Recursos Linguísticos e Processamento Morfológico do Português: o PALAVROSO e o projecto LE-PAROLE

Luzia Helena Wittmann
Grupo de Linguagem Natural do INESC
Instituto de Engenharia de Sistemas e Computadores
Apartado 13069
1000 LISBOA
luzia.wittmann@inesc.pt

Ricardo Daniel Ribeiro
Grupo de Linguagem Natural do INESC
Instituto de Engenharia de Sistemas e Computadores
Apartado 13069
1000 LISBOA
ricardo.ribeiro@inesc.pt

Resumo

Neste artigo descrevemos o dicionário e o funcionamento do analisador morfológico do Palavroso para a língua portuguesa, e o processo de reutilização dos recursos linguísticos nele contidos para a construção de um léxico para o Português, no âmbito do Projecto LE-PAROLE, ou seja, integrado na política da União Europeia para o desenvolvimento da Engenharia da Linguagem.

Introdução

Neste trabalho apresenta-se em traços genéricos o sistema de processamento morfológico do Português, Palavroso – desenvolvido pelo Grupo de Linguagem Natural do INESC –, e a reutilização da informação morfológica nele contida para a codificação morfossintáctica do léxico de Português do Projecto europeu LE-PAROLE.

No item 1 são focados os objectivos gerais e o enquadramento do Projecto LE-PAROLE no quadro da política europeia para a engenharia da linguagem.

A descrição do Palavroso e das várias ferramentas que o compõem, com especial incidência sobre o analisador morfológico, constituem os itens 2 e 3.

Descreve-se no item 4 o modelo de codificação e armazenamento do nível morfológico do Léxico PAROLE, e, no item 5, a conversão da informação morfológica do Palavroso (abordagem por regras) para o modelo Parole (abordagem paradigmática).

No item 6 faz-se uma breve referência aos outros aspectos do LE-PAROLE, enquanto no item 7 referem-se trabalhos em curso com o objectivo de dar continuidade ao desenvolvimento desses recursos.

1. O projecto LE-PAROLE

A criação de recursos linguísticos básicos reutilizáveis é actualmente uma das apostas da CE no campo da Engenharia da Linguagem. Dado o elevado custo deste trabalho, por sua morosidade e exigência de recursos humanos especializados, a criação de bases de dados harmonizadas constitui um incentivo para o desenvolvimento e comercialização de aplicações pela indústria da linguagem. Além de reforçar a competitividade industrial, projectos como este vêm de encontro à política da União Europeia no sentido de harmonizar o desenvolvimento dos diferentes estados membros, sobretudo em se tratando do perigo de exclusão da sociedade da informação.

O projecto LE-PAROLE obteve financiamento pelo programa europeu *Telematics Application of Common Interest.* Conta com a participação de instituições de 15 países europeus, que garantem o desenvolvimento de léxicos e *corpora* para 12 línguas europeias. O Português está representado pelo Centro de Linguística da Universidade de Lisboa (CLUL) e pelo Grupo de Linguagem Natural do Instituto de Engenharia de Sistemas e Computadores (INESC), como parceiro principal e como parceiro associado, respectivamente.

Cada um dos doze léxicos contém 20 mil palavras, codificadas no nível morfossintáctico e sintáctico, seguindo um modelo comum, e utilizando a mesma ferramenta de gestão lexical – o AlethGD – desenvolvido pelo GSI-ERLI S.A.

O modelo prevê, ainda, a codificação semântica das entradas. Esta fase deu origem a um novo projecto, denominado SIMPLE, com os mesmos parceiros, iniciado a 01.04.98.

2. O sistema Palavroso

O Palavroso é um sistema de processamento morfológico do Português, concebido de forma a cobrir a língua escrita com a maior abrangência possível.

Os objectivos genéricos que orientaram a sua concepção compreendem tanto uma utilização directa, sobretudo na investigação linguística em geral, quanto indirecta, ou seja, sua integração como elemento de outras aplicações da área da Engenharia da Linguagem.

