioria nomes e advérbios compostos. A estrutura das entradas é idêntica à das palavras simples, como se ilustra a seguir.

DELAC

de mão beijada, ADV+PCA
a respeito de,PREP
artista(N101) plástico(N001),N+NA
livro(N201) branco(A201),N+NA
direitos(N292) de autor,N+NDN

Os compostos pertencentes a categorias invariáveis (a maioria dos advérbios, preposições, conjunções e certos determinantes) são seguidos da informação sobre a sua categoria gramatical, a que se segue a especificação categorial dos elementos que estão na sua origem.

Em relação aos nomes e adjectivos, embora haja casos de invariabilidade total, como acontece, por exemplo, com o nome composto *direitos de autor* (um nome masculino exclusivamente plural), a maioria pode flexionar em género e/ou número, como acontece com o nome *artista plástico* :

artista(N101) plástico(N001),N+NA

A informação categorial do composto (N) é seguida da especificação da sua estrutura interna: *NA* (*Nome* e *Adjectivo*). Os elementos que podem flexionar estão marcados com o código de flexão correspondente. Neste caso em particular, os constituintes do composto têm um comportamento morfológico idêntico ao que apresentam como palavras simples. Assim, eles têm os mesmos códigos de flexão do DELAS. Porém, na maioria dos casos, os constituintes dos compostos possuem um código de flexão diferente daquele que os caracteriza enquanto unidades lexicais simples. O nome composto *direitos de autor*, por exemplo, ilustra bem esse comportamento. No dicionário de palavras compostas, o primeiro elemento do nome composto está marcado como sendo uma forma do plural (**direito de autor*); a não atribuição de um código de flexão ao segundo indica, por outro lado, que se trata de uma forma que, como constituinte do composto, não flexiona nem em género nem em número: **direitos de autora*; **direitos de autores;* **direitos de autoras*.

Segue-se uma amostragem do DELACF, gerado automaticamente a partir do DELAC, constituída pelas entradas nominais acima representadas:

DELACF

artista plástica,artista plástico.N+NA:fs
artistas plásticas,artista plástico.N+NA:fp
artista plástico,artista plástico.N+NA:ms
artistas plásticos,artista plástico.N+NA:mp
livro branco,livro branco.N+NA:ms
livros brancos,livro branco.NA:mp
direitos de autor,direitos de autor.N+NDN:mp

À excepção de *livro branco*, que, para além de nome composto, pode ser analisado como uma combinação livre de palavras, todas as unidades lexicais da amostragem são não ambíguas. É, portanto, desejável que o seu processamento seja feito tão cedo quanto possível, a fim de que sejam mais tarde analisadas como combinações livres de palavras e categorias gramaticais. Os compostos não ambíguos estão, pois, listados num dicionário próprio, que é aplicado na fase do pré-processamento do texto. A sua aplicação, nesta fase, permitirá atribuir a cada uma das sequências por ele identificadas uma etiqueta lexical única, impedindo, deste modo, uma futura análise dos seus constituintes como palavras simples.

O mesmo procedimento não é, naturalmente, adequado para tratar as combinações lexicais que, como *livro branco*, tanto podem ser analisadas como fixas (palavras compostas) ou livres (sequência de palavras simples). Neste caso, 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[vi].

2.3. Dicionário de siglas

Além destes léxicos gerais, estão a ser elaborados outros léxicos mais específicos. As siglas, por exemplo, constituem um desses léxicos. Estas unidades lexicais são frequentes em diversos tipos de textos, em particular nos textos jornalísticos. Ocupam posições sintácticas idênticas às dos nomes, por exemplo, as posições de sujeito ou de complementos de predicados verbais, nominais e adjectivais. Dada a sua especificidade linguística (sequências de letras que representam sequências de palavras), as siglas colocam problemas particulares de formalização. De um ponto de vista estritamente formal, siglas como: *ONU*, *UE*, são palavras simples. Contudo, o seu desenvolvimento: *Organização das Nações Unidas*, *União Europeia*, corresponde a uma sequência de palavras que se presta a ser analisada como composta. É preciso, pois, relacionar entre si estes dois tipos de objectos linguísticos.

