LE-PAROLE - Do corpus à modelização da informação lexical num sistema-multifunção

MARIA FERNANDA BACELAR DO NASCIMENTO, PALMIRA MARRAFA,
LUÍSA ALICE SANTOS PEREIRA, RICARDO DANIEL RIBEIRO,
RITA VELOSO e LUZIA WITTMANN

0 Introdução

Neste trabalho apresentam-se as grandes linhas do projecto LE-PAROLE - um projecto europeu de Engenharia da Linguagem (EL), no âmbito específico dos chamados Recursos Linguísticos -, incidindo-se, em particular, na investigação desenvolvida para o Português.

Dos objectivos gerais do projecto, fundamentos e contextualização no quadro dos projectos europeus de EL dá conta a secção **1**.

Nas secções **2** e **3** procede-se a uma caracterização das diferentes subáreas do projecto, com referência às especificidades e ao estado de desenvolvimento da componente relativa ao Português. Assim, a secção **2** respeita à constituição, distribuição e tratamento do *corpus*. Em **3** apresentam-se os critérios e procedimentos adoptados na constituição do léxico, o seu formato geral e os modelos de especificação da informação morfológica e da informação sintáctica que caracterizam as diferentes unidades lexicais.

Estando prevista a associação de informação semântica às descrições sintácticas definidas no LE-PAROLE, no quadro do projecto SIMPLE, procede-se ainda a uma breve caracterização deste outro projecto e evidencia-se a aplicabilidade dos resultados, na secção **4**.

Em **Conclusão**, sublinham-se as virtualidades do LE-PAROLE e a sua potenciação pelo trabalho a desenvolver no SIMPLE.

1 Enquadramento

O projecto LE-PAROLE é um projecto de Engenharia Linguística financiado pela Comissão Europeia - Direcção-Geral XIII, que se situa na continuação de outros projectos comunitários realizados com vista ao desenvolvimento, normalização e aplicação de recursos linguísticos europeus, particularmente de *corpora* e léxicos. Dada a existência do projecto *Corpus de Referência do Português Contemporâneo* (*CRPC*) no Centro de Linguística da Universidade de Lisboa, este Centro foi convidado a integrar as parcerias dos projectos que precederam o LE-PAROLE - o *Network of European Reference Corpora* (*NERC*), que teve como principal objectivo fazer recomendações à União Europeia sobre o futuro dos *corpora* linguísticos existentes na Europa a fim de se harmonizarem as metodologias de recolha, tratamento e análise dos materiais, e *Preparatory Action for Linguistic Resources Organization for Language Engineering* (*PP-PAROLE*), que promoveu e preparou o estabelecimento de infraestruturas para a criação, reutilização e harmonização de recursos na área dos *corpora* e da criação de ferramentas compatíveis, particularmente para a extracção de léxicos.

O LE-PAROLE dá, pois, continuação a estes dois projectos, integrado no Programa europeu *Telematics Application of Common Interest*. Nele participam instituições de 14 países europeus, estando a língua portuguesa representada pelo Centro de Linguística da Universidade de Lisboa (CLUL), como parceiro principal, e pelo Instituto Nacional de Engenharia de Sistemas e Computadores (INESC), como parceiro associado.

O projecto consiste no aproveitamento de recursos linguísticos e informáticos disponíveis nos países da União Europeia, tendo como objectivo a construção de léxicos de acordo com um modelo comum, de modo a facilitar as aplicações multilingues.

O léxico conterá níveis básicos de descrição linguística - ortografia, flexão, morfossintaxe, sintaxe e semântica lexical -, constituindo um modelo coerente e integrado, capaz de servir de base a um grande número de aplicações no âmbito do processamento automático das línguas naturais.

Actualmente o léxico contém informações de carácter morfológico e sintáctico; será ainda acrescentada informação semântica relevante, no âmbito do Projecto SIMPLE, a desenvolver na sequência do LE-PAROLE.

Estes projectos da União Europeia têm contribuído significativamente para a revisão e a actualização dos processos de trabalho dos vários parceiros envolvidos e para harmonizar práticas, aproximando aqueles cuja postura teórica de há muito vinha conduzindo, nos vários países, ao desenvolvimento de estudos linguísticos baseados em *corpora* orais e escritos. Da participação nestes projectos, resulta, pois, grande enriquecimento, quer em conhecimentos teóricos quer em aplicações práticas dos estudos sobre *corpora*. Para o *Corpus de Referência do Português Contemporâneo* (actualmente com 77 milhões de palavras) tornou-se, também, extremamente importante o estabelecimento da Rede Portuguesa de Fornecedores e Utilizadores de Dados, hoje constituída por mais de 30 instituições.

2 Corpus

2. 1 Definição

Tendo em conta os objectivos deste projecto, o *corpus* PAROLE é um *corpus* contemporâneo com características que permitem abranger o maior número possível de aplicações.

Inicialmente pensados para 50 milhões de palavras, os *corpora* das diferentes línguas acabaram por conter, cada um, 20 milhões de palavras (com excepção do *corpus* do Irlandês, que contém apenas 15 milhões).

Para alcançar os objectivos de harmonização no que se refere à composição dos diferentes *corpora* e para que estes possam funcionar como *corpora* comparados, os critérios definidos foram: meio, dimensão percentual por meio e ano de publicação.

