Bases digitais lexicais na União Europeia*

Margarita Correia

(FLUL / ILTEC / SILEX URA 382 CNRS)

Nota prévia

Antes de iniciar, gostaria de felicitar os organizadores deste Simpósio pela sua realização, a vários títulos oportuna. Oportuna pela necessidade de promover os estudos do léxico da língua portuguesa, que tão carecida deles se revela, sobretudo quando comparada com outras línguas de difusão semelhante; oportuna pelo estreitamento de contactos propiciado entre os nossos dois países; oportuna, finalmente, por constituir a oportunidade de sensibilizar a comunidade científica brasileira para os desenvolvimentos que a Lexicografia Computacional tem conhecido na Europa nos últimos anos, área de estudos, que, tanto quanto me é dado conhecer, não se encontra desenvolvida presentemente no Brasil.

Gostaria ainda de tornar público o meu agradecimento aos organizadores pelo convite que me foi endereçado, bem como pelo empenho e esforço demonstrado de forma a garantir a minha presença entre vós. Muito obrigada.

0. Introdução

A presente comunicação tem como principal objectivo dar conta dos desenvolvimentos que a Lexicografia Computacional tem conhecido no Velho Continente, designadamente nos países integrados na União Europeia. A Lexicografia Computacional ocupa-se da descrição das unidades lexicais de uma língua em bases informáticas altamente estruturadas; em geral, as unidades terminológicas não merecem uma atenção particular nesta disciplina, constituindo, sim, o objecto da Terminografia Computacional ou Terminótica, ou sendo tratadas, no quadro de projectos como o MULTILEX, ao mesmo nível que as unidades que constituem o vocabulário corrente da língua.

Devido às condicionantes de tempo, a presente comunicação não mais será que uma primeira abordagem dos objectivos e métodos desta disciplina, bem como de alguns projectos de bases de dados contendo informação lexical, representativos do panorama da Lexicografia Computacional contemporânea, a nível europeu. Não será dada atenção especial a pormenores técnicos, sobretudo os que dizem respeito à vertente propriamente computacional destes projectos.

Num primeiro momento, esboçarei uma tipologia de bases digitais lexicais (1.), tentando delimitar os conceitos de *dicionário 'machine-readable', dicionário 'machine-tractable'* (propondo ao mesmo tempo equivalentes portugueses para estes termos) (1.1. e 1.2.), de *base de dados lexicais* (1.3.) e de *base de conhecimento lexical* (1.4.). Para a delimitação destes conceitos, tomarei como critérios os utilizadores destes produtos, os tipos de utilização possíveis e, ainda, a estruturação e formalização da informação. Em seguida explicitarei o conceito de *bases digitais lexicais*, que constitui o título desta apresentação (1.5.).

* Agradeço à Paula Guerreiro, investigadora do ILTEC, pela leitura atenta de versões prévias deste artigo, pelas valiosas sugestões e comentários feitos, bem como pelas frutuosas discussões havidas durante o nosso trabalho conjunto no Projecto GENELEX. No entanto, qualquer falha ou incorrecção é apenas da minha responsabilidade.

Num segundo momento referirei as vantagens que podem advir da construção de bases digitais lexicais para as várias línguas (2.).

Num terceiro momento, exporei algumas das ideias fundamentais que orientam na actualidade os projectos deste tipo desenvolvidos no seio da União Europeia, nomeadamente os que têm a ver com as possíveis metodologias que possam conduzir a uma economia dos custos envolvidos, passando pela definição prévia dos usos a fazer destas bases (3.1.), as metodologias para carregamento das bases (3.2.), a possibilidade de reutilização da informação armazenada (3.3.) e os formalismos utilizados (3.4.). Na sequência da exposição darei conta da iniciativa EAGLES, promovida pela União Europeia, cujo principal objectivo é a normalização dos projectos a desenvolver no seu seio (3.5.).

Num quarto momento, serão sumariamente apresentados três projectos a decorrer neste momento na Europa: o MULTILEX (4.1.), o ACQUILEX (4.2.) e o GENELEX (4.3.). Maior atenção será dedicada ao GENELEX, por ser aquele em que me encontrei directamente envolvida, embora, por limitações de tempo, apenas me possa deter numa descrição muito sumária do modelo (4.3.1.) e de algumas particularidades da sua camada morfológica (4.3.2.).

1. Esboço de uma tipologia: dicionários '*machine-readable*', '*machine-tractable*', bases de dados lexicais e bases de conhecimento lexical

Um dos problemas que se colocam a quem inicia a sua investigação no domínio das descrições computacionais do léxico é o da indeterminação terminológica. A necessidade de estabelecimento de uma terminologia na área da Lexicografia Computacional é já referida em CORREIA, M. & P. GUERREIRO (1993c), onde, de resto, são apresentadas tipologias de dicionários em suporte informático, com base no *critério processual*, no *critério utilizador final* e, ainda, *no critério possibilidade de alteração de dados*. Porém, tal como acontece com os dicionários impressos, estas diferentes tipologias tendem a sobrepor-se em função da concepção e características de cada objecto em particular.

As denominações usadas na bibliografia desta área não correspondem, em geral, aos mesmos conceitos (cf., por exemplo, BRISCOE, T. (1991); CALZOLARI, N. (1990); MARTIN W. & M. WOLTERING (1989); WILKS, Y et al. (1988 e 1993)). Porém, dado a discussão da terminologia não ser o objectivo fundamental da presente comunicação, utilizarei como denominações básicas as de dicionário legível com máquina (termo equivalente que aqui proponho para o termo inglês *machine-readable dictionary*), dicionário tratável por máquina (equivalente proposto para *machine-tractable dictionary*), base de dados lexicais (de *lexical database*) e base de conhecimento lexical (equivalente proposto para o inglês *lexical knowledge base*) cujas características passo a apresentar.

1.1. Dicionários legíveis com máquina

A oposição entre dicionário legível com máquina (de agora em diante, DLM) e dicionário tratável por máquina (de agora em diante, DTM) é proposta por WILKS, Y. *et al*. (1988 e 1993).

Os DLMs são dicionários coligidos por lexicógrafos e concebidos para uso humano. São basicamente dicionários que, ou foram inicialmente digitados em suporte magnético (seja em processadores de texto ou em bases de dados concebidas para tal), ou foram coligidos em verbetes de papel e posteriormente transferidos para suporte magnético. Destes dicionários são publicadas versões impressas e versões em suporte informático, armazenadas em bases de dados (seja o seu

suporte físico o CD-ROM, seja a disquete)[1]. A denominação DLM pode, portanto, corresponder a produtos diferentes em termos de concepção e metodologia de trabalho, apresentando todos, porém, como denominador comum as características de serem concebidos para uso humano e de se encontrarem disponíveis em suporte informático (sob a forma de produtos comercializáveis ou de bases para investigação e/ou aprendizagem).

A estrutura interna destes dicionários é semelhante à dos dicionários impressos, isto é, basicamente as unidades lexicais são descritas em artigos distintos, apresentando estes a estrutura tripartida clássica:

entrada - categoria - definição (eventualmente, exemplificação)

O índice de formalização da informação nestes dicionários é bastante baixo, limitando-se geralmente à utilização de abreviaturas, de tipos de caracteres diferentes e de determinadas convenções para, por exemplo, distinguir acepções distintas de uma mesma entrada, introduzir informação colocacional, etc. A descrição das unidades é sempre feita em linguagem natural e a forma de abordagem é predominantemente semasiológica (como acontece com os dicionários impressos).

O facto de serem armazenados em bases de dados contribui para potenciar toda uma rede de relações morfológicas, sintagmáticas ou paradigmáticas entre diferentes unidades lexicais, que nos dicionários impressos é apenas perceptível através de remissões, de definições sinonímicas, etc. Por outro lado, a ordenação alfabética das entradas, único meio para a sua localização nos dicionários impressos, perde em parte a sua validade nos DLMs, dado que se pode aceder à informação lexical por outras vias que não a entrada lexical.

Porém, embora beneficiando das virtudes da apresentação em suporte informático, que se traduzem em grande diversificação e aumento das possibilidades de consulta, estes dicionários não são susceptíveis de ser utilizados directamente em sistemas de processamento da linguagem natural (PLN), devido fundamentalmente ao facto de serem concebidos para uso humano[2]: como é bem sabido, existe um 'pacto implícito' entre o lexicógrafo e o consulente - o primeiro conta com a competência linguística do segundo para para colmatar toda a informação omissa ou apenas dada implicitamente.

Além disso, a informação é dada em linguagem natural, pouco formalizada, não reconhecível pelos programas de PLN, que pressupõem grande formalização da informação.

1.2. Dicionários tratáveis por máquina

Embora WILKS, Y. *et al*. (1993) integrem no domínio dos DTMs projectos cuja metodologia prevê uma completa codificação manual da informação (cf. a referência ao Projecto CYC, p. 343) - a *demo-approach* (cf. 3.2.), estes autores definem um DTM como um DLM transformado, apresentando um formato que o torne apto a ser usado em sistemas de PLN (cf. p. 341). Esta

[1]Para o português do Brasil, existe a versão informatizada da 2ª edição do *Novo Dicionário da Língua Portuguesa*, sob o título *Aurélio Electrônico* (cf. Bibliografia), editada em disquetes para ambiente DOS. Embora a informação nela contida tenha sido extraída da versão impressa do dicionário, a forma como foi desenhada a base de dados que o suporta permitiu um considerável aumento das possibilidades de acesso à informação, tornando explícita muita da informação que na versão impressa se encontrava apenas implícita.
Para o português europeu, está anunciada a publicação em disquetes da versão informatizada da 7ª edição do *Dicionário da Língua Portuguesa*, da Porto Editora (cf. Bibliografia).

[2]Cf. CORREIA, M. & P. GUERREIRO (1993c) para uma exposição mais detalhada das virtudes da apresentação dos dicionários em suporte informático, bem como das consequências para a descrição lexical que derivam do facto de o produto ser concebido para utilizadores humanos.

aptidão resulta basicamente na descrição do conhecimento lexical num formalismo que o sistema possa facilmente reconhecer, traduzindo a informação que nos dicionários humanos surge em linguagem natural, bem como na explicitação de todo o conhecimento que nos dicionários para uso humano permanece implícito na descrição (como sejam a informação sintáctica, a colocacional, a estilística, ou outras). Os DTMs são, à partida, apenas utilizáveis em sistemas de PLN.

É opinião quase unânime dos especialistas neste domínio que uma das principais metas a atingir num futuro próximo pela Lexicografia Computacional é a obtenção de descrições de grandes porções do léxico de uma língua, dado que os produtos até agora existentes se limitam a protótipos, onde as entradas se encontram descritas de modo bastante pormenorizado, mas cujo número é extremamente limitado (cf. BOGURAEV, B. & T. BRISCOE (1989), p. 10)[3].

Para WILKS, Y. *et al*. (1993), a construção de um DTM deverá depender largamente da informação contida nos DLMs, por outras palavras, defende-se a *book approach* (cf. 3.2.), dado que os DLM contêm efectivamente uma quantidade não desprezável de informação acerca das unidades lexicais de uma língua e, dado a alternativa a este procedimento ser a codificação manual de extensas listas de unidades, a metodologia defendida afigura-se a mais rápida e a mais económica. Assim, a proposta destes autores é a de conceber mecanismos e metodologias que permitam a extracção, tão automática quanto possível, da informação contida nos DLMs e a sua importação para DTMs com formalismo pré-determinado e adaptável às aplicações a que servirão de base. Estas metodologias consistem na criação de analisadores (do inglês *parsers*) construídos para o efeito, capazes de extrair a informação contida nos DLMs sem dar azo a perdas significativas de informação. Qualquer destas metodologias pressupõe, no entanto, algum trabalho do lexicógrafo, seja na codificação manual de alguma informação, seja na verificação das suas validade e coerência.

O trabalho de WILKS, Y. *et al.* (1988 e 1993) tem por base a versão legível com máquina do *Longman Dictionary of Contemporary English* (LDOCE), dicionário estruturado e construído, segundo os seus autores, de forma mais coerente e sistemática que os restantes dicionários do mesmo tipo, o que se revela designadamente através do vocabulário controlado com que são redigidas as suas definições.

Dado este facto, pode verificar-se que os DTMs, tal como concebidos por WILKs, Y. *et al*. (1993), serão largamente dependentes do DLM que constitui a fonte de informação.

1.3. Bases de dados lexicais

Uma base de dados lexicais (de agora em diante, BDL) é uma estrutura computacional concebida de modo a ser capaz de suportar os mais variados tipos de conhecimento sobre cada unidade lexical, permitindo estabelecer conexões, quer entre unidades lexicais distintas, quer entre características pertencentes a unidades lexicais distintas. Por outras palavras, a estruturação de uma BDL permitir-nos-á observar as unidades lexicais sob os mais variados prismas e aceder a elas das mais variadas formas possíveis.