As siglas são, de acordo com a sua constituição formal, tratadas em vários módulos de dicionários interligados, os mais importantes dos quais estão representados nas seguintes amostragens:

SiglaS
ONU,onu.N+Sig:fs
UE,ue.N+Sig:fs
EUA,eua.N+Sig:mp

SiglasD
Organização das Nações Unidas,onu.N+DSig:fs
União Europeia,ue.N+DSig:fs
Estados Unidos da América,eua.N+DSig:mp

Cada entrada do dicionário *SiglaS* é, como se vê, constituída (i) pela sigla propriamente dita, grafada em maiúsculas; (ii) pela sua forma canónica, que se convencionou ser a própria sigla grafada em minúsculas; (iii) por um código morfossintáctico, que representa a parte do discurso da entrada (*N* para nome; *Sig* para sigla) e (iv) pela sua informação morfológica, que corresponde à flexão do determinante que precede a sigla.

As entradas do dicionário *SiglaD* são, neste caso, constituídas pelos nomes compostos que deram origem às siglas. De modo a relacionar os dois módulos de dicionários, convencionou-se que a forma canónica das entradas do *SiglaD* seria idêntica à das entradas do *SiglaS*, ou seja, a própria sigla grafada em minúsculas.

2.4. Tratamento de expressões numéricas

Certas unidades lexicais, simples e compostas, que não seria nem fácil nem natural enumerar em extensão e representar em dicionários, são descritas em transdutores. É, por exemplo, o caso dos números romanos: *II*, *III*, *IV*, *XXII*, *MCM*..., que formalmente são palavras simples, e dos determinantes numerais (cardinais e ordinais), a maior parte dos quais são palavras compostas: *vinte e um*, *vinte e duas*, *seiscentas e vinte e três*, *décima terceira*, *quadragésimo quarto*.

3. Homografia e ambiguidade lexical

Numerosas entradas do dicionário de formas canónicas são ambíguas. Vimos antes o caso de *central*, que, descontextualizado, tanto pode ser um nome (*uma central hidroeléctrica*) como um adjectivo (*uma questão central*). Nas 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. A título ilustrativo, veja-se a seguinte frase:

Ele comprou uma revista muito técnica

Embora não levante quaisquer problemas de interpretação a nenhum falante, quando processada por um sistema como o INTEX, utilizando somente as informações linguísticas dos dicionários, esta frase recebe várias análises possíveis, como se verifica pelo autómato construído automaticamente pelo sistema, após a aplicação dos dicionários:

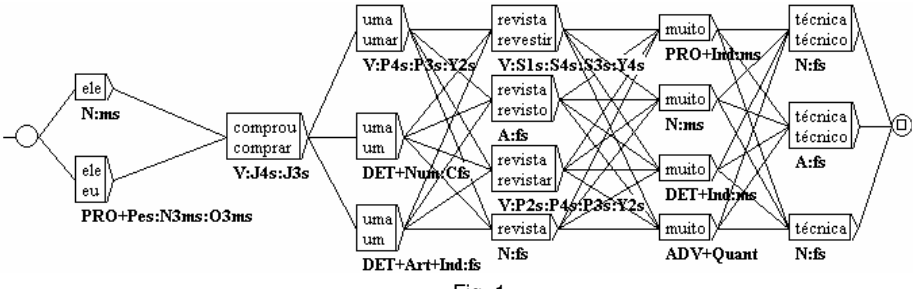

Fig. 1

Esta multiplicidade de análises advém do facto de todas as palavras desta frase terem atributos linguísticos diferentes e quase todas estarem associadas a distintas formas canónicas. Por exemplo, a palavra *revista*, além de um substantivo, com o significado que tem na frase, pode ter outros significados (*teatro de revista*, por exemplo); a mesma forma pode ainda corresponder ao adjectivo *revisto* (ele próprio associado aos verbos *revestir*, *rever* e *revistar*), a uma forma do presente do conjuntivo e do imperativo de *revestir* ou a uma forma do presente do indicativo e do imperativo de *revistar*.

Como se demonstrará a seguir, as ambiguidades lexicais condicionam a fiabilidade dos resultados de qualquer operação de processamento de texto, mesmo as mais simples. Assim, para eliminar, ou pelo menos reduzir, as análises fornecidas pelos dicionários que não são adequadas num determinado contexto, é necessário aplicar aos textos gramáticas de resolução de ambiguidades que especifiquem as restrições sintácticas e combinatórias que se estabelecem entre as várias categorias gramaticais (Ranchhod & Carvalho, neste volume).

4. Utilização das informações dos dicionários em processamento de *corpora*

Os primeiros recursos linguísticos a aplicar a um texto, quer tenha sido ou não objecto de pré-processamento, são os que estão formalizados nos dicionários de palavras flexionadas (simples e compostas) e nos FST lexicais.