Assim, no que se refere ao meio, o *corpus* total tem a seguinte composição:

Meio	Dimensão	%
Livro	4 milhões	20%
Jornal	13 milhões	65%
Periódico	1 milhão	5%
Miscelânea	2 milhões	10%
TOTAL	20 milhões	100%

Entende-se por **Livro** toda a publicação com número ISBN; em **Jornal** apenas foram incluídos diários; para **Periódico** foi seleccionada uma revista semanal; em **Miscelânea** encontram-se textos de publicidade ou divulgação.

No que se refere a aspectos cronológicos, o *corpus* é composto de materiais publicados a partir de 1970, não excedendo os 10% na década de 1970-80.

Aquando do estabelecimento dos critérios a ter em consideração no desenho do *corpus*, foi amplamente debatida a necessidade de serem tidos em conta aspectos temáticos identificados por "género" e "tópico". Constatou-se, no entanto, a impossibilidade de conseguir quer uma distinção objectiva entre estes dois parâmetros quer o consenso entre os parceiros quer, ainda, a obtenção de bons resultados a partir de uma análise externa dos materiais.

Na decisão final, optou-se por seguir apenas os critérios mais objectivos de dimensão, meio de divulgação e período de tempo, dado que, como afirma J. Sinclair (cf. Sinclair (1996: 43)), uma classificação temática coerente decorre da análise linguística interna dos textos, o que não seria possível no tempo de execução deste projecto.

As decisões relativamente aos textos a seleccionar para o *corpus* estiveram dependentes de autorizações por parte dos titulares dos Direitos de Autor.

O *corpus* português para o Projecto LE-PAROLE tem a seguinte distribuição:

Meio	Data	Nº Palavras (milhões)	TOTAL (milhões)
Livro	1970-1990	1	4
	1995-1997	3	
Jornal	1996-1998	3.75	13
	1996-1998	2.5	
	1996-1998	6.75	
Periódico	1996-1997	1	1
Miscelânea	1970-1990	0.5	2
	1996-1997	1.5	
TOTAL		20	20

No que se refere a Livro, foram incluídas obras de autores portugueses e obras traduzidas, tomando-se, no mínimo 30-40% de cada livro; relativamente aos jornais, escolheram-se as edições da Internet, por facilidade de conversão, tendo-se seleccionado 24 números por mês, durante 4 meses de 1996, 12 meses de 1997 e 2 meses de 1998, de diversos jornais. De periódicos, foram seleccionados 25 números de uma revista semanal, a partir de Agosto de 1996; na Miscelânea estão textos publicitários e informativos e entradas dos volumes de actualização de uma enciclopédia.

Todos os textos são convertidos em linguagem SGML - Standard Generalized Markup Language -, que permite um conjunto de marcações nos textos, aos níveis estrutural, tipográfico e linguístico. A aplicação do SGML é feita pelo CES - Corpus Encoding Standard, de acordo com as especificações do "TEI-Guidelines for Electronic Text Encoding an Interchage". O TEI, baseado no EAGLES - Expert Advisory Group on Language Engineering Standards, constitui uma iniciativa da União Europeia, e tem como objectivo a criação de padrões de:

- organização de recursos linguísticos em larga escala;
- manipulação desses recursos;
- acesso e avaliação de recursos, ferramentas e produtos.

O CES dá indicações sobre o nível de marcações que os *corpora* deverão atingir para serem considerados standardizados.

Um *subcorpus* de 3 milhões de palavras virá a estar acessível via Internet, sendo o restante apenas passível de consulta, mediante autorização específica do CLUL.

2. 2 Anotação

Dos *subcorpora* de 3 milhões de palavras que serão disponibilizados através da Internet, foram extraídos novos *subcorpora* com 250 mil palavras cada (um para cada língua representada) para anotação morfológica. Destas, 50 mil são desambiguadas com granularidade máxima. Toda a desambiguação é verificada manualmente.

O *subcorpus* anotado do Português reúne textos extraídos de jornais, livros, periódicos e miscelânea. A anotação é feita automaticamente através do analisador morfológico PALAVROSO (Medeiros (1995)), desenvolvido no INESC e adaptado para o modelo EAGLES/PAROLE. A desambiguação é semi-manual, tendo sido desenvolvida uma interface especial para o efeito, pelo INESC, denominada DINT-Desambiguador Interactivo (Pinto (1997)), que, além de garantir um nível de qualidade mais elevado, agiliza a execução.

Para harmonizar a informação morfossintáctica introduzida no *corpus*, foram tomadas como ponto de partida as recomendações do EAGLES (RE6100, CE). Partindo dessa base comum, os parceiros de cada língua efectuaram as devidas adaptações e acréscimos, conforme as especificidades de cada língua. A adaptação para o Português foi, portanto, feita pelos representantes do Português, CLUL e INESC (Bacelar, Bettencourt e Wittmann (1995)).

Para facilitar a consulta e uso genérico dos 12 *subcorpora* anotados, foi ainda estabelecido um conjunto de etiquetas (tagset) comum a todas as línguas representadas. Apresentamos, no quadro 1, a informação morfossintáctica anotada no *corpus* do Português: categorias, subcategorias e atributos.

