Dicionários electrónicos de léxicos terminológicos. "Seguros"

ELISABETE MARQUES RANCHHOD[*], CRISTINA MOTA[**]

Abstract

In this paper we discuss the issues raised by the integration in the dictionary modules of the system DIGRAMA of a set of Portuguese technical terms belonging to the area of Insurance.

We first refer to the linguistic analysis of those terms, to the formalisation of their properties and to how they were introduced in the system. Afterwards, we describe their implementation for Windows 95/NT, mentioning the main features of the database associated with the terminology.

Given the recognised importance of terminologies and, in particular, of those specifically developed for automatic use, it is our purpose to continue this work and to develop other terminologies.

[*] Faculdade de Letras da Universidade de Lisboa e Centro de Automática da Universidade Técnica de Lisboa, Instituto Superior Técnico
elisabet@label.ist.utl.pt

[**] Centro de Automática da Universidade Técnica de Lisboa, Instituto Superior Técnico
cristina@label2.ist.utl.pt

I WORKSHOP DA APL

SOBRE

LINGUÍSTICA COMPUTACIONAL

Lisboa, Maio de 1998

Dicionários electrónicos de léxicos terminológicos.

"Seguros"

Elisabete Ranchhod (FLUL/CAUTL)

Cristina Mota (IST/CAUTL)

1998

Dicionários electrónicos de léxicos terminológicos. "Seguros"[*]

Elisabete Ranchhod, FLUL/CAUTL[**]

Cristina Mota, IST/CAUTL[**]

Palavras-chave: dicionários electrónicos, terminologias, documentação automática.

1. Introdução

Na prática lexicográfica corrente, é costume distinguir os dicionários de termos técnicos dos que não têm essa especificidade. Assim, são, por um lado, elaborados dicionários de língua, ou de uso, e, pelo outro, dicionários especializados, ou terminologias, que contêm léxico próprio de uma área científica ou técnica. Esta separação justifica-se devido às dificuldades de ordem prática que colocaria a dicionarização do elevado número de termos técnicos existentes em qualquer língua, número que cresce incessantemente com o desenvolvimento científico e tecnológico.

Em relação aos dicionários electrónicos, contudo, essa dificuldade não se coloca, ou, pelo menos, não se coloca da mesma forma. Os actuais instrumentos informáticos permitem tratar e manipular grandes quantidades de dados lexicográficos, desde que estejam devidamente formalizados.

Neste trabalho, referir-nos-emos aos métodos de tratamento automático de um conjunto de termos pertencentes à área técnica dos Seguros, a fim de serem

---

[*] Estudo parcialmente financiado pelo Programa PRAXIS XXI (Proj. 2/2.1/CSH/775/95).

Gostaríamos de deixar aqui um agradecimento muito especial a Vítor Franca, que colaborou na apresentação oral que serviu de base a este artigo e na elaboração de algumas das informações específicas que constam da base de dados.

[**] 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/

[**] Instituto Superior Técnico/ Centro de Automática da Universidade Técnica de Lisboa.

Av. Rovisco Pais – P1096 LISBOA.

cristina@label2.ist.utl.pt

integrados nos módulos de dicionários do sistema DIGRAMA. A apresentação está organizada em duas partes: na primeira, proceder-se-á a uma breve análise dos termos e indicar-se-á o modo como foram introduzidos no sistema; na segunda, será descrita a sua implementação em Windows 95/NT, mencionando em particular as características da base de dados associada à terminologia.

2. Léxicos terminológicos do sistema DIGRAMA

O léxico técnico (terminologia) de que iremos falar é constituído por 632 termos. Estes dados foram recenseados em vários tipos de apólices por Vítor Franca, que os analisou e formalizou (Vítor Franca, 1997). As soluções encontradas para a sua integração no sistema de dicionários DIGRAMA[1] são igualmente válidas para tratar dados terminológicos pertencentes a outras áreas técnicas ou científicas.

Antes de abordar a questão da formalização dos dados e do seu posterior tratamento informático, referiremos o modo como se articulam os dicionários terminológicos e os dicionários gerais (não terminológicos).