O módulo de análise morfológica, por exemplo, é um elemento fundamental para o desenvolvimento de qualquer aplicação que inclua um módulo de processamento de linguagem natural: gramáticas computacionais, ferramentas de tradução automática, análise de texto – especialmente em *text retrieval* e resumo automático. O gerador morfológico pode ser uma base para um programa de aprendizagem do Português para estrangeiros ou de apoio ao ensino nas escolas.

Em torno do analisador morfológico, descrito mais pormenorizadamente no item 3, reúne-se um conjunto de ferramentas eficientes, rápidas e práticas de tratamento de *corpus*, essenciais para a investigação linguística que procura fundamentar-se em *corpora* de pequena, larga, ou mesmo, muito larga dimensão. O quadro 1 apresenta um breve resumo das funções dessas ferramentas.

Compila	Ferramenta usada para a criação de dicionários
Léxico	Angariador de léxico. Usa o Palavroso para fazer análise morfológica em corpus, registando todas as palavras ou lemas que não existem no dicionário utilizado.
Encontra	Consiste numa ferramenta de pesquisa de informação em corpora (não necessariamente anotados). Usa o Palavroso para encontrar palavras ou sequências de palavras, com um determinado padrão morfológico.
Estatic	Permite fazer vários tipos de contagens em corpora, que envolvam características morfológicas.
Correcto	Corrector ortográfico baseado no Palavroso. Funciona em MS WinWord (versões 2.0, 6.0 e 7.0).
Gerando	Gerador morfológico. É capaz de gerar todas as formas de cada verbo, nome, adjectivo e advérbio presentes no dicionário.
Desambig	Interface para desambiguação semi-automática de textos anotados de forma não interactiva.

Quadro 1

3. Analisador Morfológico – Morfolog

O analisador morfológico Morfolog constitui o núcleo do Palavroso. A sua abordagem baseia-se fundamentalmente em regras – por oposição à abordagem paradigmática –, e permite que lhe seja acoplado um léxico de dimensão variável.

O léxico *standard* do Palavroso contém cerca de 65 mil entradas de Português Europeu, regras de análise morfológica e informação concernente à correcção ortográfica.

O sistema contém ainda um léxico de Português Brasileiro, resultante da adaptação do léxico de Português Europeu e introdução de palavras específicas da variante brasileira.

O Morfolog pode fornecer toda e qualquer informação morfológica para o Português, destacando-se o processamento de palavras compostas, de clíticos e do grau dos adjectivos através de regras linguísticas.

É possível utilizá-lo tanto de modo interactivo (quadros 2 e 3), quanto como anotador morfológico automático de texto (quadros 4 e 5), ou seja, inserindo ficheiros com texto como *input* e obtendo o mesmo texto anotado como *output*.

Análise interactiva:

```
canto
canto: nome_S_M v_PI_1S
  canto, nome, masculino, singular,
100%
  cantar, verbo, 1ª pessoa, singular,
Presente Ind., 100%
```

Quadro 2

Onde a palavra introduzida foi *canto*, a primeira resposta é dada de forma abreviada e contendo todas as classificações possíveis, a segunda corresponde ao lema canto e a terceira ao lema cantar. O percentual de 100% informa que a palavra encontra-se no dicionário em uso, e é uma palavra de uso frequente.

Análise interactiva:

```
far-se-lhos-ão
far-se-lhos-ão: v_I_3S -cl -cl
 farão-se-lhe*-o*
  fazer, verbo, 3. pessoa, plural,
Futuro Ind., 100%
  se, clítico, 3. pessoa, número
indeterminado, apassivante/reflexivo,
100%
  lhe, clítico, 3. pessoa, número
indeterminado, objecto indirecto, 100%
  o, clítico, masculino, número
indeterminado, objecto directo, 100%
```

Quadro 3

Onde se tenta evidenciar o tratamento dos clíticos, simples e contraídos.

Para uma etiquetagem não interactiva, o Morfolog pode ser provido com qualquer conjunto de etiquetas, conforme as necessidades do utilizador. O excerto do quadro 3, foi anotado com o conjunto de etiquetas (tagset) definido no âmbito do projecto LE-PAROLE.