Do ponto de vista teórico, uma das características fundamentais das BDLs é o facto de corresponderem a uma concepção de léxico bastante diferente da dos dicionários: numa BDL, o léxico é entendido como uma complexa rede de relações (morfológicas, sintagmáticas, semânticas, paradigmáticas), onde o conhecimento sobre uma unidade lexical é composto a vários níveis ou camadas. Pelo contrário, nos dicionários em geral (dicionários impressos, DLMs ou DTMs), o léxico é encarado como uma listagem de unidades a descrever de forma atomística, não sendo potenciandas (ou, pelo menos, não de modo sistemático e/ou exaustivo) as relações interlexicais (cf. CALZOLARI, N. (1990)).

[3]Este é o caso, por exemplo, do dicionário construído para base das gramáticas do Projecto EUROTRA-P.

Deste modo, ao construir uma BDL, frequentemente um dos primeiros passos é pré-definir estruturas de especificações (do inglês *templates*) e regras lexicais gerais, cujos índices serão atribuídos, num momento posterior, às unidades lexicais a descrever. De certo modo, pode dizer-se que a descrição do léxico em BDLs é tendencialmente onomasiológica, uma vez que se parte do estabelecimento da definição das propriedades (morfológicas, sintácticas, semânticas, pragmáticas) das unidades lexicais, para, seguidamente, se proceder ao estabelecimento das correspondências entre essas propriedades e as unidades lexicais correspondentes.

Ao serem concebidas como estruturas multiplamente organizadas, com base em conhecimentos distintos sobre as unidades lexicais, as BDLs afiguram-se implicitamente como réplicas dos modelos de armazenamento lexical na memória humana, fornecidos pela Psicolinguística e em particular pela Psicolexicologia (cf. CORREIA, M. & P. GUERREIRO (1993c)).

No entanto, BRISCOE, T. (1991) situa a distinção entre BDLs e bases de conhecimento lexical precisamente ao nível do comprometimento teórico e do dinamismo: isto é, para ele, uma BDL representa a informação através de uma sintaxe menos rígida e coerente, tipicamente com uma semântica apenas implícita (cf. p. 58) e, além disso, é uma representação estática da informação lexical, por oposição ao dinamismo das bases de conhecimento lexical (cf. 1.4.).

O carregamento de uma BDL pode ser feito com base em DLMs, em DTMs, em corpora de dados textuais (orais ou escritos) em suporte informático, ou finalmente, pode ser digitada directamente pelo lexicógrafo computacional. A utilização de DLMs ou de corpora para carregamento das BDLs implica, porém, a construção de analisadores capazes de extrair o máximo de informação não formalizada; porém, é frequente haver perdas de informação significativas no processo de extracção.

A informação carregada numa BDL é primeiramente destinada a ser utilizada em sistemas de PLN (correctores sintácticos e/ou estilísiticos, programas de tradução automática, programas de reconhecimento de textos, etc.). Este facto faz com que a informação seja codificada segundo formalismos rígidos, não facilmente legíveis por humanos, geralmente representada em termos de pares 'atributo / valor', aos quais correspondem índices convencionais.

A possibilidade de acesso à informação nelas contida por parte de utilizadores humanos não é, no entanto, geralmente desprezada, podendo servir de base a produtos impressos resultantes do seu descarregamento em suporte-papel (dicionários gerais ou especializados, manuais de conjugação, etc.), ou a produtos em suporte informático (DLMs, correctores vários, programas de aplicação pedagógica, etc.). Para tal é necessário que sejam concebidas interfaces facilmente manejáveis pelo consulente, implicando a construção de programas de conversão de informação. É necessário, portanto, 'traduzir' o formalismo utilizado para linguagem acessível ao consulente. A este é, contudo, interdita a possibilidade de alterar a informação contida na BDL.

O objectivo dos projectos MULTILEX e GENELEX é a construção de BDLs, embora ambos apresentem algumas diferenças concretas que tentarei sumarizar adiante.

1.4. Bases de conhecimento lexical

O termo base de conhecimento lexical (equivalente proposto para o inglês *lexical knowledge base*, de agora em diante referido como BCL) é utilizado para designar o objectivo do projecto ACQUILEX (cf. 4.2.).

Segundo BRISCOE, T. (1991), já referido em 1.3., uma BCL representa explicitamente uma teoria do léxico, sendo, por isso, um corpo de informação representada num tipo de notação especial - a LRL (*lexical representation language*) -, que contém uma sintaxe e uma semântica explícitas e que suporta operações lexicais capazes de realizar transformações válidas dessa informação (cf. BRISCOE (1991: 58)).

Por outras palavras, enquanto que uma BDL é concebida como uma representação estática das propriedades das unidades lexicais extraível de DLMs, na perspectiva dos participantes no projecto ACQUILEX[4], uma BCL, por outro lado, é concebida como uma representação dinâmica, na medida em que, além de conter informação lexical estruturada, pressupõe a construção de LRL capaz de analisar essa informação e de gerar produções linguísticas. A definição dessa LRL é feita explicitamente de acordo com uma teoria semântica determinada. Além disso, se em relação às BDLs é já notória a ligação aos modelos propostos pela Psicolexicologia, em relação às BCLs essa ligação é explicitamente assumida.

Tal facto implica, por um lado, que o nível de estruturação de uma BCL seja muito mais elevado que o de uma BDL e, por outro, que uma BCL seja um produto muito mais elaborado e sofisticado do que uma BDL (cf. BRISCOE, T. (1991: 46)), que, de resto, será a base para a sua construção, sob a forma de um dicionário-fonte (cf. p. 48)[5].

Um dos aspectos que poderão ressaltar da exposição anterior é a dificuldade em estabelecer com nitidez a fronteira entre os vários tipos de bases digitais lexicais. Efectivamente, tal como acontece com os dicionários impressos, onde cada novo produto procurará acumular características de um ou de outro tipo, de acordo com as necessidades do público-alvo, cada projecto em particular configurará uma base digital lexical consoante os objectivos do seu trabalho, as utilizações que dele pretender fazer e as fontes que poderá utilizar para o seu carregamento.

1.5. Bases digitais lexicais

Dada a variedade de termos propostos, urge especificar o que se entende então por base digital lexical, termo que constitui o título do presente trabalho, de resto, proposto pela organização deste Simpósio.

Este termo, embora não consagrado pela bibliografia da especialidade, tem a virtude de ser suficientemente abrangente para incluir todos os anteriormente referidos.

Deste modo, entendo a denominação base digital lexical como um conceito mais genérico, podendo ser definido como:

> Qualquer base de dados informática contendo informações relativas às unidades lexicais de uma ou mais línguas naturais, podendo essa informação ser apresentada em formatos diversos e ser utilizada com objectivos também diversos.

Assim, este termo poderá ser considerado um equivalente do termo léxico computacional (do inglês *computational lexicon*), que frequentemente usarei como seu equivalente ao longo do presente texto.

2. Para que podem servir estas bases?

Neste ponto da apresentação, poderemos questionar-nos sobre a utilidade das bases digitais lexicais. Ao responder a esta questão, implicitamente responderemos a uma outra com ela relacionada:

---

[4]É de notar que, neste contexto, os conceitos de DTM (dicionário tratável por máquina) e BDL (base de dados lexicais) se confundem, dado que a BDL é concebida como o resultado da extracção e estruturação da informação lexical contida num DLM (dicionário legível com máquina).

[5]Retomar-se-á a caracterização das BCLs quando for feita a apresentação sumária do projecto ACQUILEX, em 4.2.

- a que se deve o interesse crescente que este tipo de projectos tem merecido nos últimos anos em vários países, entre os quais se incluem aqueles que integram a União Europeia?

Sem querer deter-me sobre as consequências do desenvolvimento tecnológico recente, resultando na necessidade de um intercâmbio cada vez mais acelerado de informação a nível mundial, de todos sobejamente conhecidas, gostaria de referir que este tipo de produções pode contribuir decisivamente para a aceleração desse intercâmbio.

As bases digitais lexicais, como grandes repositórios de informação multifacetada sobre as unidades lexicais, estruturada e reutilizável, constituem a base indispensável para a construção de qualquer aplicação em linguagem natural.

Em primeiro lugar, qualquer programa de PLN, seja ele um analisador, uma gramática, um gerador de texto, um programa de tradução automática, apenas pode ser construído com base num repositório de informação lexical, suficientemente ampla, estruturada e explícita. Daí que, nos últimos anos, se tenha assistido a um crescente interesse, ao nível da Linguística Computacional, pela construção de bases digitais lexicais, passíveis de conter um número progressivamente maior de unidades lexicais, descritas de modo cada vez mais exaustivo.

Por outro lado, estas bases podem ainda constituir o suporte para a construção de programas de índole pedagógica que contribuam para uma aprendizagem mais acelerada de línguas, em particular, de línguas estrangeiras, cuja necessidade é também sobejamente conhecida. Esses programas poderão assumir a forma de aplicações informáticas interactivas, ou mesmo a de manuais e dicionários de aprendizagem impressos.

De resto, de uma base digital lexical, com a ajuda de programas especificamente construídos para o efeito, é sempre relativamente fácil extrair a informação necessária para a concretização de um qualquer dicionário, impresso ou legível com máquina[6]. Dado o tipo de informação registada ser tão variado e completo, facilmente se afigura a possibilidade de poder extrair de uma base tipos de dicionários diversos, adequados ao público e aos objectivos pré-definidos. Por exemplo, de uma BDL como o GENELEX, desde que contenha um número significativo de unidades descritas (o que não é o caso presente), é possível extrair vários produtos monolingues, tais como dicionários de aprendizagem (com bastante informação gramatical explícita sobre as unidades lexicais), dicionários de verbos e/ou de conjugação verbal, *thesauri*, e mesmo dicionários de língua mais gerais.

Ainda, havendo pelo menos duas línguas para as quais sejam construídas BDLs ou BCLs obedecendo à mesma estrutura representacional, afigura-se possível extrair dessas bases produtos bilingues (ou multilingues) diversos, ou acoplar-lhes programas de tradução automática.

O acesso a uma BDL, pelos vários tipos de relações lexicais explicitadas, permite-nos apreender o léxico como uma estrutura altamente organizada e complexa, mais conforme à realidade dos factos lexicais do que a visão que deles temos a partir dos dicionários impressos que, pelos condicionalismos do seu suporte, nos dão uma visão do léxico enquanto listagem compartimentada de unidades.

Finalmente, e poderemos mesmo dizer *last but not least*, o crescente desenvolvimento deste tipo de projectos tem sido o resultado de um aprofundamento continuado dos nossos conhecimentos sobre o léxico, podendo dizer-se que a Linguística Computacional e mesmo a Engenharia Linguística muito têm beneficiado do desenvolvimento da Linguística. Em contrapartida, a Linguística também muito tem beneficiado deste desenvolvimento: pela necessidade de tornar cada vez mais explícita a informação lexical, de modo a poder ser descodificada pela máquina, o nosso conhecimento da estrutura e funcionamento do léxico tem vindo progressivamente a ser alargado, quer pela

[6]Por exemplo, o GENELEX, ao nível da sua camada semântica, prevê um campo onde a definição de cada acepção (correspondente a uma USEM) pode ser expressa em linguagem natural, o que contribui decisivamente para uma fácil extracção de um ou mais dicionários impressos da BDL.

'descoberta' de propriedades das unidades lexicais apenas intuitivamente sabidas, quer pela constatação de regularidades insuspeitadas sobre o léxico, resultante do tratamento exaustivo e necessariamente coerente de grande número dessas unidades.

Em conclusão, é meu entender que uma comunidade que pretenda ver atribuído à sua língua um estatuto internacional relevante deverá ponderar a viabilidade de realização de uma base digital lexical de dimensões razoáveis, devendo-se, no entanto, ter em conta algumas prerrogativas que tentarei apresentar em seguida.

3. Aspectos a ter em conta quando da construção de bases digitais lexicais

Um aspecto que não poderá ser negligenciado é o do elevado custo, quer em termos humanos e financeiros, quer em termos de tempo, que a construção de uma base digital lexical implica, particularmente uma BDL ou uma BCL.

Assim, antes de iniciar um qualquer projecto em Lexicografia Computacional, como de resto em Lexicografia tradicional, é necessário definir critérios para a sua realização, que têm a ver com a utilização que se pretende fazer do produto final, com a metodologia e as fontes a utilizar para o carregamento da informação, com a possibilidade de reutilizar a informação armazenada para novas aplicações e, ainda, com o formalismo a usar.

São estes os aspectos que a seguir abordarei.

3.1. Definição prévia dos usos a fazer destas bases

Tal como em relação aos dicionários impressos, um dos aspectos mais importantes a considerar é a definição prévia dos usos a dar ao produto final. Tal facto prende-se, como é bem sabido, com o facto de que toda a estrutura da base de dados (e podemos em última análise considerar que um dicionário impresso é uma base lexical não-digital), a selecção das unidades a tratar, bem como o nível de pormenorização da descrição feita, dependem directamente desses usos.