Categorias	Subcategorias	Atributos
Nome	próprio comum	género e número
Verbo	principal auxiliar	modo; tempo; pessoa; género e número
Adjectivo		grau; género e número
Pronome	pessoal demonstrativo indefinido possessivo interrogativo relativo exclamativo reflexivo recíproco	pessoa; género; número; caso e formação
Artigo	definido índefinido	género e número
Advérbio		grau
Preposição		simples; contraída: género e número
Conjunção	coordenativa subordinativa	
Numeral	cardinal ordinal	género e número
Interjeição		
Único*	marcador da voz médio-passiva	
Residual	estrangeirismo abreviatura acrónimo símbolo	
Pontuação		

Quadro 1

Segue-se, a título ilustrativo, um excerto de texto anotado e desambiguado do *subcorpus* anotado do Português.

Nada (Pi=nn) como (Cs) um (Tims) passeio (Ncms) ao (S=fms) fim-de-semana (Ncms) . (O) De (S=s) referência (Ncfs) longe (R=p) de (S=s) casa (Ncfs) e (Cc) do (S=fms) trabalho (Ncms) . (O) Bem (R=p) pensado (V=p==sm) , (O) melhor (R=c) feito (V=p==sm) . (O) Nuno (Npms) Pereira (Npns) , (O) 24 (M) anos (Ncmp) , (O) engenheiro (Ncms) técnico-florestal (A=pnn) na (S=ffs) Guarda (Npfs) , (O) leu (V=is3s) algures (R=p) que (Cs) o (Tdms) Salamanca (Npns) , (O) clube (Ncms) da (S=ffs) 2ª (NULO) divisão (Ncfs) do (S=fms) país (Ncms) vizinho (A=pms) , (O) disputava (V=ii3s) , (O) a (S=s) 7 (M) de (S=s) Setembro (Npms) , (O) o (Tdms) seu (P03ms) primeiro (Moms) jogo (Ncms) em (S=s) casa (Ncfs), (O) contra (S=s) o (Tdms) Eibar (Npns) , (O) a (S=s) contar (V=n) para (S=s) o (Tdms) campeonato (Ncms) . (O)

Quadro 2

3 Léxico

3. 1 Selecção

O léxico de cada uma das línguas tem 20 000 entradas com a seguinte distribuição:

12 000 substantivos, 3 000 adjectivos, 3 000 verbos e 500 advérbios, sendo as restantes entradas constituídas por palavras gramaticais, siglas e abreviaturas.

Na selecção das entradas do Português seguiram-se critérios externos, como a extensão e os objectivos destes léxicos, e critérios internos, ou seja, critérios de natureza linguística, como os que dizem respeito às relações formais e semânticas entre itens lexicais.

Quanto às fontes, foram utilizadas fontes primárias (índices vocabulares extraídos de *corpora*) e fontes secundárias (vocabulários estabelecidos com base em dicionários).

Como se sabe, o estabelecimento de um léxico constitui tarefa tanto mais complexa quanto mais selectivo ele for. Neste caso, para além do escasso número de entradas lexicais, a subdivisão em classes de palavras, estabelecida nas especificações contratuais, não corresponde às percentagens observadas em léxicos gerais extraídos de *corpora* portugueses (cf., a este propósito, Baeclar do Nascimento *et al.* (1987:7)).

As primeiras unidades incluídas neste léxico foram as constantes do Vocabulário do Português Fundamental e algumas palavras temáticas do Inquérito de Disponibilidade que se situam imediatamente abaixo dos limiares de frequência estabelecidos para aquele projecto. Recorreu-se, em seguida, a uma lista de lemas teóricos (isto é, não desambiguados) extraída, através do PALAVROSO, analisador morfológico do INESC, de um *subcorpus* do CRPC de 5 milhões de palavras, essencialmente constituído por discurso

jornalístico. Desta lista, fez-se uma selecção com base em dados de frequência, tendo sido estabelecido estatisticamente um limiar adequado ao número final de entradas lexicais. A partir destas duas fontes, obteve-se uma lista de cerca de 19 000 palavras, que foi objecto de análise manual. Depois de classificados através do PALAVROSO, observou-se que havia lemas cuja frequência estava inflacionada por falta de desambiguação das suas formas constituintes e, ainda, que as percentagens por categorias gramaticais não correspondiam às requeridas pelo projecto.