Análise não interactiva – etiquetagem:

```
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 4

Note-se que este conjunto de etiquetas apresenta uma granularidade maior do que o conjunto *standard* do Palavroso. No caso dos clíticos, por exemplo, inclui os lemas do verbo e do(s) clítico(s), para evitar que, após a desambiguação, um item mantenha mais do que uma etiqueta.

Convém esclarecer que o conjunto de etiquetas PAROLE foi utilizado para anotar *corpora* de 12 línguas europeias, tendo sido baseado nas recomendações de harmonização do projecto EAGLES (cf. item 3.1).

Pode-se, ainda, obter do analisador apenas os lemas possíveis para cada palavra, conforme o exemplo do quadro 4.

```
No (no) fundo (fundo fundar fundir)
trata-se (tratar se) de (de) pôr (pôr)
ao (ao) serviço (serviço) do (do)
município (município) um (um)
verdadeiro (verdadeiro) " (") Big (big)
Brother (brother) " (") . (.)
```

Quadro 5

3.1. Arquitectura

O analisador morfológico do sistema Palavroso tem como características fundamentais:

1. O processamento baseado em regras.
2. A separação bem definida que existe entre as regras de análise morfológica e o léxico.

Deste modo, o analisador pode dar sempre uma resposta, ainda que o seu léxico não esteja suficientemente preenchido.

Existem duas fases no processamento de cada palavra submetida a análise morfológica. Primeiro, são aplicadas à palavra várias regras morfológicas que produzem análises hipotéticas. Na segunda fase, estas análises são verificadas no dicionário. Dependendo do modo de operação do analisador, obter-se-ão todas ou apenas as correctas.

Observe-se o exemplo da palavra *batas*. Na primeira fase do processamento, ou seja, após a aplicação das regras morfológicas obtiveram-se cinco hipóteses de análise. Na segunda fase, foram confirmadas pelo dicionário apenas duas hipóteses de análise.

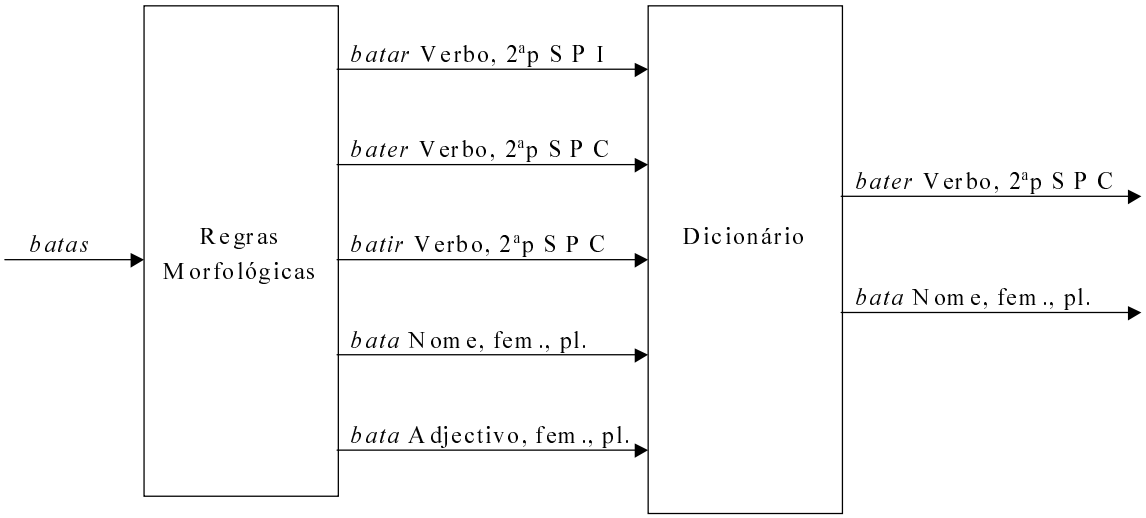

Figura 1 – Fases do processamento do analisador morfológico

Tanto as regras morfológicas, como o dicionário foram organizados em quatro classes, escolhidas em função da categoria gramatical. Essas classes são as seguintes: classes fechadas, advérbios, nomes e adjectivos, e verbos.