A separação de um e de outro tipo de dicionários é, como se aludiu acima, frequentemente adoptada. Em consequência, a análise automática de um texto técnico necessita de recorrer à utilização de um dicionário terminológico do domínio, mas não pode evitar a consulta de um dicionário geral. Na verdade, embora não existam estatísticas sobre o assunto, não é difícil verificar que qualquer texto técnico contém, em maior ou menor grau, vocabulário corrente. A consulta aos dois tipos de dicionários, não parece trazer qualquer inconveniente, se se tiver em conta que nos nossos dias os computadores já podem tratar com relativa rapidez quantidades de dados de dimensões apreciáveis. No entanto, esta solução encontra outro tipo de dificuldade. Se é verdade que os textos técnicos fazem uso do léxico geral, não é menos verdade que os textos não técnicos podem incluir léxico terminológico. Há termos que são frequentemente usados e aparecem com naturalidade em qualquer tipo de texto, prestando-se, assim, a figurar num dicionário geral. Se aí não estiverem, eles não serão reconhecidos em operações de análise de texto. É o caso, para dar exemplos da terminologia dos Seguros, de *companhia de seguros* ou *seguro*

[1] Para uma breve caracterização dos dicionários de palavras simples e compostas do sistema DIGRAMA, ver o artigo de E. Ranchhod neste volume.

de vida. Para evitar o inconveniente do não reconhecimento de uma palavra que, sendo um termo, o uso consagrou como vocábulo corrente, as entradas que estivessem nesta situação teriam de figurar no dicionário geral e nos dicionários terminológicos. Isso duplicaria indesejavelmente as entradas e criaria outra dificuldade não trivial: a de decidir quais os termos que devem fazer parte do léxico geral e quais os que aí não têm cabimento. O problema da duplicação de entradas ver-se-ia agravado pelo facto de alguns termos pertencerem a mais do que um domínio técnico ou científico, sem que haja argumentos convincentes para a sua hierarquização.

No que respeita à introdução de terminologias no sistema DIGRAMA, tomámos a decisão de integrar os léxicos terminológicos, devidamente identificados, nos correspondentes dicionários de palavras simples e compostas. Esta opção evita alguns dos inconvenientes que se mencionaram, mas só poderá ser plenamente justificada pelos resultados experimentais obtidos por análise de textos técnicos e não técnicos de grandes dimensões (vários milhões de palavras).

2.1. *Introdução dos termos nos dicionários do sistema*

No sistema DIGRAMA, os léxicos terminológicos recebem, pois, um tratamento em tudo idêntico ao do restante léxico. No que diz respeito à terminologia dos *Seguros*, os termos por que é actualmente constituída foram, depois de adequadamente identificados como tal, integrados nos módulos dos dicionários a que pertencem: (i) os termos constituídos por palavras simples, como *apólice*, *sobreprémio*, etc. figuram no DIGRAS; (ii) os termos que correspondem a palavras compostas: *apólice aberta*, *companhia de seguros*, etc., foram incluídos no DIGRAC. Adicionalmente constituiu-se uma base de dados com informações terminológicas específicas, que vão sendo progressivamente melhoradas e completadas, cujo funcionamento, autónomo ou não, será descrito mais adiante.

Como acontece frequentemente com os léxicos terminológicos, os elementos da terminologia dos seguros são na sua esmagadora maioria constituídos por nomes compostos. A codificação destes termos para utilização automática não difere

substancialmente da que foi definida para tratamento dos nomes compostos[2] em geral. Os nomes compostos técnicos foram, de acordo com a sua constituição interna, integrados em classes formais, o que, entre outras vantagens (por exemplo, especificação de várias zonas de pesquisa), permite prever comportamentos morfológicos típicos. Formalizaram-se e codificaram-se as regras de variação flexional de cada termo; especificou-se o género global do composto. No que respeita aos termos que são palavras simples, os métodos de tratamento são também idênticos aos das palavras simples correntes.

Todos os termos, simples e compostos, são identificado com o seu domínio técnico ou científico através de um código específico; no caso dos Seguros, a esta informação corresponde o código *TS*. Estes códigos identificadores podem ser usados em combinação, a fim de permitir dar conta da pertença de um termo a mais do que uma terminologia (situação frequente).