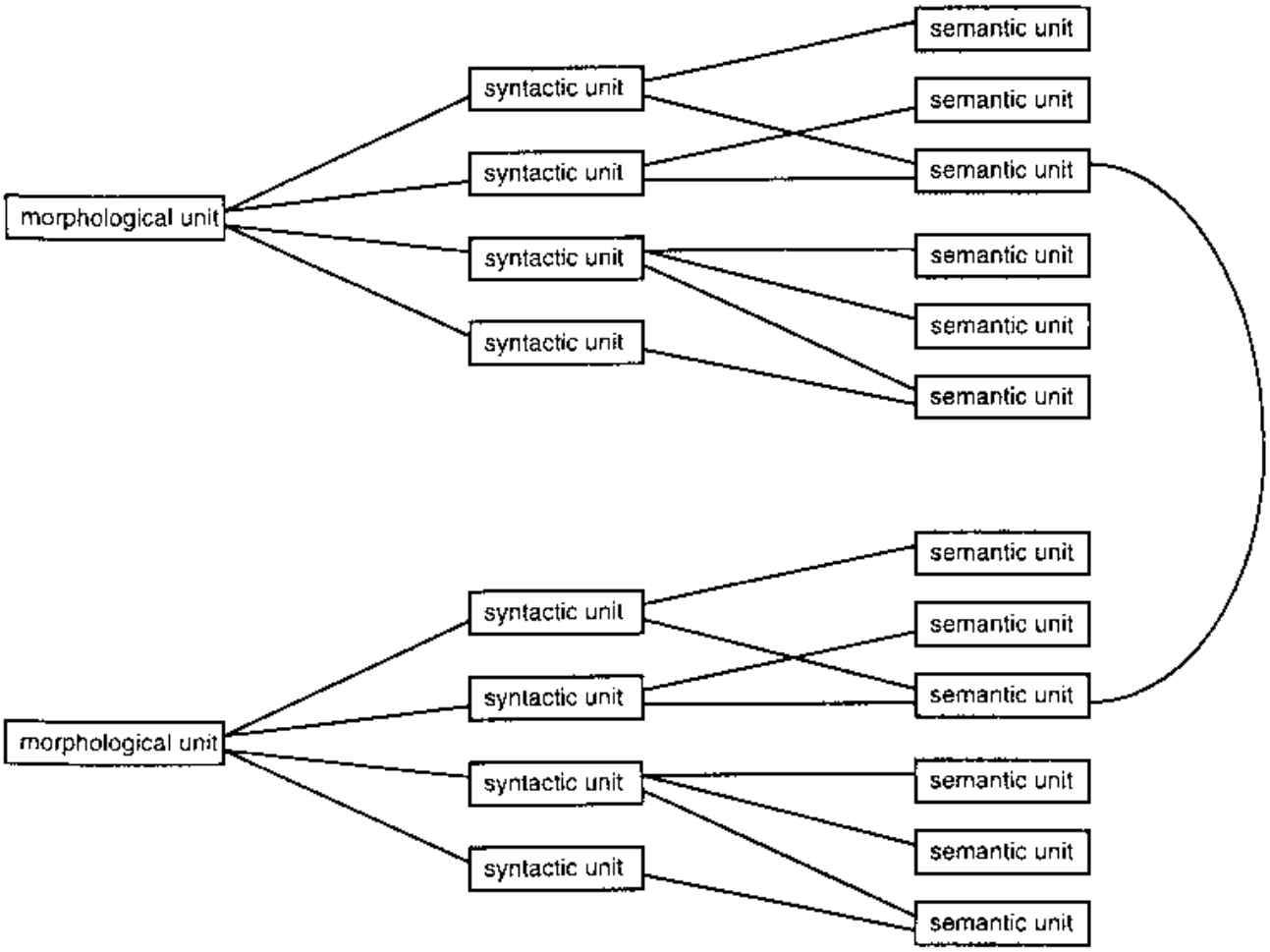
Recorreu-se, então, ao dicionário do PALAVROSO (que teve como fontes dicionários e *corpora* diversos), extraindo-se cerca de 2 000 nomes para substituir lemas anulados aquando da análise manual e adequar as categorias às percentagens previamente estabelecidas.

O léxico assim obtido foi, finalmente, objecto de mais duas análises: uma, de carácter quantitativo, consistiu na validação das entradas por confronto com os lemas teóricos extraídos automaticamente, mas, desta vez, de um *subcorpus* do CRPC de 12 milhões de palavras, o *corpus* do *Dicionário de Combinatórias do Português* que, quer pela dimensão que atinge quer pela sua actualidade e equilíbrio interno, considerámos adequado à aferição final de um léxico desta natureza. A análise seguinte teve carácter qualitativo. Observaram-se todas as unidades lexicais, no sentido de verificar a categorização proposta pelo analisador automático, uma vez que nem todas as classificações correspondiam ao nível de frequência adequado às finalidades e dimensão deste léxico.

Foram, igualmente, incluídas siglas e abreviaturas extraídas do *corpus* PAROLE e de fontes secundárias.

3. 2 Modelo, arquitectura e ferramenta de gestão lexical

Seguindo, genericamente, o modelo GENELEX, enriquecido com os resultados do EAGLES, a arquitectura do léxico PAROLE interliga, de forma integrada, unidades morfológicas, unidades sintácticas e unidades semânticas, como se evidencia no Quadro 3:



Quadro 3: Arquitectura do PAROLE

Adicionalmente, o modelo permite ligações multilingues.

Todos os parceiros utilizam a mesma ferramenta de gestão lexical, o AlethGD, desenvolvida pelo GSI-ERLI, de modo a garantir a uniformidade de todas as bases de dados. O AlethGD assenta sobre o sistema de gestão de dados orientada para objectos - "Object Store". Esta ferramenta dispõe das funcionalidades usuais e necessárias, incluindo importação e exportação de informação, para além da possibilidade de manipulação interactiva do modelo adoptado. Para garantir a abrangência e portabilidade dos dados, o modelo foi implementado em SGML, tendo sido definido um DTD – Document Type Definition – (que permite a importação e exportação de dados), comum a todas as línguas.

3. 3 Nível morfológico

3. 3. 1 Modelo de codificação

Tomando como base as recomendações EAGLES, os representantes de cada uma das línguas actualizaram os dados para a sua própria língua, acrescentando os traços específicos necessários. Uma vez harmonizada a nomenclatura, expressa em Inglês, foi definido um DTD comum, a ser utilizado na ferramenta comum (AlethGD).