Em seguida, ilustram-se alguns tipos de pesquisas simples que podem ser efectuadas em *corpora*, tirando proveito das informações desses recursos. Os resultados das pesquisas podem ser apresentados

pelo sistema INTEX de dois modos: destacados no texto (por exemplo, a negrito e a sublinhado) e/ou incluídos numa concordância (neste caso, as sequências reconhecidas estão individualmente destacadas numa linha e inseridas no seu contexto).

4.1. *Reconhecimento e contextualização de palavras*

O sistema permite, entre outras operações, visualizar todas as palavras simples e compostas de um dado texto, previamente identificadas pelos recursos lexicais. A título de exemplo, solicitou-se ao sistema que destacasse as unidades lexicais complexas que ocorrem no *Texto 1*, correspondente a um extracto de um artigo publicado na edição on-line do jornal «Expresso»[vii], de 7 de Setembro de 2002:

> O Tribunal Penal Internacional - para julgamento de crimes de guerra, contra a humanidade e de genocídio - nasceu a 1 de Julho de 2002, sob os auspícios da ONU, o boicote dos EUA e a não adesão da China, Israel, Índia, Indonésia, Paquistão, entre outros. Em 2000, os EUA tinham assinado, com mais 139 países, o Estatuto de Roma que cria o TPI. Por seu lado, Portugal aderiu, apesar de o TPI contemplar a pena de prisão perpétua que não tem cabimento no nosso ordenamento jurídico […].
> As liberdades e direitos dos cidadãos e dos povos, que haviam obtido estatuto de irreversibilidade universal e justificariam mesmo a introdução, na caótica «ordem internacional», do direito de ingerência humanitária, estão agora sob o fogo da maior potência mundial, num auge de agressividade ostensiva e assumida, ameaçando com os bombardeamentos (autêntico fogo de Deus) quem quer que a desgoste.

Texto 1

Como se pode observar, este pequeno texto contém um número considerável de compostos, o que vem confirmar, uma vez mais, a ideia de que uma boa parte do léxico de qualquer língua é constituído por palavras compostas, tal como ficou demonstrado por M. Gross (1986). E, coisa não desprezável, que uma parte do sentido de qualquer texto pode estar ancorada nas palavras compostas, sobretudo nos nomes, que contém.

Para fazer pesquisas mais específicas, podem utilizar-se as informações linguísticas formalizadas nas entradas dos dicionários. As siglas, por exemplo, podem ser extraídas do texto, através deste simples comando: <N+Sig>. No primeiro parágrafo do texto acima, os resultados são:

> nasceu a 1 de Julho de 2002, sob os auspícios da ONU, o boicote dos EUA e a não adesão da China, Israel, Índia, Indonésia, Paquistão, entre outros. Em 2000, os EUA tinham assinado, com mais 139 países, o Estatuto de Roma que cria o TPI. Por seu lado, Portugal aderiu, apesar de o TPI contemplar a pena de prisão perpétua que não tem cabimen

Para obter as ocorrências de qualquer forma do verbo *ter*, basta incluir o seu lema entre angulares: *<ter>*. Apesar de extremamente simples, este tipo de pesquisa revela-se de grande utilidade; permite, nomeadamente, determinar os diferentes valores sintácticos de um verbo, em função do tipo de complementação que tiver. Na concordância abaixo estão expressos dois valores do auxiliar *ter*:

```
aquistão, entre outros. Em 2000, os EUA tinham assinado, com mais 139 países,
mplar a pena de prisão perpétua que não tem cabimento no nosso ordenamento ju
```

Na primeira linha, *ter* é um auxiliar temporal do verbo *assinar*, com o qual forma um tempo composto. Na segunda, *ter* é também um verbo auxiliar, mas de um tipo diferente: trata-se de um verbo-suporte que forma com o nome predicativo *cabimento* um predicado nominal, cujo valor é idêntico ao dos predicados verbais (Ranchhod, 1990).

4.2. Identificação de expressões linguísticas descritas por expressões regulares e autómatos

Vários tipos de estruturas linguísticas (morfo-sintácticas, léxico-sintácticas) podem ser representados sob forma de expressões regulares ou autómatos. Por exemplo, a expressão regular:

(<ter> + <haver>) <V:K>

permite identificar nos textos os tempos compostos formados com os auxiliares *ter e haver*, seguidos do particípio de um verbo qualquer. Aplicada ao Texto 1, esta expressão extrai as seguintes construções:

```
dos cidadãos e dos povos, que haviam obtido estatuto de irreversibilidade univ
entre outros. Em 2000, os EUA tinham assinado, com mais 139 países, o Estatuto
```

Em todas estas operações de processamento de texto, simples exemplos do que é possível obter, foram exclusivamente utilizadas as informações contidas nos vários módulos de dicionários. Contudo, para pesquisas mais rigorosas e que envolvam maior complexidade sintáctica e semântica, os dicionários têm de ser aplicados aos textos em combinação com gramáticas adequadas aos fins pretendidos.