No âmbito de uma base digital, particularmente no caso das que forem destinadas a sistemas de PLN, essa definição prévia adquire redobrada importância, dado que toda a informação é introduzida em termos de estrutuuras de especificações, às quais são atribuídos valores, estruturas essas que serão definidas em função do tipo de informação que se considera relevante.

Considere-se, por exemplo, que se pretende construir uma base com informação lexical para um analisador e/ou anotador de corpora textuais[7], passível de ser utilizado como base de um corrector sintáctico bastante sofisticado. Em função da língua tratada será relevante definir especificações relativas à subcategorização nominal (`massivo/contável`, `abstracto/concreto` ou `comum/próprio`), que permitirão um tratamento mais efectivo das concordâncias entre os argumentos de uma estrutura sintáctica. Esse tipo de especificações poderão ser consideradas irrelevantes para programas menos sofisticados de tratamento de dados linguísticos.

Considere-se ainda que a base lexical é construída para suportar uma gramática específica. Torna-se evidente que terá que existir coerência entre as especificações definidas para o léxico e as especificações definidas para a gramática.

A situação ideal é a definição de uma estrutura de especificações que seja suficientemente flexível e pormenorizada para permitir, por um lado, uma conversão fácil dessas especificações para um

[7] Entende-se aqui 'texto' de forma ampla, incluindo, portanto, produções escritas e orais.

novo formato e que, por outro, permita a construção de extensões de especificações facilmente incorporáveis no estrutura global da base, de modo a prever a reutilização do repositório de informação lexical para novas aplicações (o problema da reutilizabilidade das bases (do termo inglês *reusability*) será retomado em 3.3.). Porém, a Lexicografia Computacional ainda não atingiu este objectivo, o que tem condicionado a promoção de iniciativas como a EAGLES, sumariamente apresentada em 3.5.

3.2. Metodologias a considerar para o carregamento das bases

O carregamento das bases é um dos problemas que mais tem ocupado a comunidade científica, sobretudo porque, se a definição da estrutura modelar é já um trabalho moroso e difícil, o carregamento da informação no modelo definido envolve grandes custos, quer em termos do tempo dispendido, quer em termos da mão-de-obra necessária, quer, em suma, em termos financeiros.

WILKS, Y. *et al.* (1993) distinguem basicamente dois tipos de metodologias:

- a *demo-approach*, que consiste na digitação manual de toda a informação contida na base;
- a *book approach*, que consiste na extracção automática ou semi-automática contida em produtos legíveis com máquina, corpora textuais ou DLMs, que segundo os autores, implica a sua transformação em DTMs (como já foi referido em 1.2.).

Parece evidente que a *book approach* se apresenta como a mais económica, sobretudo se se pretender construir uma base contendo um elevado número de unidades lexicais e não um mero repositório prototípico de informação lexical. À partida tal metodologia implica a existência dessas fontes para a língua tratada. Por outro lado, a mera existência de corpora digitais e de DLMs não é uma condição suficiente:

- no que respeita aos corpora, a sua utilização implica que eles se encontrem anotados com base em especificações que estejam de acordo com a informação que se pretende extrair deles;
- no que respeita aos DLMs, como é sabido e já foi, de resto, referido, o facto de serem concebidos para uso humano traduz-se numa estruturação muito mais 'fluida', na qual parte da informação necessária para o sistema de PLN permanece implícita e, portanto, não descodificável pelo sistema.

Além disso, mesmo se a *book approach* for possível, ela exige sempre uma verificação / rectificação da informação por parte do lexicógrafo.

Deste modo, no caso, por exemplo, da participação portuguesa no GENELEX, a única opção possível foi a *demo-approach*, isto é, toda a informação contida na base foi digitada manualmente.

3.3. A possibilidade de reutilização da informação armazenada

A possibilidade de reutilização da informação armazenada em bases digitais lexicais para novas aplicações é uma das principais metas estabelecidas para a Lexicografia Computacional na actualidade.

CALZOLARI, N. (1991) atribui dois significados básicos à noção de 'reutilizável' em termos de Lexicografia Computacional:

- **reutilizável_1**: explorar e reutilizar informação lexical implícita ou explicitamente existente em fontes lexicais pré-existentes (DLMs, bases de dados terminológicas, corpora textuais, etc.), como auxiliar na construção de grandes léxicos computacionais do tipo reutilizável_2;

- **reutilizável_2**: construir léxicos computacionais de tal forma que vários utilizadores (diferentes sistemas de PLN - definidos no quadro de diferentes teorias e para diferentes aplicações - , mas também utilizadores humanos, como lexicógrafos, linguistas, consulentes comuns) possam utilizar - com interfaces apropriadas - informação lexical relevante (cf. p. 188).

Esta preocupação prende-se, como é claro, com os elevados custos envolvidos na construção de qualquer descrição de informação lexical.

Para atingir a meta da 'reutilizabilidade', quatro estratégias se afiguram, na minha opinião, determinantes:

- ao nível dos produtos para uso humano, o tratamento progressivamente mais estruturado e coerente da informação lexical, traduzível, por exemplo, em vocabulários controlados para a elaboração das definições, no respeito sistemático pela estrutura frásica característica de cada tipo de definição, pelo refinamento e utilização sistemática das várias convenções utilizadas ao nível da micro-estrutura para distinguir acepções distintas, separar definições perifrásticas das correspondentes definições sinonímicas, etc.;
- ao nível dos formalismos, a escolha de um formalismo facilmente convertível em outros formatos;
- ao nível das especificações definidas, a definição de estruturas suficientemente detalhadas, mas ao mesmo tempo flexíveis, de modo a possibilitarem a selecção do ponto de vista mais indicado para a extracção da informação que se pretende;
- ao nível das instâncias responsáveis pelos projectos, o estabelecimento de normas (ou *standards*) amplamente aceites pela comunidade científica, de modo a serem seguidas pelos participantes nos vários projectos.

3.4. Formalismos a usar

Tal como já foi referido, a selecção do formalismo escolhido, ou, por outras palavras, da linguagem de representação da informação, tem fortes implicações a nível das possibilidades de reutilização da informação armazenada em bases digitais lexicais.

Assim, a escolha dos formalismos mais adequados foi alvo de estudo por parte das instâncias responsáveis ao nível da União Europeia. No Relatório Final do Estudo *Eurotra-7* (também conhecido por *ET-7*) (*apud*. MULTILEX Consortium (1992b: 18)) são mencionados dois tipos de formalismo, que têm vindo a ser progressivamente adoptados:

- a *typed feature logic*, que podemos brevemente definir como um tipo de linguagem na qual às estruturas de especificações (*feature structures*[8]) são atribuídos tipos (*types*) ordenados. Um tipo permite definir condições de validade para uma determinada classe de estruturas de especificações, implicando que as estruturas que se apresentarem como subtipos do tipo principal devem obedecer a essas condições. Este sistema de tipos pode ser usado para definir uma hierarquia de especificações, segundo a qual as estruturas de especificações situadas num nível inferior da hierarquia são automaticamente enriquecidas com informação derivada das condições de validade definidas em níveis superiores (cf. BRISCOE, T. (1991: 43));
- o *SGML* (*Standard Generalized Markup Language*), que é uma linguagem de estruturação de informação, não necessariamente linguística mas geral (como o próprio

[8]As *feature structures* são frequentemente usadas em sistemas baseados em teorias da gramática mono-estráticas e lexicais para tratar a categorização sintáctica.

nome indica), normalizada (Norma ISO 8879[9]), permitindo um mais fácil intercâmbio de informação armazenada em estruturas de formatos distintos, sem perdas significativas. Trata-se basicamente de uma linguagem de balizagem da informação contida em documentos, permitindo delimitar de forma clara, em cada um deles, os seus vários campos constituintes. Para tal, o SGML fornece uma sintaxe coerente e não-ambígua de representação de qualquer balizagem que o utilizador pretenda marcar num documento. A escolha da linguagem SGML, por parte do projecto GENELEX, prende-se, pois, com a necessidade de garantir uma mais fácil exportação da informação lexical nele descrita para bases de dados de tipos distintos.

3.5. A iniciativa EAGLES (vertente lexicográfica)[10]

No seio da União Europeia, no que se refere a produções em Lexicografia Computacional, a situação actual pode genericamente resumir-se nos seguintes tópicos:

- nove línguas nacionais de trabalho (com o próximo alargamento da União, mais línguas de trabalho passarão a ser consideradas);
- para cada língua, uma ou mais (muito) pequenas bases digitais lexicais;
- uma tradição lexicográfica distinta praticamente para cada língua;
- algumas (poucas) organizações (privadas ou públicas) a produzir na área da engenharia linguística;
- uma ausência quase total de emparelhamento de línguas.

A situação desejável seria aquela que apresentasse as características seguintes:

- a existência de grandes repositórios lexicais para as diversas línguas;
- a realização de grandes dicionários (com 50 000 ou mais entradas) contendo informação detalhada.

Ora, os elevados custos envolvidos na concretização destes projectos, já referidos anteriormente, obrigam a um grande esforço de normalização do trabalho, de modo a garantir a reutilizabilidade das bases, a partilha de informação, a economia dos meios e o mais rápido desenvolvimento desta disciplina. É neste contexto que é lançada, em Fevereiro de 1993, a iniciativa EAGLES, que, não sendo a primeira deste género, é a mais recente[11].

EAGLES é a sigla para *Expert Advisory Groups on Linguistic Engineering Standards* e constitui uma das iniciativas do *Linguistic Research and Engineering Program* (Programa de Investigação em Linguística e Engenharia), promovido pela Comissão Europeia.

O seu principal objectivo é, pois, abrir caminho para a construção de directivas, metodologias e ferramentas amplamente aceites, capazes de orientar os trabalhos em curso na área do processamento da linguagem natural e da fala.

O procedimento preconizado para esta normalização é, não o da normalização do facto consumado, mas o da normalização resultante da aproximação consensual entre as várias equipas europeias a trabalhar em PLN ou em projectos a ele ligados.

O trabalho do EAGLES baseia-se nos seguintes princípios:

---

[9]ISO: International Organization for Standardization.

[10]Agradeço a Ulrich HEID, editor interno do EAGLES, a gentil cedência dos acetatos por ele apresentados no *Workshop on Acquisition and Representation of Lexical Information* (cf. Bibliografia), que muito facilitaram a organização da informação relativa a esta iniciativa.

[11]Outras iniciativas com objectivos normalizadores são a *Text Encoding Initiative*, o projecto *ET-7* (*Standards for Reusable Lexical and Terminological Resources*) e o projecto *NERC* (*Network of European Reference Corpora*) (cf. CORREIA, M. & P. GUERREIRO (1993c)).

- o realismo, traduzindo-se na necessidade de aceitar o estabelecimento de diferentes graus de pormenorização para cada nível da descrição (por exemplo, uma descrição mais detalhada ao nível da morfossintaxe do que ao nível da semântica);
- tratamento igual para todas as línguas da União Europeia;
- a modularidade e a extensibilidade, isto é, aceita-se o princípio de que o EAGLES poderá agir a nível das especificações que são comuns a todas as línguas, mas que cada língua definirá extensões próprias, capazes de dar conta das suas particularidades;
- a tomada em consideração dos dados já existentes.

No quadro do EAGLES estão definidos cinco grupos de trabalho em função das cinco áreas abrangidas:

- léxicos computacionais;
- corpora textuais;
- formalismos para PLN;
- avaliação das ferramentas e componentes de PLN;
- linguagem oral.

Cada grupo encontra-se dividido em subgrupos vários, a fim de desempenhar tarefas específicas. O ILTEC participa no EAGLES, no quadro dos léxicos computacionais, designadamente nos subgrupos que trabalham a área da morfossintaxe e da semântica.

A primeira fase de trabalhos do EAGLES decorreu entre o começo desta iniciativa e Junho de 1994, tendo o primeiro grupo (aquele que agora nos interessa referir) desenvolvido o seu trabalho em cooperação com o grupo dos corpora. Esta cooperação entre os dois grupos de trabalho encontra a sua justificação no facto de apresentarem uma base conceptual / descritiva comum, que surge evidenciada nos seguintes factos:

- na prática, a anotação de corpora consiste na aplicação de léxicos morfossintacticamente descritos;
- os anotadores de dados textuais podem ser definidos como conjuntos de especificações, tal como os léxicos;
- os anotadores podem beneficiar da inclusão de informação adicional: distribucional e mesmo lexical.

O trabalho relativo aos léxicos desenvolveu-se em três vertentes:

- normalização da descrição das formas lexicais através de propriedades morfossintácticas (e de traços de concordância);
- estudo da sintaxe (em particular do fenómeno da subcategorização);
- estudo da(s) possível(eis) normalização(ões) dos dicionários multilingues.

A estas tarefas aliam-se duas actividades infra-estruturais, a desenvolver ao longo das duas primeiras fases de desenvolvimento do EAGLES:

- estabelecimento de uma metodologia para a normalização lexical;
- directivas relativas à arquitectura dos léxicos computacionais.