Uma unidade morfológica corresponde, genericamente, à noção de lema, ou de entrada dos dicionários tradicionais. No léxico do Português, cada unidade morfológica é registada na sua forma gráfica e respectivas variações (como por exemplo *ouroloiro*), e codificada com a categoria, subcategoria (em alguns casos) e valores gramaticais, os paradigmas flexionais – incluindo o tratamento de clíticos –, e formas abreviadas.

Seguindo a política de reutilização de recursos anteriormente desenvolvidos pelos parceiros, a informação morfossintáctica para a codificação do léxico do Português foi extraída do analisador morfológico PALAVROSO, do INESC.

3. 3. 2 Conversão de classificações e regras do modelo PALAVROSO para o modelo LE-PAROLE

A conversão do modelo PALAVROSO para o modelo PAROLE exigiu o desenvolvimento de um gerador morfológico que classifica as entradas da selecção lexical, calculando em seguida o seu paradigma de flexão. Após terem sido calculados todos os paradigmas necessários, é gerada a informação no formato SGML, de acordo com o DTD PAROLE. Este DTD traduz o modelo PAROLE em SGML. Finalmente, utiliza-se um *filler* para a introdução do léxico e sua descrição morfossintáctica na ferramenta de gestão lexical, AlethGD.

Tomemos como exemplo a palavra **belo**. Inicialmente, esta palavra é identificada pelo PALAVROSO como um **adjectivo**, do género **masculino**, número **singular** e grau **positivo**. Em seguida, o módulo de cálculo dos paradigmas de flexão recebe esta informação e, com base nas regras de análise morfológica do PALAVROSO, gera como formas possíveis (paradigma de flexão) para esta palavra as seguintes:

Grau Positivo

Masculino, Singular	: belo
Feminino, Singular	: bela
Masculino, Plural	: belos
Feminino, Plural	: belas

Grau Superlativo

Masculino, Singular	: belíssimo
Feminino, Singular	: belíssima
Masculino, Plural	: belíssimos
Feminino, Plural	: belíssimas

A codificação consiste em representar esta informação de forma compacta, como se exemplifica:

(Singular):	M(,)F(o,a)	M(o,íssimo)F(o,íssima)
(Plural):	M(,s)F(o,as)	M(o,íssimos)F(o,íssimas)

Desta forma, a comparação dos paradigmas de flexão, com vista ao cálculo dos mais adequados e económicos para a cobertura de todas as entradas lexicais, é realizada com maior eficiência.

3. 4 Nível Sintáctico

3. 4. 1 Fundamentos

Como já foi referido, o projecto LE-PAROLE surge na sequência de outros projectos europeus de Engenharia da Linguagem. No que diz respeito à codificação da informação sintáctica, constitui, em grande medida, uma instanciação das recomendações do projecto EAGLES, para a categoria *verbo*, e do modelo geral do projecto GENELEX, para as restantes categorias.

Os modelos dos referidos projectos são teoricamente independentes, embora não teoricamente neutros, visto serem, de modo substancial, inspirados em modelos teóricos não derivacionais, com elevados níveis de declaratividade, decidibilidade e completude. Assim, é possível expressar, de forma transparente, descrições com diferentes motivações teóricas, susceptíveis de inclusão em sistemas integrados de processamento automático das línguas naturais e de utilização em aplicações práticas específicas.

3. 4. 2 Informação Codificada

Embora exista um relativo consenso quanto ao tipo de informação sintáctica a integrar em léxicos desta natureza, naturalmente que a codificação da informação encontra obstáculos nos limites do poder expressivo dos formalismos de implementação e das ferramentas de gestão. No caso concreto do AlethGD, a ferramenta utilizada no LE-PAROLE, como já referido, põem-se algumas dificuldades no que respeita à codificação de relações de alternância, a um nível de granularidade satisfatório. A despeito desta condicionante, a componente sintáctica do léxico PAROLE integra informação que permite uma caracterização fina das diferentes unidades lexicais. A saber: (i) restrições com reflexo morfológico; (ii) subcategorização; (iii) restrições sobre a estrutura interna dos argumentos; (iv) ordem linear; (v) função sintáctica dos argumentos; (vi) obrigatoriedade *vs* opcionalidade de realização dos argumentos; (vii) alternâncias (a um nível não exaustivo).

Restrições com reflexo morfológico. As propriedades léxico-conceptuais dos itens lexicais reflectem-se, como se sabe, tanto nos contextos sintáctico-semânticos que caracterizam a sua distribuição como em certos aspectos da sua morfologia.

Há, naturalmente, que exprimir apenas as idiossincrasias. É o caso, por exemplo, da restrição relativa à flexão de verbos que não atribuem qualquer função temática. Estes verbos (salvo no caso de empregos metafóricos) ocorrem apenas na 3ª pessoa do singular, uma vez que, como é sabido, o sujeito expletivo em Português, seja nulo ou lexicalmente realizado, é 3ª pessoa do singular (cf. *(Ele) chove a cântaros!/*(Eles) chovem a cântaros!*).

Este tipo de informação é codificada no objecto que respeita à unidade lexical, *stricto sensu*, o *Self*, como se ilustra na secção **3. 4. 3**.