5. Conclusão

A crescente necessidade de aplicações cada vez mais complexas da linguística computacional tem posto em evidência a importância de dados linguísticos formalizados, em particular, de léxicos e gramáticas de grande cobertura. Em relação ao léxico, os dicionários têm de ser tão completos quanto possível, tanto do ponto de vista da sua cobertura lexical, como do ponto de vista da sua granularidade (descrição rigorosa dos atributos lexicais das entradas). Se os dicionários tiverem lacunas lexicais ou informação linguística insuficiente, isso comprometerá não só a análise lexical do texto, mas todas as fases de tratamento subsequentes.

Com o objectivo de dar resposta a essas necessidades, têm vindo a ser elaborados, para o português, léxicos computacionais de palavras simples e compostas, que apresentámos em breve síntese. As informações lexicais, formalizadas em transdutores de estados finitos (FST), podem ser imediatamente aplicadas à análise lexical de textos. As utilizações dessas análises são variadas; o processamento lexical é o primeiro passo de qualquer processamento linguístico de texto. Além destas aplicações gerais, os léxicos computacionais, com as características dos que apresentámos, podem ser utilizados com vantagem na própria investigação em Linguística: exploração de grandes *corpora* para estudar fenómenos de natureza morfo-sintáctica e léxico-sintáctica. 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 de textos põe em evidência dados linguísticos interessantes, que de outro modo seriam dificilmente detectáveis.

Referências

Baptista, Jorge (1995). *Estabelecimento e formalização de classes de nomes compostos*. Tese de mestrado, Faculdade de Letras da Universidade de Lisboa.

Carvalho, Paula. (2001). *Gramáticas de Resolução de Ambiguidades Resultantes da Homografia de Nomes e Adjectivos*. Tese de mestrado, Faculdade de Letras da Universidade de Lisboa.

Eleutério, Samuel; E. Ranchhod; H. Freire; J. Baptista (1995). «A System of Electronic Dictionaries of Portuguese», *Lingvisticae Investigationes*, XIX:2 Amsterdam/Philadelphia: Jonh Benjamins.

Gross, Maurice (1986). «Lexicon-Grammar. The Representation of Compound Words», *COLING-86*, Bona.

Moura, Paulo (2000). *Dicionário Electrónico de Siglas e Acrónimos*. Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa.

Ranchhod, Elisabete (1990). *Sintaxe dos Predicados Nominais com Estar*, Lisboa: INIC.

Ranchhod, E., Mota, C., Baptista, J. (1999). «A Computational Lexicon of Portuguese for Automatic Text Parsing», In *Proceedings of SIGLEX' 99: Standardizing Lexical Resources*, 37th Annual Meeting of the ACL, Maryland.

Ranchhod, Elisabete (2001), «O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais», In: E. Ranchhod (org.): *Tratamento das Línguas por Computador*, Lisboa: Caminho.

Silberztein, Max (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris: Masson.

Silberztein, Max (1997). «The Lexical Analysis of Natural Language», *Finite-State Language Processing*, Cambridge, Mass./London: MIT Press.

[*] Este trabalho foi, em parte, financiado pela FCT no âmbito do Projecto ENLEX – *Enhancement of Large-scale Lexicons*. Ref. POSI/PLP/34729/99.

[i] Alguns destes recursos estão disponíveis em: http://label.ist.utl.pt/

[ii] Ver http://www.nyu.edu/pages/linguistics/intex/

[iii] DELA é uma notação para **d**icionários **e**lectrónicos do **La**bEL; S corresponde ao módulo das palavras simples, F ao das formas flexionadas. Adoptou-se a mesma notação para os dicionários dos compostos (Cf. 2.2.): DELAC e DELACF, respectivamente.

[iv] Os verbos são representados pelo infinitivo; os nomes e adjectivos pela forma do masculino singular, quando aplicável, ou pelo feminino singular quando só têm essa forma. As categorias invariáveis (a maioria dos advérbios, preposições, conjunções e certos determinantes) são, naturalmente, representadas pela sua própria forma.
[v] As notações e formalismo são os do sistema INTEX.
[vi] Ver, neste volume, *Unidades lexicais complexas. Problemas de análise e etiquetagem*.
[vii] http://www.expresso.pt/