"Conceptualmente, os quatro tipos de regras são aplicados em paralelo, com acessos independentes ao dicionário (fig. 2). Na práctica, trata-se de processamento sequencial e, como será visto mais adiante, existem algumas opções de ordenação que tornam o sistema mais eficiente." (Medeiro, 95)

Outro dos aspectos relevantes em termos de arquitectura é a comunicação com o exterior. O núcleo do analisador morfológico é o módulo de análise, denominado Palavroso, não autónomo que pode ser usado por outras aplicações (fig. 2). Assim, pode haver a necessidade de realizar operações de pré e pós-processamento, ambas da responsabilidade da aplicação que usa os serviços do Palavroso. No entanto, de modo a facilitar a integração do módulo de análise morfológica noutras aplicações, foram já criadas algumas rotinas.

Uma das operações de pré-processamento mais usuais é a da normalização das palavras. Esta operação consiste em dois processos: passagem das letras maiúsculas a minúsculas e conversão dos caracteres acentuados para o formato usado pelo Palavroso. Outra operação que se pode considerar de pré-processamento é a escolha do modo de operação do analisador. Esta escolha reflectir-se-á no seu modo de funcionamento e, consequentemente, no tipo de resultados por ele produzidos.

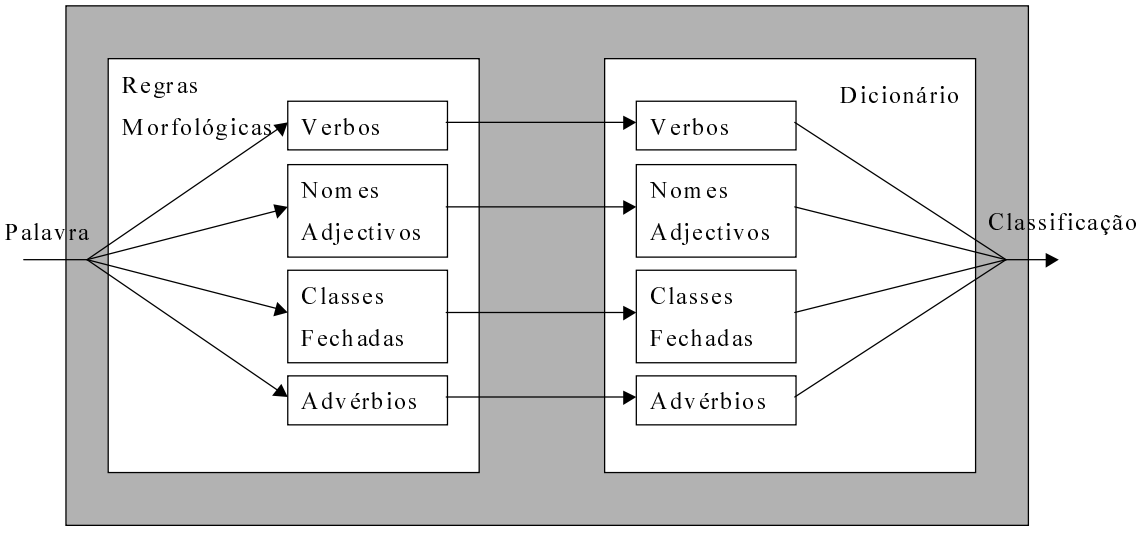

Figura 2 – Divisão das regras e do dicionário do analisador morfológico

Quando usado para anotação automática de textos, o resultado do analisador é uma etiqueta (ou conjunto de etiquetas) que descreve cada palavra analisada. Foi construído um módulo de pós-processamento que transforma o resultado do módulo de análise morfológica nessas etiquetas, de acordo com uma especificação dada pelo utilizador. Para tal, o utilizador, usando um editor de texto que manipule documentos em formato ASCII, constrói duas tabelas, output.tab e codigos.tab. A tabela output.tab permite definir as etiquetas em função da palavra, origem, tempo, classe, número, pessoa e género. A tabela codigos.tab permite definir agrupamentos de etiquetas, isto é, definir novas etiquetas em função das etiquetas obtidas através da tabela output.tab.

3.2. Modos de Operação

Existem várias opções que definem o modo de funcionamento do analisador morfológico. Usando o Morfolog no modo interactivo, sem nenhuma opção activa, os resultados teriam a seguinte forma:

```
bonezinho
bonezinho:
  nome, masculino, singular, Diminutivo
(bone', 100%)
  verbo bonezinhar, 1. pessoa do
singular do Presente Ind. (0%)
```

Quadro 6

Podemos observar neste exemplo duas particularidades: o acento em *boné* aparece no formato interno do Palavroso e uma classificação morfológica inesperada (verbo).

Normalização e acentuação. A primeira particularidade motiva esta opção – *opção a*. Quando activa, faz com que as palavras venham acentuadas no formato usado no exterior. Serve apenas para o analisador morfológico saber se a palavra submetida a análise já foi normalizada ou não.

Quantidade de informação. A segunda particularidade motiva as *opções u e v*, controladoras da quantidade de informação que o analisador dá como resposta. Quando alguma destas opções estiver activa, o resultado será apenas o confirmado pelo dicionário. Se a *opção u* está activa e não houver nenhuma hipótese de análise confirmada pelo dicionário, o resultado é constituído por todas as hipósteses de análise. Se a *opção v* está activa e não houver nenhuma resposta confirmada pelo dicionário, o analisador não dá qualquer resposta.

Utilização como dicionário booleano. Para uma utilização como dicionário Booleano, deve ser activada a *opção d*, que pára o processamento após ter sido encontrada uma análise válida, para além da *opção v* que, como vimos, dá como resultado apenas as análises válidas.

Depuração. Esta opção – *opção l* – quando activa, faz com que o analisador escreva num ficheiro (palav.log) uma descrição de todas as operações realizadas.

3.3. Componentes e Sequência de Processamento

O Palavroso é constituído por seis componentes essenciais:

1. Componente responsável pelas palavras fechadas.
2. Componente dos nomes e adjectivos.
3. Componente de processamento verbal.
4. Componente de processamento adverbial.
5. Componente de processamento de verbos com clíticos.
6. Componente de processamento de palavras compostas.

As várias componentes são processadas na sequência seguinte:

0. Iniciar variáveis.
1. Verificar se a palavra contém algum hífen. Se contiver, será efectuado o processamento relativo aos enclíticos e palavras compostas por justaposição. Senão, prosseguirá com o passo 2.
2. Efectuar processamento das palavras fechadas. Algumas destas podem ser homógrafas com palavras pertencentes a classes gramaticais abertas. Se não for o caso, o processamento parará aqui. Senão, continua com o passo seguinte.
3. Efectuar processamento relativo aos nomes e adjectivos.
4. Processar as palavras quanto à possibilidade de serem verbos.
5. Processar as palavras terminadas em *mente*, verificando se são advérbios.
6. Ordenar os resultados e formatá-los de acordo com a palavra analisada.

3.4 Dicionário – o léxico e as regras

O dicionário do sistema Palavroso está dividido em tabelas, conforme as categorias gramaticais das palavras, e contém a