Subcategorização. O conceito é aqui entendido no sentido alargado que os modelos lexicalistas lhe atribuem. Desta forma, é especificada a informação categorial relativa a todos os argumentos, incluindo o argumento externo. Pode-se, assim, distinguir, entre sujeitos nominais e sujeitos frásicos[1]. Adicionalmente, é possibilitada a expressão de restrições entre argumentos, bem como a associação de propriedades semânticas aos argumentos. Fica ainda facilitada a expressão da relação entre a distribuição sintáctica dos predicados e a sua estrutura conceptual, em aplicações em que tal se revele necessário.

Restrições sobre a estrutura interna dos argumentos. O modelo permite "entrar" nos diversos complementos (aqui incluído o sujeito, como atrás se referiu) e exprimir diferentes tipos de restrições: natureza dos complementadores (*que* + frase temporalizada; complementador *nulo* + infinitivo), modo, dependências referenciais obrigatórias (controlo ou simples co-referência), quantificação, entre outras.

Ordem linear. Os diferentes argumentos são associados a posições numa lista, pelo que a ordem por que aí ocorrem é relevante. Nesta fase, optou-se pela expressão da ordem básica por que os argumentos se realizam, na medida em que as alterações à ordem básica são sintáctica e/ou pragmaticamente motivadas. Contudo, e uma vez que se pretende que o sistema sirva aplicações diversas, independentemente de ser ou não associado a um módulo gramatical, ulteriormente será introduzida informação relativa a alterações sistemáticas, como acontece em combinatórias em que opera sempre *Heavy NP Shift*.

Função sintáctica dos argumentos. As funções consideradas são as funções canonicamente associadas às diferentes posições da lista de subcategorização.

No caso dos predicados complexos (como, por exemplo, *torna alegres*, em *A música torna as pessoas alegres*), não prevendo o modelo, nesta fase, a expressão de descontinuidades, optou-se por incluir o elemento não verbal do predicado na lista de subcategorização[2], associando-se-lhe a função de *predicativo*.

Obrigatoriedade *vs* opcionalidade de realização dos argumentos. Embora o modelo permita incluir informação relativa à possibilidade de não realização dos argumentos, optou-se por não exprimir essa possibilidade, na medida em que se assume a ideia comummente aceite de que a não realização dos argumentos é sintáctica ou pragmaticamente legitimada, não se tratando assim de uma propriedade idiossincrática dos predicados.

Nos casos em que um argumento interno pode ter ou não realização sintáctica (cf., por exemplo, *O João já comeu a sopa*/*O João já comeu*) o verbo surge associado a duas construções distintas.

Alternâncias. Representando as alternâncias um contributo importante para a determinação da relação entre o significado e a sua expressão sintáctica (a este propósito, ver, por exemplo, Levin (1993), para o Inglês, e Saint-Dizier e Marrafa (1995), para o Francês), o modelo LE-PAROLE prevê a associação dos diferentes contextos de realização dos argumentos (exs.: construção causativa/construção incoactiva: *O João ferveu a sopa*/*A sopa ferveu*; alternância locativa: *O João carregou as malas no carro*/*O João carregou o carro com as malas*). Contudo, contrariamente ao que é assumido em Levin (1993) e em alguns trabalhos subsequentes, não é aqui considerada qualquer relação derivacional entre esses contextos. Tal relação não encontra, aliás, motivação teórica nem empírica, nem é computacionalmente relevante, como demonstrado em Marrafa (1996).

O nível de granularidade da informação relativa às alternâncias não é ainda muito fino, em parte devido a algumas limitações do AlethGD, acima referidas, em parte por, nesta fase, não se incluir ainda informação semântica.

3. 4. 3 Objectos

O modelo conceptual do LE-PAROLE envolve um conjunto de objectos complexos, que são directamente traduzidos no AlethGD. Em termos gerais, o sistema inclui um objecto básico designado *Description*, em que é codificada toda a informação relativa à distribuição de cada unidade sintáctica, e vários subobjectos - *Self* (item lexical), *Construction* (construção), *Position* (posição), *Syntagma* (sintagma), *Frameset* (conjunto de descrições relacionadas) -, que permitem codificar a informação relativa às propriedades dos diferentes elementos que integram a *Description* e a relação entre diferentes *Description*. Assim, em *Self* codificam-se as restrições relativas às possibilidades de realização da unidade lexical, em *Construction* definem-se os contextos sintácticos, em *Position* codifica-se a informação respeitante à ordem e à função sintáctica dos argumentos e em *Syntagma* exprime-se o estatuto categorial e a estrutura interna dos constituintes. *Frameset* permite a expressão das alternâncias, quando as há, através da associação de diversas *Description*.

Toda a informação é codificada em pares *atributo: valor*, em concreto, sob a forma *featurename="α"/ value="Y"*.