2.1.1. Formato das entradas

Tomando como exemplos os termos utilizados antes: *apólice*, *sobreprémio*, *apólice aberta* e *companhia de seguros*, as correspondentes entradas lexicais têm o seguinte formato:

apólice,N300,TS
sobreprémio,N200,TS
apólice aberta, N,NA1301,TS
companhia de seguros, N,NDN300,TS

Todas são nomes (N); *apólice* é exclusivamente feminino (300), *sobreprémio* é exclusivamente masculino. As entradas compostas pertencem também aos nomes, sendo a sua classe formal indicada pelo código alfabético que, separado por vírgula, segue imediatamente o código categorial (NA: nome + adjectivo; NDN: nome + de + nome). A zona numérica contém informações sobre a flexão do composto e sobre o seu género global. Os dois constituintes do composto *apólice aberta* são flexionáveis no plural, sendo o género global do composto idêntico ao do primeiro elemento (1301); quanto a *companhia de seguros*, só a primeira palavra é susceptível de

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

variação, variação do mesmo tipo da que apresenta quando não faz parte de um composto; o género global do termo é também idêntico ao do primeiro constituinte.

Para além desta descrição gramatical, a base de dados correspondente, contém vários tipos de informação complementar, organizada por diferentes campos. Parte desta informação (nomeadamente a que está contida no campo *notas técnicas*) não está codificada e destina-se a ser exclusivamente usada por utilizadores humanos (em ambiente informatizado, naturalmente).

3. Características gerais da base de dados terminológica

Tal como já foi referido anteriormente, os dicionários electrónicos terminológicos encontram-se integrados no sistema DIGRAMA, fazendo parte dos dicionários electrónicos gerais. A formalização de um termo contém um código que indica a que terminologia(s) pertence esse termo. A informação terminológica está organizada numa base de dados que contém a informação específica de cada termo, colocada nos seguintes campos:

- **Nome**: identificação do termo na sua forma canónica. Exemplo: "*seguro de saúde*".
- **Notas Técnicas**: breve descrição do valor do termo, que corresponde, tanto quanto possível, a uma definição precisa do mesmo. A nota técnica de *seguro de saúde* contém a seguinte informação: "*Modalidade que garante a comparticipação em despesas com a doença ou, em caso de internamento hospitalar, por doença ou acidente, para além de um período de carência estipulado, o reembolso de despesas inerentes à necessária assistência cirúrgica, medicamentosa e outras.*".
- **Variação gramatical**: especifica as possibilidades de variação morfológica dos termos. No caso que nos serve de exemplo, o termo permite a flexão plural do primeiro constituinte: "*seguros de saúde*".
- **Termos Equivalentes**: lista de termos que têm o mesmo valor do termo que estiver a ser descrito, no caso, "*seguro de doença*". De notar que nem sempre é possível preencher este campo, porque, na maior parte dos casos, cada termo corresponde a um conceito único. Nesta situação frequente, o campo será preenchido com "*(Nenhum)*".

– **Termos Relacionados**: lista de termos que estão de alguma forma relacionados com o termo que estiver a ser descrito e cuja consulta poderá ser indispensável para uma melhor compreensão do mesmo. Trata-se de um conjunto de termos unidos pelo mesmo campo nocional. Retomando ainda o mesmo exemplo, encontram-se neste campo os termos: "*seguro de internamento hospitalar*", "*seguro de assistência médica hospitalar*", "*seguro indemnizatório*", "*seguro de indemnização*", "*seguro social*", "*seguro de acidentes e doença*", "*seguro de cuidados de saúde*". Caso esta lista seja vazia, o campo será preenchido igualmente com "*(Nenhum)*".

4. Acesso aos dicionários e à base de dados

A Figura A representa esquematicamente o modo como se inter-relacionam os vários módulos de dicionários do sistema e sua articulação com a base de dados.