classificação morfológica do lema. As principais tabelas de entradas lexicais são a dos nomes e adjectivos, a dos verbos, a dos advérbios, a das palavras gramaticais, a dos números, a dos estrangeirismos, e a das abreviaturas e siglas. No quadro 7, exemplificamos com uma tabela de palavras gramaticais, FECH.TAB:

Palavras gramaticais – entrada no dicionário:

Entrada	Número	Género	Classe	Subclasse	Probabilidade
a	S	F	art	def	1.0
cerca	I	F	adv	adv	1.0
sua	S	F	poss	3	1.0

Quadro 7

A tabela dos nomes e adjectivos é constituída por cinco campos, correspondentes a: lema, número, género, classe e probabilidade. A tabela dos verbos é composta apenas pelo lema do verbo e a informação probabilística relativa a este lema. De modo idêntico é constituída a tabela dos advérbios, contendo cada entrada, o advérbio e a informação probabilística associada.

As regras relativas à flexão também estão organizadas em tabelas e foram concebidas de forma a operarem com um alto nível de eficiência e com o máximo de economia.

Vejamos como exemplo as regras de pluralização de nomes e adjectivos:

Plural	Singular	Número
as	a	P
s	o	P
is	l	P
teis	til	P

Quadro 8

Onde a terminação *l* no singular vai originar *is* no plural (*funil/funis*), mas para obter o plural correcto de palavras como *fútil*, foi necessária uma nova regra, indicando que a terminação *til* dá origem a *teis* no plural (*útil-úteis*).

As regras para a flexão verbal são bem mais complexas. Compõem-se de três conjuntos de regras, denominados de regras produtoras, regras restritivas e regras de transformação. As regras produtoras sugerem classificações morfológicas possíveis para a palavra em análise. As regras restritivas excluem as classificações não passíveis de confirmação no dicionário. As regras de transformação apenas modificam algumas formas, como a transformação de *ce* em *que* (lema *ficar*: *ficei, ficaste...* → *fiquei, ficaste...*).

Padrão					Informação morfológica			Aplicabilidade				Probab.
ar	er	ir	or	Terminação	Temp.	Pess.	Núm.	ar	er	ir	or	Prob.
a	e	i	po	mos	pi	1	P	1	1	1	1	1.0
.	.	.	.	ímos	pi	1	P	0	0	1	0	1.0
av	i	i	punh	a	pii	3	S	1	1	1	1	1.0

Quadro 9

Esta tabela apresenta o formato das regras dos verbos regulares. Cada regra é composta por quatro partes: Padrão, Informação Morfológica, Aplicabilidade e Probabilidade. A componente Padrão refere-se à terminação da palavra a que vai ser aplicada a regra e ao conjunto de letras que antecede a terminação. Este conjunto de letras permite identificar a que conjugação pertence a forma. A segunda parte da regra, contém a Informação Morfológica que

vai ser atribuída à palavra. A terceira componente da regra permite definir a sua Aplicabilidade. Por, fim a última parte, Probabilidade, tem o significa já atrás referido.