O AlethGD permite o estabelecimento de ligações entre os diversos objectos, gerando, através de um *mapper*, um ficheiro com toda a informação organizada em diferentes níveis. Vejam-se, a título ilustrativo, os exemplos que se seguem[3]:

V117 afiançar VERB Usyn231 Usyn232

```
(232 )   description="DescriptionV47"
   example="O Rui afiançou ao Pedro que estava inocente"
   representativemu="afiançar"
   self="SelfV1"
         intervconst="IntervConst15"
               syntagmatl="Syntagme_T8"
                     syntlabel="V"
                     <SyntFeatureClosed
                           featurename="PASSIVIZABLE"
                           value="PASYES"

   construction="Syntagme_NT_C60"
         syntlabel="Clause"
         selfinsertion="1"
         <InstantiatedPositionC
               range="0"
               positionc="Position_CV0"
                     function="SUBJECT"
                     syntagmacl="Syntagme_NT_C3"
                           syntlabel="NP"
         <InstantiatedPositionC
               range="1"
               positionc="Position_CV30"
                     function="OBJECT"
                     syntagmacl="Syntagme_NT_C22"
                           syntlabel="Clause"
                           <SyntFeatureClosed
                                 featurename="SYNSUBCAT"
                                 value="THATCL"
                           <SyntFeatureClosed
                                 featurename="MOOD"
                                 value="INDICATIVE"
         <InstantiatedPositionC
               range="2"
               positionc="Position_CV1"
                     function="INDIRECTOBJECT"
                     syntagmacl="Syntagme_NT_C24"
                           syntlabel="PP"
                           featurel="Trait_Lex0"
                                 featurename="INTROD"
                                 value="a"
                                 mu="S1"
```

V160 ajeitar VERB Usyn414 Usyn415 Usyn416 Usyn417

---

(417) description="DescriptionV65"
example="A Ana ajeita-se a tratar de crianças"
representativemu="ajeitar"
self="SelfV9"
intervconst="IntervConst24"
syntagmatl="Syntagme_T12"
syntlabel="V"
<SyntFeatureClosed
featurename="NPRONOMINAL"
value="SE"
<SyntFeatureClosed
featurename="CONTROLT"
value="SUBJECTCONTROL"

construction="Syntagme_NT_C84"
syntlabel="Clause"
selfinsertion="1"
<InstantiatedPositionC
range="0"
positionc="Position_CV44"
function="SUBJECT"
syntagmacl="Syntagme_NT_C4"
syntlabel="NP"
<SyntFeatureClosed
featurename="COREF"
value="COI"
<InstantiatedPositionC
range="1"
positionc="Position_CV42"
function="PREPOBJ"
syntagmacl="Syntagme_NT_C85"
syntlabel="Clause"
featurel="Trait_Lex0"
featurename="INTROD"
value="a"
mu="S1"
<InstantiatedPositionC
range="0"
positionc="Position_CV43"
function="SUBJECT"

syntagmacl="Syntagme_NT_C457"
syntlabel="NP"
<SyntFeatureClosed

featurename="COREF"
value="COI"
<SyntFeatureOpen
featurename="NP"
value=" PRO"
<SyntFeatureClosed
featurename="SYNSUBCAT"
value="SSINFINITIVE"
<SyntFeatureClosed
featurename="MOOD"
value="INFINITIVE"

Como se observa, na primeira linha, identifica-se o item lexical, através de um código - *V117*, para *afiançar*, e *V160*, para *ajeitar* -, atribui-se-lhe um estatuto categorial - *VERB* -, e dá-se conta do número de contextos sintácticos que admite - para os quais remetem os códigos identificadores *Usyn* (cf. nota 3). Toda a restante informação é integrada no objecto básico *Description*. As restrições relativas às possibilidades de realização e a outras idiossincrasias da unidade lexical são codificadas no *Self*, o primeiro subobjecto da *Description*, e a informação relativa aos contextos sintácticos em questão, é expressa, de forma estruturada (por forma a dar conta das relações hierárquicas dos constituintes), nos outros subobjectos.

Assim, por exemplo, para *afiançar*, *Self* inclui informação relativa à possibilidade de ocorrência desta unidade em construções passivas - *featurename="PASSIVIZABLE"/ value="PASYES*. O traço em questão não é declarado para *ajeitar*, dado que, no caso do contexto seleccionado, tem valor negativo[4]. Do mesmo modo, o *Self* de *ajeitar*, (mas não o de *afiançar*) inclui o traço *featurename="NPRONOMINAL"/value="SE"* e o traço *featurename="CONTROLT"/ value="SUBJECTCONTROL*, que exprimem informação relativa ao facto de *ajeitar*, no contexto em causa, ser um verbo pronominal e um verbo de controlo de sujeito. Respeitando esta última propriedade a uma relação de dependência referencial entre o sintagma sujeito de *ajeitar* e o sintagma sujeito da frase encaixada que constitui o seu complemento, esta informação é também expressa nos subobjectos que lhes correspondem no interior de *Construction* - *syntagmacl="Syntagme_NT_C4"* e *syntagmacl="Syntagme_NT_C457"*, respectivamente -, através do traço *featurename="COREF"/value="COI"*, que é associado a ambos. A impossibilidade de realização fonética do sujeito da frase objecto é expressa através do traço *featurename="NP"/ value=" PRO"*. O sistema tem, pois, poder expressivo para exprimir restrições relativas à estrutura interna dos argumentos. Veja-se, ainda, a este propósito, a codificação das restrições impostas ao *modo* do complemento frásico. Em qualquer dos casos, a caracterização deste constituinte inclui um traço com o atributo *MOOD*, que, para *afiançar*,assume o valor *INDICATIVE* (cf. *featurename="MOOD"/value="INDICATIVE"*) e, para *ajeitar*, o valor *INFINITIVE* (cf. *featurename="MOOD"/ value="INFINITIVE"*).

A informação relativa à ordem e à função dos diferentes constituintes é directamente associada ao subobjecto *Position* e expressa através dos traços *range="n"* e *function="X"*.