Ao nível da morfossintaxe, foram comparadas as descrições existentes nos vários projectos europeus, tendo-se chegado à formulação das seguintes recomendações:

- obrigatórias: classificação das partes do discurso (ex.: `N, ADJ, V`);
- recomendadas: definição de proriedades morfossintácticas relevantes para descrever os fenómenos de concordância (ex.: para os nomes, as especificações `comum / próprio, contável / massivo`)[12];

[12]É de notar que a gramática tradicional portuguesa não considera a subcategorização da categoria N em termos dos valores 'contável /massivo'.

- opcionais: definição de propriedades específicas de um determinado fenómeno, língua ou aplicação (ex.: para o português a segmentação do traço pessoa em dois: `personne_deixis` e `personne_accord` - cf. 4.3.2.).

Ao nível da arquitectura dos dicionários computacionais, o objectivo do EAGLES é, não a definição de uma única estrutura possível, mas sim a definição de princípios que permitam a definição de estruturas para dicionários, de modo a que estes se tornem multifuncionais.

A segunda fase de trabalhos do EAGLES decorrerá entre Julho deste ano e Junho de 1995 e no seu decurso está prevista a execução das seguintes tarefas:

- aprofundamento do estudo das especificações morfossintácticas e sua validação através da aplicação a dados concretos (extraídos, por exemplo, de corpora textuais);
- continuação do estudo das descrições sintácticas, bem como das aplicações multilingues;
- possível estudo da descrição semântica.

4. Alguns projectos de bases digitais lexicais na União Europeia - breve apresentação

Foram seleccionados como representativos de projectos a decorrer na União Europeia o MULTILEX, o ACQUILEX e o GENELEX. Dada a limitação de tempo, outros projectos, tão ou mais representativos, tiveram, portanto, que ser deixados de lado.

4.1. MULTILEX

O projecto MULTILEX (*A Multilingual Standardized Lexicon for the European Community Languages*) é um projecto integrado no Programa ESPRIT (5304).

O MULTILEX pretende construir uma BDL e, tal como o seu nome indica, centra a sua atenção no estabelecimento de um nível mínimo de normalização (*standardisation*), que possa permitir meios para obter, intercambiar, etc. aquela que é considerada a informação lexical básica. As propostas elaboradas por este projecto poderão ser usadas para orientar a criação de novos recursos lexicais, na medida em que se informarem futuros projectos de léxicos computacionais sobre o conjunto mínimo de descrições necessário para construir um léxico reutilizável, razoavelmente aceitável e englobante, utilizável em aplicações multilingues.

Assim, o MULTILEX deverá ser pensado, na perspectiva dos seus participantes, como um 'molde' no qual podem ser fundidas cópias de léxicos das várias línguas comunitárias previamente existentes. O resultado dessa fusão será conforme a vários conjuntos de parâmetros, assegurando que o objecto pode ser reconhecido por outros e ser usado de forma apropriada. Qualquer léxico moldado conforme os parâmetros do MULTILEX deverá ser considerado como um componente a ser carregado ou estruturado no seio de um léxico mais vasto, por englobar um maior número de unidades lexicais (ULs) e por englobar léxicos de várias línguas (cf. MULTILEX Consortium (1992a: 2)).

O MULTILEX pode então ser entendido como uma macro-estrutura onde caibam descrições monolingues do léxico das várias línguas comunitárias, associadas a sistemas de transferência que permitam que a cada unidade lexical X de uma dada língua sejam associadas tantas unidades lexicais de outras línguas quantas as traduções possíveis de X nessas línguas.

O modelo MULTILEX é aplicável às unidades do léxico corrente e às unidades terminológicas, partilhando ambas o mesmo tipo de descrição, sob a forma de ULs. O conteúdo semântico e pragmático que caracteriza as unidades terminológicas é tratado, ao nível da camada semântica, através de especificações distintivas, através de informação pragmática adicional e através da

utilização do sistema geral de referências cruzadas *(cross-references*) para estabelecer relações conceptuais (cf. MULTILEX Consortium (1992c: 3)).

A linguagem de representação usada no MULTILEX é uma variedade de *typed feature logic* (cf. 3.4.).

O MULTILEX apresenta uma estrutura quadripartida, sendo a informação estruturada nos seguintes níveis ou camadas:

- ortografia;
- morfologia;
- sintaxe;
- semântica.

A informação relativa à pragmática aparece disseminada ao longo dos vários níveis de análise. A informação fonológica, embora inicialmente prevista não foi trabalhada no MULTILEX, segundo é referido nos relatórios disponíveis.

Para além dos níveis anteriormente referidos, são projectados dois campos particulares:

- o campo das especificações de referências cruzadas, onde são definidas as relações semânticas entre ULs. Uma referência cruzada é definida por dois elementos: um elo típico (*typelink*) e uma lista de ULs, onde os valores dos tipos de elo se referem às relações de sinonímia, oposição, genericidade, etc. É sugerida a utilização de especificações deste tipo para dar conta das relações derivacionais;
- o campo das especificações de transferência (*transfer*), associadas directamente às ULs. Existirão tantas especificações de transferência associadas a uma UL de partida quantas as suas traduções possíveis nas línguas envolvidas. A transferência estabelece as modalidades de relação entre a UL da língua de partida e as ULs de chegada, mas deixa as descrições da unidade lexical de partida no dicionário monolingue.

Dado o tempo disponível não me permitir alongar-me na descrição do projecto, referirei apenas dois aspectos que me parecem particularmente interessantes:

I. Uma UL é definida como um item que permite identificar informação lexical relativa a um e apenas um significado. Por outras palavras, a homonímia é, no quadro deste modelo, levada às últimas consequências, abordagem que difere substancialmente da utilizada no quadro de projectos como o GENELEX (cf. 4.3.2.). Este tipo de abordagem é, no entanto, compreensível sobretudo considerando os princípios basilares do projecto: de facto, esta abordagem homonímica (a cada UL, um significado) é, como é sobejamente conhecido, a que mais se adequa à descrição das unidades terminológicas, sendo ao mesmo tempo aquela que poderá viabilizar o sistema de transferências, isto é, de localização de equivalentes interlinguísticos.

II. A estrutura de especificações prevista para a descrição das variações ortográficas das ULs deveria merecer uma atenção particular se se pensasse na construção conjunta de uma grande BDL que desse conta do léxico das diversas variedades do português (europeia, brasileira, angolana, etc.).

A cada forma ortográfica ficam associadas diversas especificações relativas a informação diatópica, diafásica e diastrática que permitem uma melhor localização / explicitação dos seus usos efectivos.

As especificações associadas a cada forma ortográfica podem ser observadas na figura 1. A figura 2 apresenta-nos o tratamento dado à UL de língua inglesa **nosy** (com as variantes: *nosy* / *nosey*)[13].

4.2. ACQUILEX

O projecto ACQUILEX (*The Acquisition of lexical knowledge for Natural Language Processing systems*) está integrado no Programa ESPRIT-BRA (3030) da União Europeia.

Tal como foi anteriormente referido, o produto resultante deste projecto será uma BCL (cf. 1.4.).

Os objectivos últimos do ACQUILEX são, por um lado, contribuir decisivamente para a construção de léxicos computacionais de grandes dimensões (uma das principais metas da Lexicografia Computacional na actualidade) e, por outro, conseguir uma representação na qual a informação semântica seja explícita e acessível.[14]

O ACQUILEX centra os seus esforços no desenvolvimento de técnicas e metodologias para utilizar e interpretar os DLMs existentes de modo a construir componentes para sistemas de PLN. Pretende-se extrair informação lexical - sintáctica e semântica de múltiplos DLMs, em contexto multilinguístico, com o objectivo de construir uma única BCL multilingue. As línguas presentemente envolvidas no projecto são o italiano, o inglês, o neerlandês e o espanhol. Com base na informação semântica contida nesses DLMs, pretende-se construir um modelo de especificações que permita fazer a descrição semântico-conceptual do conhecimento lexical.

A informação a extrair dos DLMs é, não só a informação que neles se encontra explícita (listas de palavras, categoria, etc.), mas sobretudo a informação que neles se encontra apenas implicitamente apresentada, não sendo directa e imediatamente acessível (ex.: taxinomias, estruturas argumentais, relações semânticas entre derivados, etc.).

No interior da BCL, será possível 'navegar' através do léxico, acedendo a ele através de conceitos ou através de relações semânticas diversos.

Assim, o ACQUILEX pretende formalizar o conhecimento básico geral contido nos DLMs, sob a forma de conceitos e de relações semânticas. O método utilizado para a definição dessa formalização é heurístico e basicamente indutivo, procedendo-se a generalizações progressivas a partir de elementos comuns.

Pretende-se, com este procedimento, estabelecer normas de representação semântica e do conhecimento do mundo, normas essas compatíveis, por um lado, com as representações existentes em dicionários e línguas diferentes (no ACQUILEX, até ao momento, cerca de 10 dicionários, envolvendo as quatro línguas mencionadas), e, por outro lado, com as abordagens semânticas da Lexicografia e da teoria linguística.

Uma das relações semânticas mais explorada no quadro do ACQUILEX tem sido a relação de hiponímia, por permitir construir taxinomias conceptuais. Para a extracção dessa informação taxinómica, foram construídos analisadores específicos, capazes de distinguir automaticamente, numa definição substancial o incluinte lógico (*genus*) e a(s) diferença(s) específica(s) (*diferentiae*) (cf. CALZOLARI, N. (1991)).

O mais interessante nas taxinomias, do ponto de vista do ACQUILEX, é, não os nós mais baixos da estrutura (*leaf nodes*), correspondentes a palavras específicas, mas sim os nós dos níveis médio e

[13]O GENELEX também permite a associação de informação pragmática a cada variante ortagráfica e/ou fonética, desde que não se escolha nenhuma delas como lema ou vedeta (do francês *vedette*, termo vulgarmente utilizado na bibliografia sobre lexicografia nesta língua) (cf. 4.3.2.).

[14]Esta focalização particular na descrição da semântica das ULs faz com que o modelo não apresente uma estrutura global organizada em níveis ou camadas, como no caso do MULTILEX e do GENELEX.

alto, dado que são eles que representam os conceitos lexicais mais básicos e, portanto, comuns às várias línguas.

A linguagem de representação utilizada no ACQUILEX é *a typed feature structure* (apresentada sumariamente em 3.4.).

Como foi já referido em 1.4., uma das características do ACQUILEX é precisamente o seu comprometimento explícito com uma teoria linguística. Assim, pretende-se utilizar a abordagem proposta por PUSTEJOVSKY, sob a forma de *'qualia structures'* (cf. PUSTEJOVSKY (s.d.) e (1993)).

Neste contexto, assume-se a hipótese de existirem tipos de significado (*meaning types*) comuns a todas línguas e usa-se a noção de estrutura de especificações como um mecanismo estruturador da informação semântica, alargando esta noção de modo a incluir e representar informação semântica não coberta pelos principais papéis definidos (*Constitutive, Formal, Telic, Agentive*) e informação enciclopédica mais genérica relativa aos conceitos.

Neste modelo, as taxinomias e as estruturas de especificações conceptuais constituem um ponto de convergência entre diferentes fontes e línguas, e entre as abordagens empíricas e as teóricas.

Nas figuras 3 e 4 apresentam-se as estruturas de especificações gerais, definidas com base na extracção da informação contida nas definições lexicográficas, para nomes de substância e para nomes de líquidos, respectivamente (note-se que, dado o conceito `substância` ser mais genérico que o conceito `líquido`, o número de especificações previstos para o primeiro ser maior, resultando as especificações definidas para `líquido` num subconjunto extraído do conjunto previsto para o conceito hierarquicamente superior).

4.3. GENELEX

O projecto GENELEX (acrónimo de *GENeric LEXicon*) é um projecto EUREKA e iniciou-se em França em 1990, tendo sido progressivamente alargado a outros países. A participação portuguesa teve o seu início em Março de 1992, tendo sido concluída em Junho de 1994 (o projecto está ainda em curso em França e em Itália).

Do consórcio GENELEX fazem parte, além do Instituto de Linguística Teórica e Computacional, o ILTEC (EU 524), por Portugal, os seguintes membros:

- por França, a GSI-Erly, o ASSTRIL-LADL, a SEMA-Group e a IBM-France;
- por Itália, o CONSORZIO LEXICON RICERCHE e a SERV.EDI e SOGESS sri.

O objectivo do projecto GENELEX é a criação de BDLs monolingues de várias línguas europeias, informatizadas segundo uma modelização comum. A informação a inserir nessas bases diz respeito, principalmente às unidades do léxico comum, não havendo, portanto, nenhuma focalização particular nas terminologias científicas e/ou técnicas.

O facto de essas BDLs serem construídas segundo uma modelização comum poderá permitir, no futuro, a sua utilização como base para sistemas de PLN quer monolingues, quer bi- ou plurilingues.

O grande objectivo do GENELEX é basicamente, portanto, a construção de uma estrutura modelar permitindo construir BDLs de várias línguas. Assim o modelo de especificações dos dados lexicais terá que ser não só amplo e flexível, como, ao mesmo tempo, rigoroso e coerente. Neste sentido, foi opção assumida pelo projecto a utilização do formato SGML.