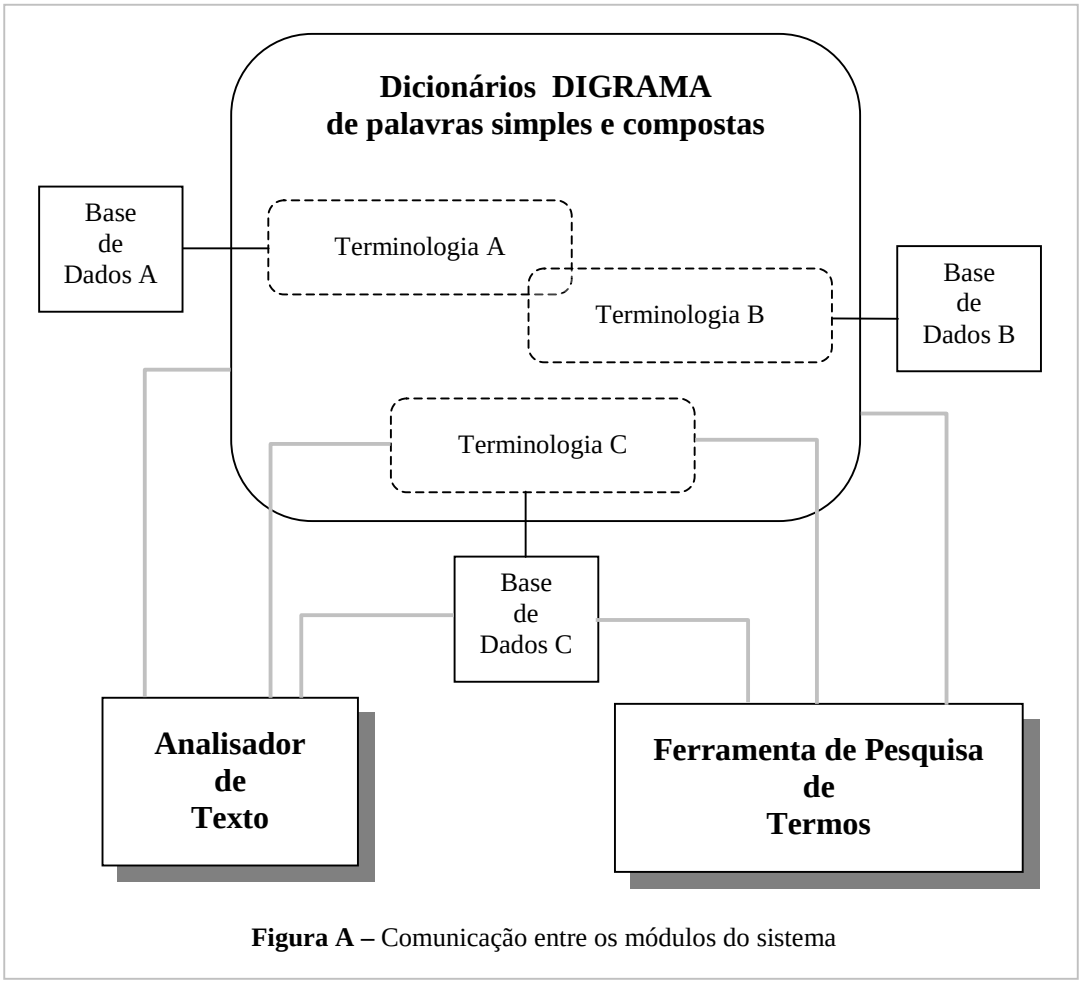

Figura A – Comunicação entre os módulos do sistema

Como se verifica, os vocábulos de uma terminologia fazem parte dos dicionários gerais respectivos, possuindo um código próprio que indica em que base de dados se encontra a informação terminológica específica. Dado que um termo pode pertencer a mais do que uma terminologia, essa informação é dada pela combinação de mais do que um código.

4.1 *Ferramentas para utilização das terminologias*

As ferramentas elaboradas para manipulação das terminologias podem pesquisar os dicionários de duas formas: acedendo ao dicionário geral e procurando todos os termos que têm um código que os identifica como elementos da(s) terminologia(s) pretendida(s) ou, então, acedendo apenas a um sub-dicionário (constituído, no caso dos Seguros, pelas entradas marcadas *TS*), extraído automaticamente dos dicionários gerais. A pesquisa na base de dados pode ser feita de forma autónoma ou passando primeiro por uma pesquisa nos dicionários (ver 4.1.1 e 4.1.2).

Passaremos, então, a descrever o funcionamento da ferramenta de pesquisa de termos desenvolvida para Windows 95/NT e que foi testada com uma terminologia de seguros, embora esteja concebida para qualquer outra terminologia.

4.1.1 Pesquisa de termos

Como se pode ver pela Figura B, a interface é bastante simples e fácil de usar. O utilizador apenas tem de ir seleccionando termos e consultando a informação respectiva (operações descritas nos pontos 2, 5 e 6). Em alternativa, este poderá introduzir um termo para que a ferramenta o encontre (operação descrita no ponto 7).

Estes dois tipos de utilização correspondem a dois tipos de pesquisa diferentes. Enquanto que no primeiro, a ferramenta pesquisa somente a base de dados, fazendo um acesso directo à mesma, no segundo a ferramenta faz a pesquisa em duas fases: inicialmente consulta os dicionários, extraindo a forma canónica do termo introduzido e, em seguida, com base nessa forma, faz, então, a pesquisa na base de dados.