Como fica evidente, o sistema é suficientemente expressivo para codificar toda a informação sintáctica relevante. Adicionalmente, é flexível o bastante para admitir a introdução de correcções e de informação nova, o que permite, sempre que necessário, o melhoramento da base de dados, quer em termos de actualização da informação expressa quer em termos incrementais.

4 O SIMPLE, no futuro do LE-PAROLE

Permitindo a arquitectura do LE-PAROLE a associação de um nível de informação semântica aos níveis que o sistema actualmente comporta (cf. secção 3. 2: Quadro 3), está já previsto o desenvolvimento desse nível, no âmbito do projecto SIMPLE, a iniciar-se no termo do LE-PAROLE.

Este projecto tem como objectivo, em termos gerais, a inclusão de informação relativa ao domínio informacional e ao "tipo" semântico das unidades lexicais, bem como às propriedades semânticas e às funções temáticas dos seus argumentos.

Desta forma, dispor-se-á de um instrumento susceptível de utilização num vasto leque de áreas do Processamento Automático das Línguas Naturais e da Engenharia da Linguagem, de que se destacam, para referir apenas alguns exemplos, sistemas de (i) etiquetagem, busca e aquisição automáticas da informação lexical; (ii) sumário inteligente de texto; (iii) tradução automática multilingue.

5 Conclusão

O LE-PAROLE representa, como se evidenciou, um importante incremento dos recursos linguísticos europeus, em geral, e, tendo em conta a componente portuguesa, dos recursos linguísticos portugueses, em particular.

A inclusão dos três níveis de descrição lexical num modelo de arquitectura unificada - que o SIMPLE vem proporcionar - potenciará, por certo, as virtualidades do sistema, amplificando o seu impacto científico e tecnológico.

Agradecimentos

Florbela Barreto, Fernando Batista, José Bettencourt, Maria João Ferro e Amália Mendes, da equipa do LE-PAROLE, contribuíram para a elaboração deste trabalho com comentários e sugestões úteis. Aqui fica o agradecimento.

NOTAS:

* Classificação única para o Português

1. Sendo o modelo teoricamente independente, os argumentos frásicos que evidenciem propriedades de sujeito são tidos como tal, independentemente de, em quadros derivacionais, poderem ser interpretados como objectos em estrutura profunda.

2. À semelhança, aliás, do que acontece em análises em que não é assumida uma relação isomórfica entre grelha de subcategorização e grelha temática (cf., a este propósito, Marrafa (1997)).

3. Tanto *afiançar* como *ajeitar* admitem mais do que uma construção sintáctica, como evidenciam os códigos identificadores das unidades sintácticas (que correspondem a diferentes comportamentos sintácticos) que lhes estão associados: Usyn231/Usyn232 e Usyn414/Usyn415/Usyn416/Usyn417, respectivamente. Contudo, não se tendo aqui como objectivo dar conta, de forma exaustiva, da informação associada a cada item, mas, antes, ilustrar o modo como o sistema organiza a informação, apenas se apresenta, em ambos os casos, uma das construções possíveis.

4. A não declaração dos traços para valores negativos é um factor de preservação da economia do sistema.

BIBLIOGRAFIA:

BACELAR DO NASCIMENTO, M. F., P. RIVENC e M. L. SEGURA DA CRUZ (1987), *Português Fundamental*, vol. II, *Métodos e Documentos*, tomo 2, Inquérito de Disponibilidade, Lisboa, INIC-CLUL.

BACELAR DO NASCIMENTO, M. F., J. BETTENCOURT e L. WITTMANN (1995), *Lexicon Morphosyntactic Specifications. Portuguese instantiation of the EAGLES*, MLAP PAROLE report. CLUL/INESC.

LEVIN, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, Chicago, the University of Chicago Press.

MARRAFA, P. (1996), Representação das Formas Predicativas Verbais do Português numa Base de Conhecimento Lexical, *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado - XIII Simpósio Brasileiro de Inteligência Artificial (SBIA' 96)*, Curitiba, Centro Federal de Educação Tecnológica do Paraná.

MARRAFA, P. (1997), Representação e Computação da Estrutura Conceptual das Construções Resultativas: Uma Abordagem Lexicalista, in A.. Brito, *et al.* (orgs.), *Sentido Que a Vida Faz - Estudos para Óscar Lopes*, Porto, Campo das Letras.

MEDEIROS, J. C. (1995), *Processamento Morfológico e Correcção Ortográfica do Português*, dissertação de Mestrado, Lisboa, IST.

PINTO, E. (1997), *Relatório de Estágio*, Lisboa, UBI/INESC.

SAINT-DIZIER, P. e P. MARRAFA (1996), Predicate-argument syntactic realizations in French, in N. Ide (ed.), *Research in Computing for the Humanities*, Vol. 6, Oxford, Oxford University Press.

SANTOS, D., J. C. MEDEIROS e L. WITTMANN (1995), *Portuguese Lexicon: Inflectional Morphology. On the compatibility of the Portuguese Lexicon of Palavroso and the Genelex model at the morphological level*, MLAP PAROLE report, Lisboa.

SINCLAIR, J. (1996), «Tipologia Textual EAGLES», *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, 1995 (Vol. I) – *Corpora*, Bacelar do Nascimento, M. F., M. C. Rodrigues e J. B. Gonçalves (orgs.), APL, Lisboa, pp. 39-91.