A participação portuguesa neste projecto teve como principal objectivo a validação do modelo para a descrição do léxico da língua portuguesa, tendo a equipa do ILTEC decidido fazê-lo através do tratamento de um conjunto de cerca de 5000 unidades do vocabulário corrente, recolhido com base no *Vocabulário do Português Fundamental* e ainda no *Dicionário do Português Básico*, de Mário

Vilela. Esta validação implicou a proposta de algumas alterações ao modelo inicial, bem como o estabelecimento de especificações necessárias para a descrição da nossa língua (cf. 4.3.2. para alguns exemplos destas especificações).

Do ponto de vista da metodologia e das fontes usadas, toda a informação foi digitada manualmente em ficheiros, geralmente no formato *delimited text file*, dado a equipa portuguesa não ter tido acesso a fontes legíveis com máquina (corpora ou dicionários). A informação digitada teve por base a contida em diversas gramáticas, dicionários de língua, dicionários de verbos; recorreu-se também frequentemente à própria competência de falantes de português[15]. Os ficheiros assim construídos foram, num momento posterior, descarregados automaticamente na base de dados (construída em sistema Unyx), tendo sido necessário levar a cabo a verificação / rectificação da informação lexical.

A estrutura modelar do GENELEX é uma estrutura que permite uma descrição estática do léxico (para usar as palavras de Ted BRISCOE já referidas anteriormente); qualquer tipo de análise ou de geração (por exemplo, de formas flexionadas) construída com base no GENELEX pressupõe, portanto, a criação de programas específicos para o efeito.

Uma das primeiras aplicações em que está prevista a utilização desta BDL como repositório de informação lexical é o conjunto das Gramáticas GRAAL, projecto em que o ILTEC está envolvido, juntamente com os parceiros franceses do GENELEX.

O facto de as bases construídas de acordo com o modelo GENELEX se destinarem principalmente a constituir a base para sistemas de PLN, aliado ao facto de se pretender que sejam aplicáveis a várias línguas, exigiu um grande rigor ao nível da formalização da informação lexical.

A informação contida nas BDL GENELEX pode também ser utilizada por humanos, pressupondo, no entanto, a criação de interfaces específicas e de programas de conversão da informação necessária para cada produto em particular.

4.3.1. Estrutura global

A BDL GENELEX tem uma estrutura tripartida, comportando uma camada morfológica, uma camada sintáctica e uma camada semântica. Cada entrada ou lema corresponde a uma unidade morfológica (UM) e encontrar-se-á descrita a estes três níveis, excepto os afixos, que correspondem a unidades morfológicas particulares no GENELEX (UMAFF) [16], não recebendo descrição sintáctica. A informação atribuída em cada nível a uma entrada determinada encontra-se em conexão com todas as informações referentes a essa entrada nas restantes camadas do modelo, o que pode ser expresso pelo modelo entidade-relação simplificado apresentado na figura 5.

A camada morfológica permite a descrição das características morfológicas de cada lema, como sejam, genericamente, a sua categoria, flexão e estrutura interna. Ao nível desta camada foram também pré-definidos campos que comportam a informação etimológica. A informação pragmática é também contemplada no GENELEX, sobretudo ao nível da camada morfológica, onde é possível marcar variações dialectais, sociolectais ou temporais, bem como níveis de frequência de uso.

Na camada sintáctica são definidas as possíveis complementações de base dos lemas (estrutura máxima atribuível a uma UM, onde se regista a opcionalidade dos vários complementos), bem como

[15]O ILTEC estabeleceu contactos com editoras de dicionários portuguesas no sentido de lhes propor a participação no projecto. Porém, todos esses contactos foram infrutíferos.

[16]Dado o modelo ter sido primeiramente concebido em França e de modo a facilitar o intercâmbio de informação entre as várias equipas, foi decidido pela equipa portuguesa manter as etiquetas dos campos e das especificações em francês.

as possíveis complementações transformadas ou de superfície; cada uma destas estruturas corresponde a uma USYNT (unidade sintáctica), cujos índices foram atribuídos aos lemas constantes do dicionário de base. A cada UM pode corresponder uma ou mais USYNTs, consoante o número de estruturas de complementação que a UM apresenta.

A descrição do comportamento sintáctico das unidades compostas (entendidas sobretudo como sintagmas lexicalizados) é alvo de tratamento especial ao nível da camada sintáctica.

A camada semântica representa a informação por dois tipos de representação:

- uma representação 'decomposicional', através de traços de significação com valores positivos ou negativos, do tipo `animado, humano, contável`, etc.;
- uma representação 'relacional', que situa uma UL dentro do sistema lexical pelas relações que ela estabelece com as restantes. Essas relações podem ser de tipo predicativo (traduzidas, por exemplo, em termos de `casos (agente, paciente, destinatário`, etc.) ou de restrições de selecção) e de tipo semântico-paradigmático (como `sinonímia, oposição`, etc.).

A cada estrutura predicativa (USÉM - unidade semântica) corresponde um determinado índice. Uma USÉM pode estar directamente ligada a uma ou mais USYNTs, isto é, as USYNTs estabelecem a mediação entre as UMs e as USÉMs.

A correspondência entre unidades morfológicas, sintácticas e semânticas no seio do modelo pode ser representada através do esquema da figura 6.

Dado o tempo disponível, não é possível fazer uma descrição mais detalhada de todas as camadas do modelo. Assim sendo, optou-se por referir apenas alguns aspectos estruturais da camada morfológica, dado que, por um lado, ela é a mais 'pesada' do modelo e, por outro, é a partir da entidade UM que toda a informação que lhe é associada se desenvolve.

4.3.2. Camada morfológica

Como foi já referido, a camada morfológica do GENELEX é o nível mais 'pesado' da estrutura, contendo um maior número de especificações, bem como de relações entre elas. Este facto deve-se, sobretudo, à escolha básica do GENELEX que consiste em proceder à descrição lexical a partir da unidade lexical: é a partir da entidade UM que se estabelecem todas as relações dentro da estrutura modelar, o que, de resto, é notório a partir do esquema apresentado na figura 6.

A delimitação das unidades que correspondem a lemas no GENELEX tem por base os critérios *categoria morfossintáctica*, *combinatória de traços morfológicos* e *padrão de flexão* (critérios como a etimologia ou a estrutura interna não são geradores de homonímia, como acontece normalmente nos dicionários tradicionais[17]). Cada lema corresponde a uma UM (unidade morfológica). As UMs são caracterizadas em função da sua autonomia e da sua estrutura interna prevendo o modelo os seguintes tipos de unidades:

[17]Uma forma como **pena**, correspondente a dois homónimos nos dicionários gerais de língua, à luz do critério etimológico, no GENELEX é tratada como uma única UM, dado apresentar apenas uma categoria (N) e um padrão de flexão (**pena** (fem../sing.) / **penas** (fem./pl.)).
Por seu turno, uma forma como **falar**, correspondente a apenas um lema nos dicionários gerais de língua, à luz do mesmo critério etimológico, corresponde no GENELEX a duas UMs (uma, enquanto N e outra, enquanto V), a segunda delas dando origem a múltiplas USYNs.
Por fim, uma forma como **capital**, corresponde no GENELEX a três UMs: uma como adjectivo, outra como substantivo masculino e outra, ainda, como substantvo feminino.

- as simples (UM_S), que podem ser autónomas ou não autónomas (ex.: **cavalitas**, que apenas ocorre na locução **às cavalitas**); as palavras derivadas constituem um subtipo de unidades simples;
- as compostas (UM_C), autónomas, permitindo descrever compostos sintagmáticos e compostos por temas;
- as aglutinadas (UM_AGG), autónomas, permitindo descrever, por exemplo, as contracções de preposição e determinante, frequentes em português;
- as unidades morfológicas afixais (UM_AFF), obviamente não-autónomas.

A camada morfológica é bipartida, isto é, a cada UM é atribuída uma descrição em função da sua forma gráfica e outra em função da sua forma fonética, encontrando-se estes dois tipos de descrição em simetria estrutural. A cada UM pode estar associada uma ou mais UMGs (unidade morfológica gráfica) e uma ou mais UMPs (unidade morfológica fonética - do francês *unité morphologique phonétique*). Assim, à UM **ouro** estão associadas duas UMGs, correspondentes às variantes gráficas da palavra, e duas UMPs, correspondentes às suas variantes fonéticas, conforme se pode observar na figura 7.

No modelo GENELEX, as informações de carácter diatópico, diafásico e diastrático aparecem reunidas na entidade CombVE (*Combinaison de Valeurs d'Emploi* - Combinatória de valores de uso) que, por sua vez pode ficar directamente associada à entidade UM (equivalente ao lema), ou às entidades UMG ou UMP (equivalentes às várias formas gráficas e/ou fonéticas que a palavra pode assumir). Em casos como o de **bíceps / bicípete** (s. m. sing.), dado que aparentemente nenhum factor diafásico, diastrático, diatópico ou de frequência de uso distingue as duas formas, a descrição dos pares atributo / valor que constituem a CombVE é feita apenas uma vez, ficando esta associada directamente ao lema. Em casos como o de **brócolos / brocos** (s.m. pl.), dado que a primeira forma corresponde ao nível-padrão da língua, ao passo que a segunda forma se enquadra claramente num nível popular / coloquial do discurso, há, portanto necessidade de atribuir informação pragmática distinta a cada uma delas, ficando então cada CombVE associada à forma gráfica / fonética correspondente (cf. Consourtium GENELEX (1993a: 30-32; 84 e 113)).

No GENELEX português a questão da variação não foi exaustivamente tratada dado que o dicionário de base não apresentou casos em nosso entender problemáticos.[18]

As categorias morfossintácticas previstas pelo modelo, bem como as especificações respeitantes a subcategorias são as constantes da figura 8. A todas as UMs contidas no GENELEX são atribuídas, num primeiro momento, uma categoria e uma subcetgoria.

As UM_S são descritas em termos da sua flexão, e da sua estrutura interna no caso de se tratar de palavras derivadas, as quais serão referidas mais adiante.

As UM_C, tal como entendidas no GENELEX não correspondem à noção tradicional de 'palavra composta': do ponto de vista formal, trata-se de expressões complexas que o lexicógrafo decidiu registar ao nível da camada morfológica (cf. Consortium GENELEX (1993a: 23-25)). Porém, a uma UM_C é sempre atribuída uma categoria. Os componentes de uma UM_C devem ser descritos como UMs, quer sejam unidades simples, autónomas ou não-autónomas, afixos, unidades aglutinadas ou unidades compostas. As UM_C não possuem UMGs e UMPs, dado que as suas formas gráfica e fonética são dedutíveis a partir das formas dos seus componentes. A descrição das UM_C dá conta, então, em relação a cada um dos seus componentes, da ordem que ocupa no seio do composto, bem como do tipo de separador gráfico (hífen, espaço em branco) ou da sua ausência. Na descrição da flexão das UM_C são dadas as várias formas que cada componente assume em função da forma flexionada da UM_C.

[18]Os exemplos de variação citados anteriormente não pertencem sequer ao dicionário do GENELEX português.

A entidade UM_AGG permite registar fenómenos de contracção gráfica das unidades. Embora não lhe seja atribuída nenhuma categoria, o aglutinado é posto em relação com cada um dos seus componentes por meio de uma relação de composição, mas, ao contrário das UM_C, às UM_AGG são atribuídas UMGs e UMPs, dado que a forma do aglutinado não é dedutível a partir da forma dos seus componentes. O GENELEX permite marcar o carácter obrigatório ou facultativo da aglutinação (cf. Consortium GENELEX (1993a: 27)).

A descrição das UM_AFF corresponde a um esboço de descrição de regras de formação de palavras, dado que cada afixo é descrito em função da categoria de bases que selecciona, da categria de derivados que permite gerar, bem como dos padrões de flexão que podem estar associados aos derivados em função do afixo (no que diz respeito sobretudo aos sufixos; no que se refere aos prefixos, o campo reservado para indicação do padrão flexional foi deixado em branco). Na figura 13 encontram-se apresentados os vários tipos de informação que são associados a cada afixo, exemplificando-se a descrição com a atribuída ao sufixo português **-eiro.**

Finalmente, o GENELEX permite ainda dar conta de abreviaturas (ex.: **sr**. por **senhor**) e abreviações (ex.: **metro** por **metropolitano**), bem como de siglas e acrónimos. Para tal, no caso das abreviaturas e das abreviações, a forma abreviada é associada à entidade UM correspondente à forma não-abreviada da UL[19]. Quanto às siglas e acrónimos, estas formas são associadas à entidade UM correspondente à forma por extenso, que é, então, descrita como uma UM_C (cf. Consortium GENELEX (1993a: 26)).

Como foi já referido, a informação sobre as unidades lexicais é descrita em BDLs e BCLs em termos de pares de 'atributo ou especificação / valor'. Ao descrever alguns aspectos da informação lexical portuguesa, foi necessária a introdução de novas especificações tendentes a descrever fenómenos típicos da nossa língua, bem como de valores a atribuir a essas especificações.