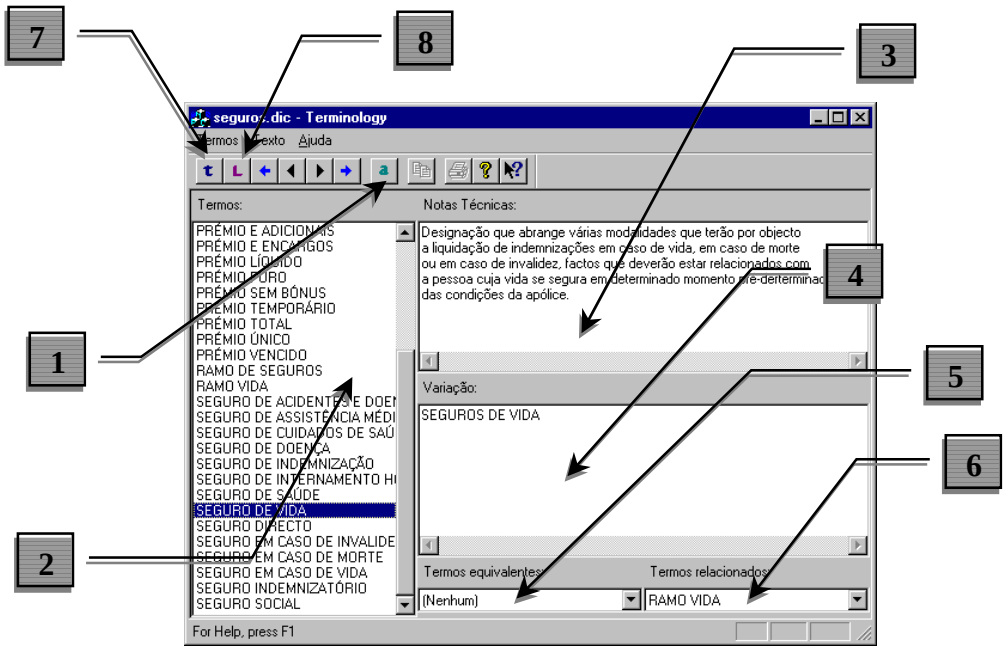

Figura B – Funcionalidade da ferramenta de pesquisa

1 -- Corre a ferramenta de análise de texto (ver 4.1.2).

2 -- Lista dos termos que constituem a terminologia. Seleccionando um termo desta lista, visualiza-se a sua informação terminológica na janela.

3 -- Descrição técnica do termo seleccionado.

4 -- Informação sobre as variações flexionais que o termo permite.

5 -- Lista de termos equivalentes ao termo seleccionado, no caso de existirem. Seleccionando um termo desta lista, equivale a seleccioná-lo na lista de termos, ou seja, passa-se a visualizar a informação referente a esse termo.

6 -- Lista de termos relacionados com o termo seleccionado. Também é possível seleccionar um termo desta lista para visualização.

7 -- Permite ao utilizador introduzir um termo para visualização, o qual passa a estar seleccionado na lista de termos. A informação específica terminológica é actualizada na janela.

8 -- Trata-se de um filtro que o utilizador pode usar para visualização de informação parcial e específica, por exemplo, alistar: (i) os termos que comecem por uma dada letra; (ii) os que estejam num dado intervalo; (iii) os que tenham uma determinada estrutura sintáctica; (iv) os que contenham uma determinada sequência de caracteres, etc.

Para além da selecção directa de um termo, é possível avançar para o termo seguinte e anterior por ordem alfabética, usando as setas (3e4) que estão na barra de ferramentas. Também é possível avançar ou recuar por ordem de consulta, usando as setas com cauda (⇦ e ⇨).

4.1.2 Análise de texto

Com a ferramenta de análise de texto, é possível aplicar qualquer dicionário do sistema DIGRAMA a um texto. Porém, quando o dicionário aplicado corresponde a uma terminologia, os termos que ficam em evidência são os que constam da terminologia escolhida. A informação terminológica específica correspondente pode ser solicitada a partir da selecção de um desses termos no texto em análise.

Por exemplo, na Figura C aplicou-se, numa primeira fase, a terminologia dos seguros ao texto, surgindo a janela 1 que corresponde à indexação dos termos encontrados (e que estarão em evidência no texto). Em seguida seleccionou-se o termo "*seguros de vida*" para visualização, surgindo a janela 2 que contém a informação específica desse termo.