Dado o português ser, como é sabido, uma língua com uma flexão muito rica[20], foi sobretudo ao nível flexional que se revelou necessária a introdução de um maior número de alterações de especificações às primeiras versões do GENELEX, bem como de novas especificações necessárias para descrever a nossa língua.

Assim, por exemplo, o modelo inicial não previa os valores casuais relativos à categoria pronome (que substituíram os valores franceses `personnel_fort` e `personnel_faible`, traços não-adequados à descrição dos nossos pronomes).

Também o valor `quantificateur` atribuído aos pronomes foi uma extensão portuguesa, justificável no facto de o valor `indéfini` inicialmente previsto não deixar claro o valor quantificativo que pronomes como **algum**, **muito** ou **alguém** conferem às palavras às quais aparecem associados.

Para descrever a flexão pronominal e a dos adjectivos possessivos, dado que a equipa portuguesa decidiu descrever as formas **você** e **vocês**, foi necessário o desdobramento do traço de flexão `personne` em dois novos traços: o traço `personne_deixis` (que define a pessoa do discurso) e o traço `personne_accord` (que define a pessoa da concordância). Para cada um destes traços foram definidos os valores `1 | 2 | 3`. Assim, por exemplo, os pronomes pessoais **eu**, **tu**, **ele** e **você** apresentarão, para os traços relativos a pessoa, os valores observáveis na figura 9. A `CombTM` (combinatória de traços morfológicos) permitindo descrever a flexão pronominal e a flexão dos adjectivos possessivos tem a configuração apresentada na figura 10.

[19]Note-se que o registo das abreviaturas convencionais não é prática corrente na lexicografia portuguesa. Porém, essa descrição é, em meu entender, necessária, sobretudo em dicionários de aprendizagem e/ou activos, dado que, se por um lado, as abreviaturas convencionalmente aceites não são previsíveis por regras, por outro, elas devem fazer parte do conhecimento mínimo exigido a um utilizador da língua.

[20]A confirmar esta afirmação pode referir-se que, enquanto no modelo definido para o francês, o número máximo de formas verbais simples definido foi 51, para o português, esse número atingiu as 91 formas.

Para descrever a flexão verbal portuguesa foi também necessário estabelecer algumas novas especificações no modelo, para além das referentes aos traços `personne_deixis` e `personne_accord`, nomeadamente o valor `GERONDIF` para a especificação `Mode`, os valores `PARFAIT` e `PLUS_QUE_PARFAIT` para a especificação `Temps`.

As formas flexionadas dos lemas podem ser geradas, no seio do GENELEX, a partir da forma equivalente ao lema, genericamente por dois processos:

- a definição de radicais combinatórios (`rad`s, gráficos ou fonéticos - `radg`s e `radp`s, respectivamente), aos quais se juntam as terminações específicas de cada uma das formas descritas;
- a utilização de um mecanismo de `retrait` (redução) e `ajout` (acrescento), que permite retirar à terminação da forma lematizada, acrescentando-lhe a terminação própria da forma flexionada.

O segundo mecanismo é o mais fácil de utilizar, além de evitar uma sobrecarga de informação ao nível da camada morfológica, resultante da definição de um ou mais `rad`s associados a cada unidade lexical (a forma do lema corresponde por defeito ao `radg0 / radp0`), que é determinado pela utilização do primeiro mecanismo. No entanto, este revelou-se o mais adequado para a descrição das formas verbais, sobretudo as dos verbos de padrão flexional mais irregular, como **ser** ou **ir**. Na figura 11, pode observar-se a descrição das formas flexionadas de quatro substantivos (a substantivos e adjectivos são atribuídos os mesmos códigos de flexão gráfica e fonética). Na figura 12, pode observar-se a descrição das formas de presente do indicativo do verbo **parar**.

Ao fazer a descrição da derivação no GENELEX português foi tomado em consideração o modelo derivacional de Danielle CORBIN e o estudo da derivação em português levado a cabo por Graça Maria RIO-TORTO (cf. Bibliografia). Tal facto foi possível pois o GENELEX, embora se assuma como neutral em relação a teorias do ponto de vista da estruturação da informação, pretende-se um modelo capaz de acolher diferentes teorias (*modèle 'théorie accueillant'*), do ponto de vista da descrição. Por outras palavras, na concepção da estrutura modelar GENELEX houve a preocupação de tornar o modelo capaz de suportar descrições de factos lexicais feitas no âmbito de teorias distintas (cf. Consortium GENELEX (1993b: 3-4))[21].

Do ponto de vista da derivação, o modelo revelou-se suficientemente eficiente para descrever os vários processos presentes na língua portuguesa, designadamente a prefixação, a sufixação, a circunfixação, a conversão e a derivação regressiva. Como exemplo da descrição conferida às palavras derivadas no GENELEX, pode observar-se a figura 14, na qual essa descrição é exemplificada com a da palavra **brincalhão**.

Porém, a descrição da conversão (inicialmente prevista para ser descrita ao nível da camada semântica), nos casos em que a base e o derivado apresentam o mesmo padrão de flexão (ex.: **dentista**$_{ADJ}$ / **dentista**$_{N}$), motivou a introdução dos valores relativos à categoria na descrição, quer da base, quer do derivado, único meio de o sistema conseguir distinguir um do outro, isto é, como meio para descrever a direcção da conversão.

Um problema foi insolúvel no quadro do GENELEX, tendo obrigado a um tipo de tratamento não satisfatório: o da descrição de palavras construídas apresentando sufixos avaliativos do tipo de **-zinho**.

[21]Por exemplo, na camada sintáctica, o modelo permite uma descrição de tipo atomicista (onde só os traços mais genéricos são atribuídos a uma UL, pressupondo que, para análises sintácticas, ao GENELEX terá que ser acoplada uma gramática), ou de tipo sintacticista (onde se prevê uma descrição pormenorizada de todas as estruturas frásicas, básicas ou transformadas, em que uma UL pode ocorrer e que é a perspectiva, por exemplo, dos léxico-gramáticas concebidos segundo o modelo do LADL).
A equipa portuguesa optou pela descrição dos factos sintácticos de tipo atomicista.

Três hipóteses de tratamento se apresentaram:

- tratar estas unidades como derivadas: no entanto, nesse caso seria impossível dar conta da flexão de unidades como **cãozinho** / **cãezinhos**, dado que o modelo apenas prevê que a flexão seja marcada na periferia dos derivados[22];
- tratar estas unidades como compostos: no entanto, dados os mecanismos propostos para a descrição da flexão dos compostos, se descrevêssemos o primeiro componente (**cão**) como invariável, o sistema geraria as formas **cãozinho** / ***cãozinhos**; se, por outro, descrevêssemos o primeiro componente como variável, as formas de **cãozinho** geradas seriam **cãozinho** / ***cãeszinhos.**
- tratar estas unidades como palavras simples: de resto, a única solução viável.

  Este meio permite dar conta das formas flexionadas por meio do mecanismo de `retrait / ajout`, descrevendo-se a forma de plural de uma palavra como **cãozinho**, resumidamente, do seguinte modo:

```
<MFGid="mfgn070" exemple="cãozinho,cãezinhos"
idr_CombTM="cgn1">#MS#<RETRAIT></RETRAIT><AJOUT></AJOUT>
idr_CombTM="cgn3">#MP#<RETRAIT>ozinho</RETRAIT><AJOUT>ezi
nhos</AJOUT>
```

Porém, esta solução contraria a verdade dos factos linguísticos, tendo sido, consequentemente, uma solução difícil de adoptar.

5. Conclusão

O objectivo da presente apresentação era dar a conhecer os desenvolvimentos que a Lexicografia Computacional tem conhecido na União Europeia.

Para tal, achei pertinente propor alguns equivalentes portugueses para termos usados em inglês, traduzindo frequentemente, porém, alguns conceitos mal delimitados. Neste contexto, em 1., propus os termos:

- dicionário legível com máquina (DLM);
- dicionário tratável por máquina (DTM);
- base de dados lexicais (BDL);
- base de conhecimento lexical (BCL);
- base digital lexical ou léxico computacional,

procurando contribuir para a definição dos conceitos correspondentes a cada um deles.

Antes de apresentar sumariamente alguns projectos de Lexicografia Computacional a decorrer na União Europeia neste momento (cf. 4.), julguei também oportuno referir as utilizações que destas produções se podem fazer (cf. 2.), bem como deixar algumas indicações sobre os temas gerais que mais ocupam os especialistas nesta área (cf. 3.).

Foi minha intenção sensibilizar para o tema, dando uma imagem o mais aproximada possível da situação actual na Europa da União.

[22]Esta restrição manifesta-se na associação dos diversos modos de flexão gráfica e fonética que ficam associados a cada afixo quando da sua descrição no GENELEX.

Dada a vastidão do tema, porém, não foi possível apresentar mais do que uma panorâmica muito genérica, e provavelmente empobrecedora, do estado da Lexicografia Computacional na Europa. Para obviar a este facto incluíram-se, pois, na bibliografia alguns dos títulos considerados mais significativos para uma introdução a esta disciplina, tendo sido marcados com asterisco.

Bibliografia

*ATKINS, B. T. S. & a. ZAMPOLLI (eds.) (1994), *Computational Approaches to the Lexicon*, Oxford, Oxford University Press.

*BOGURAEV, Bran & Ted BRISCOE (1989), "Introduction", in BOGURAEV, Bran & Ted BRISCOE (eds.), *Computational Lexicography for Natural Language Processing*, Londres e Nova Iorque, pp. 1-40.

BRANCO, António H. & Paula GUERREIRO (1994a), *Final Report on the Syntactic Specifications for Portuguese Computational Lexicons*, Lisboa, ILTEC (disponível).

BRANCO António H & Paula. GUERREIRO (1994b), *Final Report on the Semantic Specifications for Portuguese Computational Lexicons*, Lisboa, ILTEC (disponível).

*BRISCOE, Ted (1991), "Lexical Issues in Natural Language Processing", in KLEIN, E. & F. VELTMAN (eds.), *Natural Language and Speech*, Spinger-Verlag, pp. 39-68.

*CALZOLARI, Nicoletta (1988), "The dictionary and the thesaurus can be combined", in EVANS, Martha Walton (ed.), *Relational models of the lexicon*, Cambridge, Cambridge University Press, pp. 75-96.

CALZOLARI, Nicoletta (1990), "Structure and access in an automated lexicon and related issues", in *Linguistica Computazionale vol. VI - Computational Lexicology and Lexicography: Special Issue dedicated to Bernard Quemada*, Pisa, Giardini Editori e Stampatori, pp. 139-161.

*CALZOLARI, Nicoletta (1991), "Acquiring and representing semantic information in a Lexical Knowledge Base", in PUSTEJOVSKY, James & Sabine BERGLER (eds.), *Lexical Semantics and Knowledge Representation*, Association for Computational Linguistics, pp. 188-197.

*CALZOLARI, Nicoletta, Johan HAGMAN, Elisabeta MARINAI, Simonetta MONTEMAGNI, Antonietta SPANU & Antonio ZAMPOLLI (1993a), «Encoding Lexicographic Definitions as Typed Features Structures», in BECKMANN, Farnk & Gehrard HEYER (eds.), *Theorie und Praxis des Lexicons*, Berlim / Nova Iorque, Walter de Gruyter, pp. 274-315.

CALZOLARI, Nicoletta, John McNAUGHT & Tarina AYAZI (1993b), *EAGLES: First Progress Report (Draft)* (disponível).

CALZOLARI, Nicoletta & Ted BRISCOE (s.d.), «Acquisition of Lexical Knowledge from Machine-Readable Dicitonaries and Text Corpora», in VARILE, N. & A. ZAMPOLLI (eds.), *European Projects*, Pisa. [disponível na colectânea de textos de apoio do Curso *Léxicos Computacionais*, ministrado por Nicoletta Calzolary, no decorrer da *Fith European Summer School in Logic Language and Information*, Lisboa, Faculdade de Letras da Universidade de Lisboa, Agosto de 1993].

CAMEIRA, Célia, M. CORREIA & Paula GUERREIRO (1994), *Final Report on the Morphological Specifications for Portuguese Computational Lexica*, Lisboa, ILTEC (disponível).

Consortium GENELEX (1993a), *Couche morphologique*, Version 3.0, Paris (disponível).

Consortium GENELEX (1993b), *Couche syntaxique*, Version 4.0., Paris (disponível).

Consortium GENELEX (1993c), *Couche sémantique*, Version 1.0. Bêta, Paris (disponível).

CORBIN, Danielle (1987), *Morphologie dérivationnele et structuration du lexique*, 2 vols., Tubinga, Max Niemeyer Verlag.

CORBIN, Danielle (1991), «Introduction - La formation des mots: structures et interprétations», in *Lexique 10*, Villeneuve d'Ascq, Presses Universitaires de Lille, pp. 7-30.

CORREIA, Margarita (1993), "Léxico possível ou léxico atestado: dados para uma escolha de base no tratamento das palavras construídas", in *Actas do 1º Encontro de Processamento da Língua Portuguesa - Escrita e Falada*, Lisboa, INESC / UNINOVA / CLUL, pp. 107-112.

CORREIA, Margarita & Paula GUERREIRO (1993a), "GENELEX: um modelo para a construção de bases de dados lexicais do português", in *Actas do 1º Encontro de Processamento da Língua Portuguesa - Escrita e Falada*, Lisboa, INESC / UNINOVA / CLUL, pp. 147-150.

CORREIA, Margarita & Paula GUERREIRO (1993b), "Modèle morphologique du portugais", comunicação apresentada ao *Club Utilsateurs GENELEX*, Paris, IBM-France, Abril (disponível).

CORREIA, Margarita & Paula GUERREIRO (1993c), "Bases de dados lexicais", *in Actas do Seminário «Engenharia Linguísitca»*, Lisboa, Working Papers do ILTEC.

COSTA, J. Almeida & A. Sampaio e MELO (s.d.), *Dicionário da Língua Portuguesa*, 6ª ed., Porto, Porto Editora.

Dicionário Aurélio Eletrônico (baseado em FERREIRA, Aurélio Buarque da Holanda (1986), *Novo Dicionário da Língua Portuguesa*, 2ª ed. revista e ampliada) (1993), Rio de Janeiro, Editora Nova Fronteira.

EURALEX Newsletter (1994), «EC Programmes: Linguistic Research and Engineering 1», in *International Journal of Lexicography*, Volume 7, Number 3, Autumn 1994, pp. 1-3.

FERREIRA, Aurélio Buarque da Holanda (1986), *Novo Dicionário da Língua Portuguesa*, 2ª ed. revista e ampliada, Rio de Janeiro, Editora Nova Fronteira.

GUERREIRO, Paula (1993), *Report on the Morphosyntactic Specifications for the Portuguese Language (in the framework of the GENELEX PROJECT)*, Lisboa, ILTEC (disponível).

HEID, Ulrich (1994), «Towards guidelines for morphosyntactic description in lexicons and corpora - An overview of ongoing work in EAGLES», comunicação apresentada *no Workshop on Acquisition and Representation of lexical information*, Pisa, Julho de 1994 (acetatos gentilmente cedidos pelo autor).

*INTERNATIONAL JOURNAL OF LEXICOGRAPHY (1991), *Building a Lexicon*, Volume 4, Number 3, Autumn, Oxford, Orford University Press.

ISO (1986), Norma nº 8879: *Information processing - Text and Office Systems - Standard Generalized Markup Language (SGML),* 1ª ed.: 1986-10-15 (Ref. No. ISO 8879-1986(E)).

ISO (1988), *Amendment 1 to International Standard ISO: 8879-1986* (Ref. No. ISO 8879: 1986/AMENDMENT 1).

*MARTIN, Willy & Marc WOLTERING (1989), *Basic Issues in Computational Linguistics* (relatório redigido sob a responsabilidade de Bernard AL), Utreque, Van Dale Lexicografie.

MULTILEX Consortium (1992a*), Definition of the Standard: Multilingual description of Lexical items* (disponível).

MULTILEX Consortium (1992b*), Definition of the Standard: Multilingual description of Lexical items - final report* (disponível).

MULTILEX Consortium (1992c), *Definition of the Standard: Linguistic Architecture - final report* (disponível).

PUSTEJOVSKY, James (s.d.), «Linguistic Constraints on Type Coercion» [disponível na colectânea de textos de apoio do Curso *Léxicos Computacionais*, ministrado por Nicoletta Calzolary, no decorrer da *Fith European Summer School in Logic Language and Information*, Lisboa, Faculdade de Letras da Universidade de Lisboa, Agosto de 1993].

*PUSTEJOVSKY, James (1993), «Type Coercion and Lexical Selection», in PUSTEJOVSKY, James (ed.), *Semantics and the Lexicon,* Dordrecht / Boston / Londres, Kluwer Academic Publishers, pp. 73-93.

RIO-TORTO, Graça Maria (1993a), "Processamento derivacional em português", in *Actas do 1º Encontro de Processamento da Língua Portuguesa - Escrita e Falada*, Lisboa, INESC / UNINOVA / CLUL, pp. 89-92.

RIO-TORTO, Graça Maria (1993b)*, Formação de palavras em português: Aspectos da construção de avaliativos*, Dissertação de Doutoramento, Coimbra, Faculdade de Letras da Universidade de Coimbra. (inédito).

VILELA, Mário (1990), *Dicionário do português básico*, 1ª ed., Porto, ASA.

WILKS, Yorick, Dan FASS , Cheng-Ming GUO, James McDONALD, Tony PLATE & Brian SLATOR(1988), "Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing", in *Proceedings of COLLING'88*, Budapest, pp. 750-755.

*WILKS, Yorick, Dan FASS , Cheng-Ming GUO, James McDONALD, Tony PLATE & Brian SLATOR (1993), "Providing Machine Tractable Dictionary Tools", in PUSTEJOVSKY, James (ed.), *Semantics and the Lexicon,* Dordrecht/ Boston/ Londres, Kluwer Academic Publishers, pp. 341-401.

Feature	**Value**
lex	<string>
lex_type	full | abbreviation | acronym | truncation | symbol |elliptical_form
geo(graphy)	<linguistic geographical region(s)>
temp(oral)	contemporary | archaic | obsolete
social_group	<social gropus> | all
subject_field	<subject field labels> | general
style	neutral, formal, informal, standard, euphemistic, offensive, taboo, poetic

```
status                  preferred   |   progressive   |
                        accepted            |
                        unofficial | any

freq(uency)             common   |   less   common   |
                        uncommon |           rare

hyphenation             <hyphenation variants>
```

Figura 1: Estrutura de especificações que caracterizam a variação ortográfica no projecto MULTILEX (MULTILEX Consortium (1992a: 14)).

```
nosy_lu_1: [nosy_gpmu_1, nosy_gpmu_2]
                 syntax:    [...],
                 semantics: [def: quality of curiosity,
                 example: “a nosy person”, ...]
                 cross-ref: [...],
                 maintenance: [...],
                 transfer: [...].

nosy_gpum_1: ortography : [lex: “nosy”,
                          lex_type : full,
                          prag : <[geo: <usa>,
                                   temp: comtemporary,
                                   social_group: all,
                                   subject_field: general,
                                   style: neutral,
                                   status: any,
                                   freq: common,
                                   hyphenation: <no$sy>]
                                  [geo: <uk>,
                                   temp: comtemporary,
                                   social_group: all,
                                   subject_field: general,
                                   style: informal,
                                   status: any,
                                   freq: common,
                                   hyphenation:<no$sy>]
                 phonology : <“nowzi”>,
                 morphology : [ entrytype : lemma,
                                cat: syntax_cat,
                                boundness: free_form,
                                number_forms: invariant ]

nosy_gpum_2: ortography : [lex: “nosey”,
                           lex_type: full,
                           prag: <[geo: <usa>,
                                   temp: comtemporary,
                                   social_group: all,
                                   subject_field:general,
                                   style: neutral,
                                   status: any,
                                   freq: common,
                                   hyphenation:<no$sey>]
                                  [geo: <uk>,
                                   temp:comtemporary,
                                   social_group: all,
                                   subject_field: general,
                                   style: informal,
                                   status: any,
                                   freq: common,
                                   hyphenation: <no$sey>]
                  phonology: <“nowzi”>,
                  morphology: [ entrytype: lemma,
                                cat: syntax_cat,
                                boundness: free_form,
                                number_forms: invariant]
```

Figura 2: Representação da variação ortográfica, associada à informação pragmática, da unidade lexical **nosy**, no Projecto MULTILEX (em formato *structured type features* -simplificado) - extraído de MULTILEX Consortium (1992a: 21-22) .

SUBSTANCE

FUNCTION
USED_FOR (TELIC):
USED_IN:
USED_AS:
USED_BY:

PROPERTY
NATURE:
STRUCTURE:
ORIGIN:
STATE:

TASTE:
SMELL:
COLOUR:
SHAPE:
ASPECT:

LACKING:

SIMILAR_TO:

CONSTITUENCY
CONSTITUED_BY:
MAIN:
MUCH:

CONSTITUENT_OF:
MAIN:

SOURCE
DERIVED_FROM:
PRODUCED_BY:
PRODUCED_BY_MEANS OF:

LOCATION:
CAUSE_OF:
TYPIC_EXAMPLE:
NAME:

Figura 3: Exemplo de estrutura de especificações para nomes de SUBSTÂNCIA no Projecto ACQUILEX (extraído de CALZOLARI (1991)).

LIQUID

FUNCTION
USED_FOR (TELIC):
USED_IN:
USED_AS:

PROPERTY
SMELL:
COLOUR:

CONSTITUENCY
CONSTITUED_BY:
MAIN:
MUCH:
CONSTITUENT_OF:
MAIN:

SOURCE
DERIVED_FROM:
PRODUCED_BY:
PRODUCED_BY_MEANS OF:

LOCATION:
CAUSE_OF:

Figura 4: Exemplo de estrutura de especificações para nomes de LÍQUIDOS no Projecto ACQUILEX (extraído de CALZOLARI (1991)).

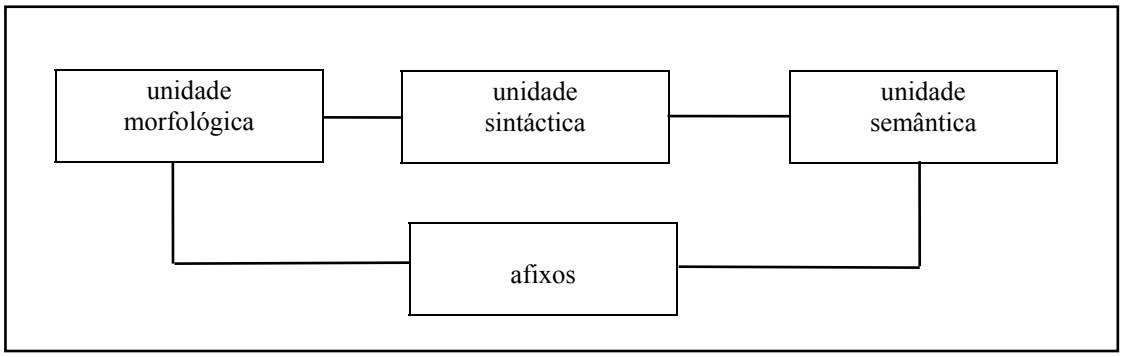

Figura 5: Esquema de entidade-relação simplificado, dando conta da interligação da informação relativa a cada unidade morfológica ao longo das três camadas do GENELEX.

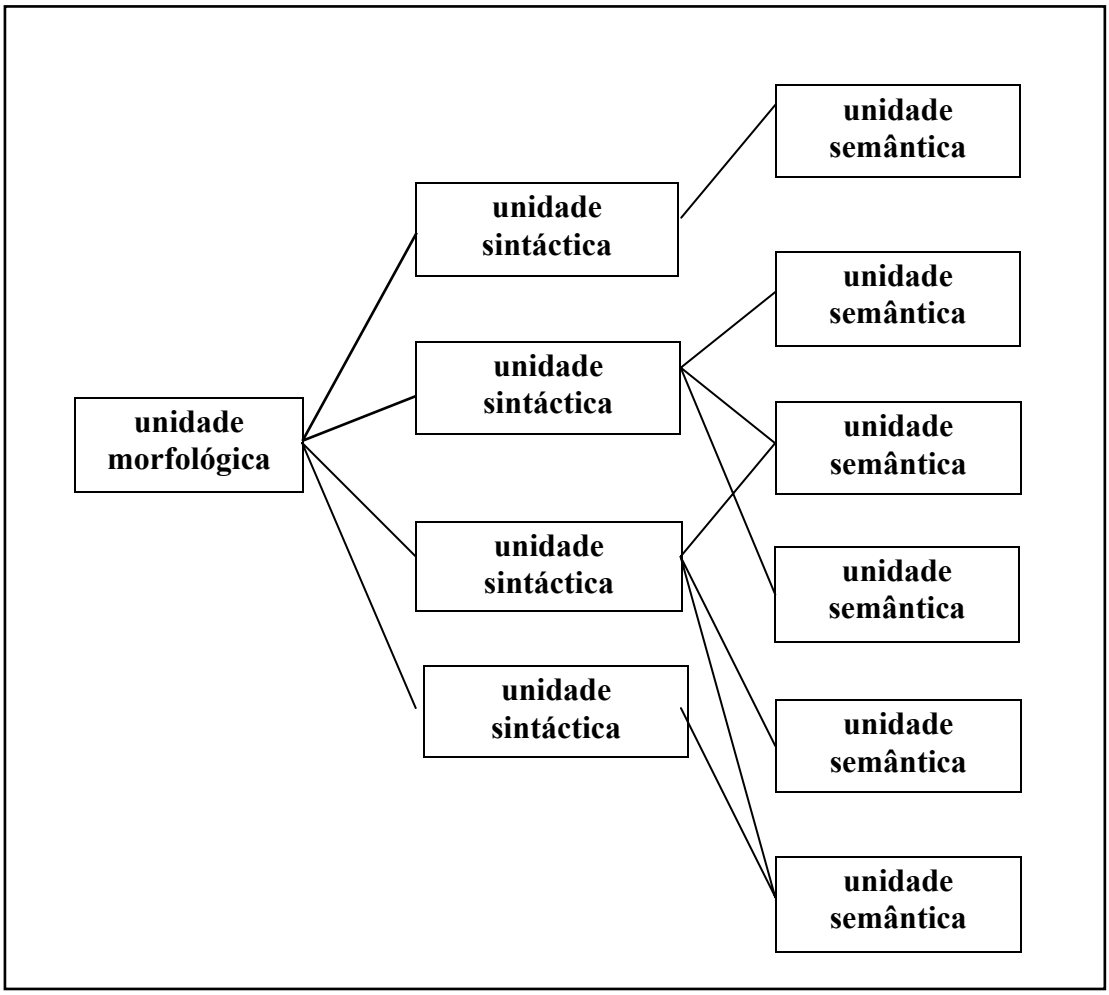

Figura 6: Esquema representando uma possível distribuição da informação relativa a uma UM por diversas USYNTS e USÉMS no Projecto GENELEX.