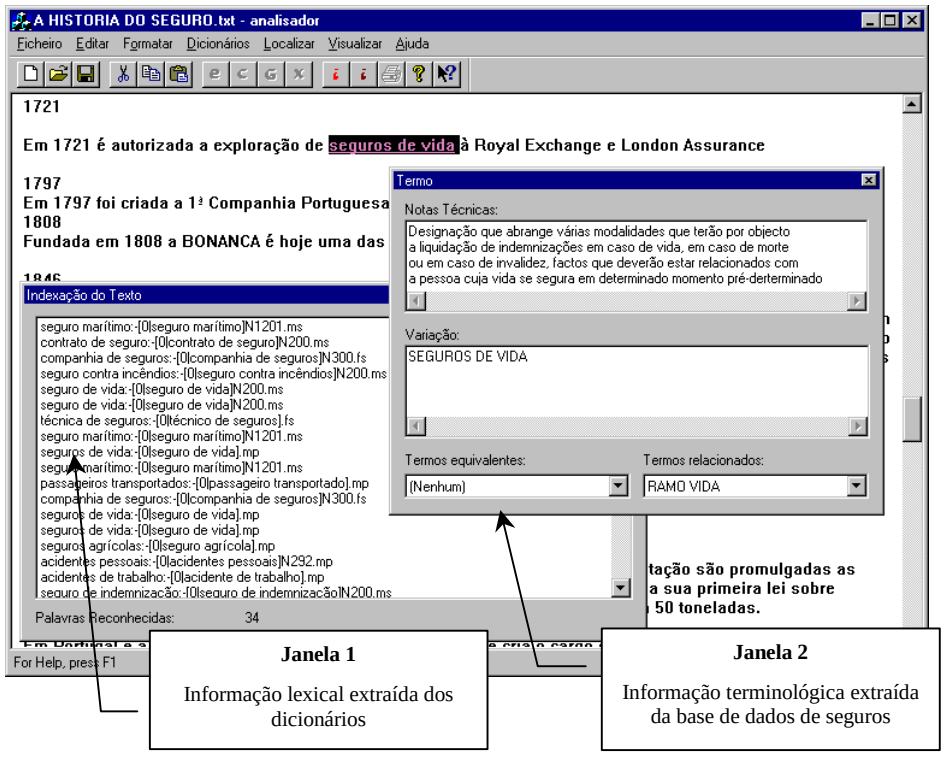

Figura C – Análise de um texto usando uma terminologia de seguros

Ao contrário do que acontece com a ferramenta de pesquisa de termos, o analisador de texto faz sempre primeiro uma pesquisa nos dicionários terminológicos para encontrar a forma canónica do termo e, só depois, é que faz a pesquisa na base de dados, nunca acedendo directamente à mesma.

5. Desenvolvimentos futuros

Dada a reconhecida importância das terminologias e, em particular, das que são elaboradas com o objectivo específico da sua utilização automática, é nossa intenção prosseguir este tipo de trabalho, alargando-o a novas terminologias e melhorando alguns aspectos da formalização linguística que ainda não estão adequadamente tratados. Mencionaremos apenas um que é da maior importância: a ocorrência de formas truncadas de termos mais extensos. Estas formas curtas, que podem resultar de truncaturas do termo tanto à esquerda como à direita, aparecem frequentemente nos textos, sendo imprescindível, por diversas razões, ligá-las à forma completa de que são uma redução.

Do ponto de vista da implementação em Windows, dever-se-ão introduzir a breve trecho funcionalidades novas, entre elas, as que permitam criar e editar terminologias, visando facilitar ao utilizador (e em particular aos linguistas) a construção e manutenção das mesmas.

Outro objectivo é equipar a base de dados com informação multimédia, a fim de tornar mais atractivo e mais claro o conteúdo das terminologias.

BIBLIOGRAFIA

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

BAUDOT, Jean (1984), A model for a Bilingual Terminology Minibank, *Lebende Sprachen*, nº2, Munique: Langensheidt (pp. 49-54).

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).

FRANCA, Vítor (1997), *Um léxico terminológico: Seguros*, Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa (206p).

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

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. (1998), Dicionários e análise lexical automática. In *Actas do Workshop sobre Linguística Computacional da APL* (no prelo).