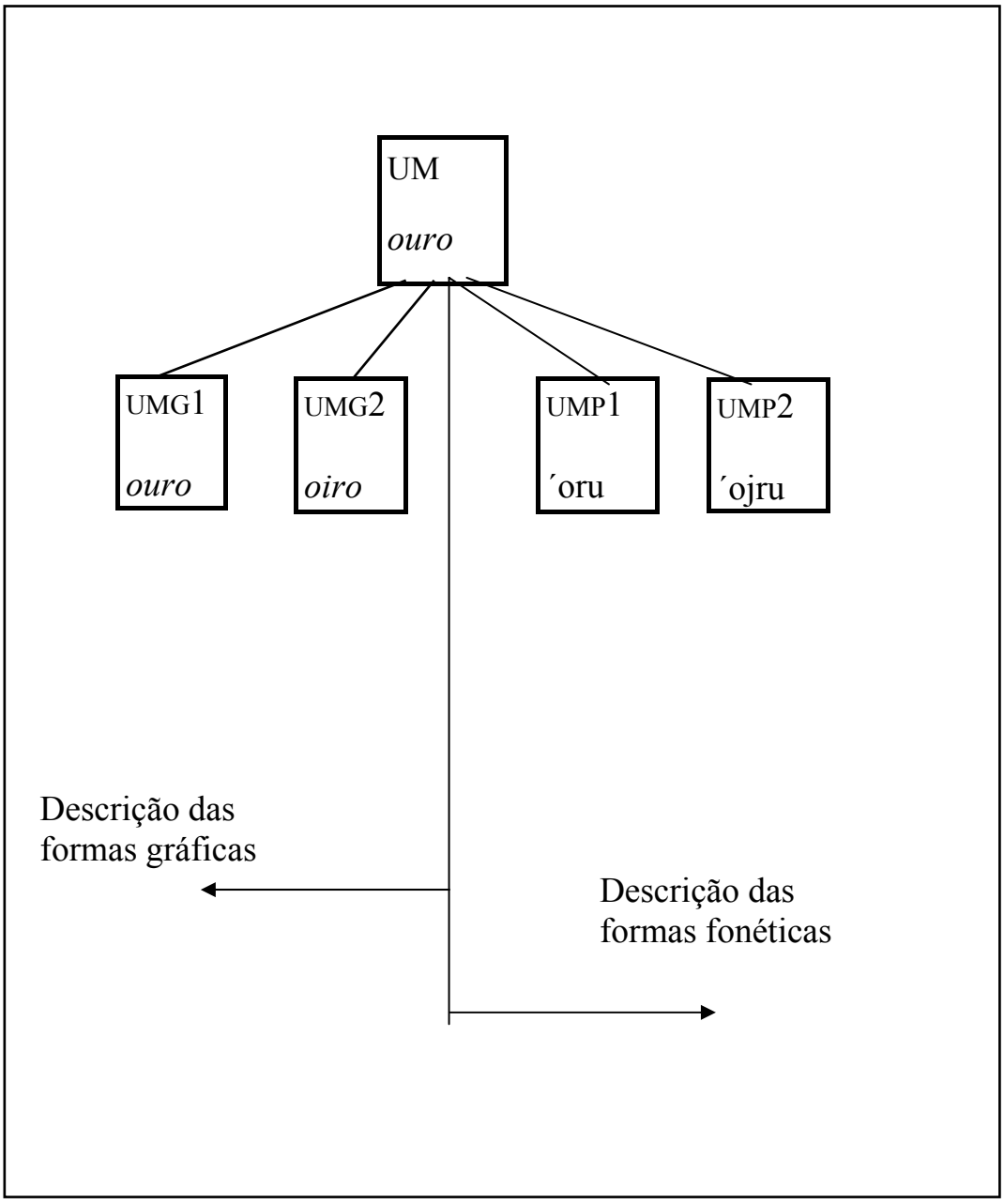

Figura 7: Esquema simplificado representando o tratamento atribuído a uma unidade (**ouro / oiro**) que apresenta duas variantes gráficas (MFGs) e duas fonéticas (MFPs) às quais foi atribuída uma mesma `CombVE`. É também visível a estrutura bipartida da camada morfológica, desenvolvendo-se em simetria estrutural, de modo a dar conta das características gráficas e fonéticas das UMs.

CatGram	**SCatGram**	**Exemplos**
NOM	COMMUN PROPRE	casa Abril
ADJECTIF	QUALIFICATIF DEMONSTRATIF POSSESSIF COMPARATIF_INF COMPARATIF_SUP SUPERLATIF CARDINAL ORDINAL	azul este seu pior melhor paupérrimo dois primeiro
ADVERBE	------ COMPARATIF_EG COMPARATIF_INF COMPARATIF_SUP	depressa tão menos mais
VERBE	------	negociar
PREPOSITION	------	para
CONJONCTION	COORDINATION SUBORDINATION COMPLETIF	e quando que
INTERJECTION	------	------
DETERMINANT	DEFINI INDEFINI QUANTIFICATEUR CARDINAL	o um nenhum um
PRONOM	QUANTIFICATEUR DEMONSTRATIF RELATIF INTERROGATIF PERSONNEL_NOMINATIF PERSONNEL_ACCUSATIF PERSONNEL_GENITIF PERSONNEL_DATIF PERSONNEL_OBLIQUE	alguém isso que qual ele o lhe lhe ele
PARTICULE	------	que

Figura 8: Valores de categoria gramatical (`CatGram`) e de subcategoria (`SCatGram`) atribuídos às UMs no GENELEX português.

eu	Personne_Deixis: 1 Personne_Accord: 1
tu	Personne_Deixis: 2 Personne_Accord: 2
ele	Personne_Deixis: 3 Personne_Accord: 3
você	Personne_Deixis: 2 Personne_Accord: 3

Figura 9: Valores atribuídos aos traços relativos a `pessoa`, usados na descrição dos pronomes pessoais e dos adjectivos possessivos, no GENELEX português (exemplo).

```
CombTM = Personne_Deixis
         Personne_Accord
         Genre
         Nombre
         Nombre_Poss
```

Figura 10: Combinatória de traços morfológicos usada para descrever a flexão de ADJECTIVOS POSSESSIVOS e de PRONOMES PESSOAIS, no Projecto GENELEX.

```
  <MFGid="mfgn001" exemple="aluno,aluna,alunos,alunos"
  idr_CombTM="cgn1">#MS#<RETRAIT></RETRAIT><AJOUT></AJO
  UT>
  idr_CombTM="cgn2">#FS#<RETRAIT>o</RETRAIT><AJOUT>a</A
  JOUT>
  idr_CombTM="cgn3">#MP#<RETRAIT></RETRAIT><AJOUT>s</AJ
  OUT>
  idr_CombTM="cgn1">#FP#<RETRAIT>o</RETRAIT><AJOUT>as</
  AJOUT>

<MFGid="mfgn013"exemple="aldeão,aldeã,aldeões(aldeãos,a
  ldeães),aldeãs"
  idr_CombTM="cgn1">#MS#<RETRAIT></RETRAIT><AJOUT></AJO
  UT>
  idr_CombTM="cgn2">#FS#<RETRAIT>o</RETRAIT><AJOUT></AJ
  OUT>
  idr_CombTM="cgn3">#MP#<RETRAIT>ão</RETRAIT><AJOUT>ões
  </AJOUT><RETRAIT></RETRAIT><AJOUT>s</AJOUT><RETRAIT>o
  s</RETRAIT><AJOUT>es</AJOUT>
  idr_CombTM="cgn4">#FP#<RETRAIT>o</RETRAIT><AJOUT>s</A
  JOUT>

  <MFGid="mfgn045" exemple="carácter,caracteres"
  idr_CombTM="cgn1">#MS#<RETRAIT></RETRAIT><AJOUT></AJOUT
  >
  idr_CombTM="cgn2">#MP#<RETRAIT>ácter</RETRAIT><AJOUT>
  acteres</AJOUT>

  <MFGid="mfgn068" exemple="óculos"
  idr_CombTM="cgn3">#MP#<RETRAIT></RETRAIT><AJOUT></AJO
  UT>
```

Figura 11: Descrição dos `MFGNs` (modos de flexão gráfica nominal - nomes e adjectivos) dos substantivos apresentando o mesmo padrão flexional que os substantivos **aluno**, **aldeão**, **carácter**, **óculos**, no modelo GENELEX (representada em formato SGML).

```
<MFG_id="mfgv108"=exemple="parar"

idr_CombTM="cmtppgn01">#IndPr11_S#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>o<AJOUT>
idr_CombTM="cmtppgn02">#IndPr22_S#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>as<AJOUT>
idr_CombTM="cmtppgn03">#IndPr23_S#
    nieme_Radg="1"
    <RETRAIT>ar</RETRAIT><AJOUT>ára<AJOUT>
idr_CombTM="cmtppgn04">#IndPr33_S#
    nieme_Radg="1"
    <RETRAIT>ar</RETRAIT><AJOUT>ára<AJOUT>
idr_CombTM="cmtppgn05">#IndPr11_P#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>amos<AJOUT>
idr_CombTM="cmtppgn06">#IndPr22_P#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>ais<AJOUT>
idr_CombTM="cmtppgn07">#IndPr23_P#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>am<AJOUT>
idr_CombTM="cmtppgn08">#IndPr33_P#
    nieme_Radg="1"
    <RETRAIT></RETRAIT><AJOUT>am<AJOUT>
```

Figura 12: Descrição do MFGV (modo de flexão gráfica verbal) das formas de presente do indicativo do verbo **parar**, no modelo GENELEX (representada em formato SGML).

Projecto GENELEX

Tipos de informação associados a cada Um_Afixo: exemplificação com o sufixo *-eiro*

Umaff: a forma representativa do afixo.
ex.: *-eiro*

Typaff: o tipo de afixo (prefixo ou sufixo).
ex.: *sufixo*

CatGram_Select: a categoria de palvras seleccionadas como bases de derivação.
ex.: *N* (*taberna* / *barba*) / *V* (*herdar*)

CatGram_Result: a categoria resultante do processo derivacional envolvendo o afixo.
ex.: *ADJ* (*taberneiro*)
N (*taberneiro / herdeiro* / *alheira*).

GenreN_Result: se o derivado for um substantivo, o seu género na forma lematizada.
ex.: *masc*. / *fem*.

umgaff: as várias formas que o afixo pode assumir nos vários derivados (alomorfes ou variantes condicionadas historicamente).
ex.: *-eiro* / *-eir-* (*tol**eir**ão*)

mfg: o modo de flexão gráfica, traduzido num índice, que os derivados apresentando dado afixo podem apresentar.
ex.: *-eiro,001/040; -eira,056; -eir-,∅*

Figura 13

Projecto GENELEX

Tipos de informação associados a cada Um_S (derivada): exemplificação com *brincalhão*$_{ADJ}$

Informações directamente relacionadas com o derivado:

Um: o lema a ser descrito.

ex.: *brincalhão*

mfg: o modo de flexão gráfica (expresso através de um índice.

ex.: *014*

Informações directamente relacionada com cada um dos seus componentes

ordre_linéaire: a posição que o componente ocupa na estrutura do derivado (1, 2 ou 3).

statut: o tipo de componente descrito.

ex.: *1 - base* ; *2 -sufixo* ; *3 - sufixo*.

um: o lema da base ou a forma canónica do afixo.

ex.: *brincar / -alho / -ão*

umg: a forma gráfica seleccionada no processo derivacional (geralmente igual à de um).

radg: o radical combinatório ou a forma truncada assumida pelo componente no processo derivacional.

ex.: *brinc / -alh- / -ão*

Figura 14