Esta breve descrição da estrutura e funcionamento do Morfolog não se pretende exaustiva, mas apenas uma exemplificação que permita vislumbrar a filosofia que rege uma aproximação com base em regras e uso mais amplo do conhecimento linguístico.

4. O nível Morfológico do Projecto LE-PAROLE

4.1 Modelo, arquitectura e ferramenta de gestão lexical

O modelo PAROLE para codificação do léxico nos níveis morfológico, sintáctico e, em breve, semântico, foi baseado no modelo GENELEX (projecto EUREKA GENELEX, EU-524, JNICT) e nas recomendações do EAGLES. Este reaproveitamento do produto de outros projectos europeus visa, por uma lado, uma economia de recursos e, por outro, partir de uma base sólida e aglutinadora para a harmonização da codificação nas várias línguas.

Assim, a arquitectura do léxico PAROLE segue genericamente o modelo GENELEX, enriquecido com os resultados do EAGLES. O modelo está estruturado como um todo, interligando Unidades Morfológicas, Unidades Sintácticas e Unidades Semânticas, além de prever ligações multilingues.

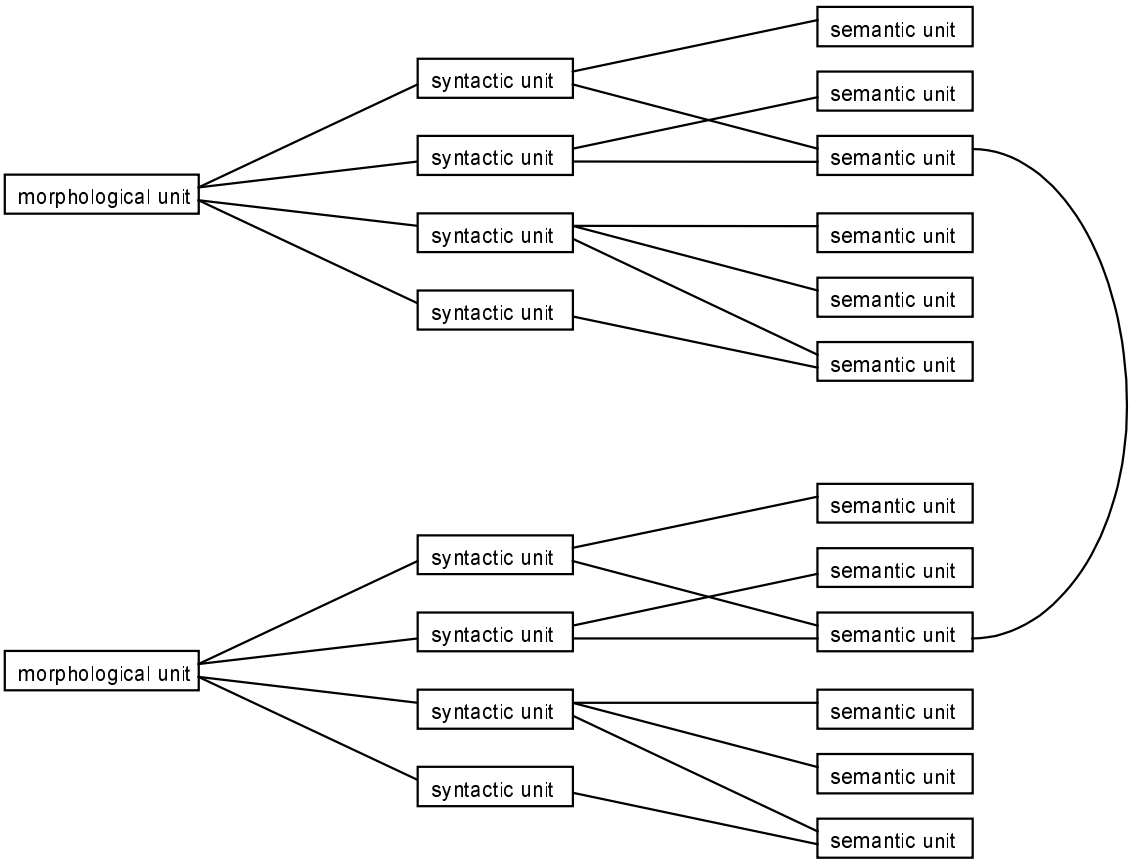

Figura 4: Arquitectura do PAROLE

Uma ferramenta de gestão lexical – AlethGD – foi desenvolvida pelo GSI-ERLI e utilizada por todos os parceiros, de modo a garantir a uniformidade de todas as bases de dados. O AlethGD assenta sobre o sistema de gestão de dados orientada a objectos, "Object Store". Contém as funcionalidades usuais e necessárias, incluindo importação e exportação de informação, para além da 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.

4.2 O modelo PAROLE para a morfossintaxe

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 possível definir um DTD comum, a ser utilizado na ferramenta comum (AlethGD).

No modelo PAROLE, a abordagem para codificar as flexões é paradigmática, em contraposição à abordagem por regras adoptado pelo sistema Palavroso. A abordagem paradigmática, embora menos económica, simplifica e torna mais eficaz a capacidade de harmonização do modelo entre as diversas línguas.

Esta abordagem consiste em criar um exemplo para cada tipo de flexão e ligar todas as outras entradas com o mesmo tipo de flexão a esse exemplo.

Uma Unidade Morfológica corresponde, genericamente, à noção de lema, ou entrada dos dicionários tradicionais. No léxico Português, uma Unidade Morfológica é uma Unidade Morfológica Simples (UMS), uma Unidade Morfológica Composta (UMC) ou uma Unidade Morfológica Contraída (UMCont). A base da codificação do léxico são as UMS. Cada UMS contém a sua categoria e subcategoria (quando aplicável) gramatical, bem como a sua forma gráfica principal e suas variações (como por exemplo *ouro/oiro*). Os Paradigmas de Flexão contêm a informação flexional associada a cada conjunto de traços morfológicos existente para a UMS a cuja forma gráfica está associado (ver figura 5). Os Paradigmas de Flexão incluiem o tratamento de clíticos. Tantos as UMC, como as UMCont são definidas com base nas várias Unidades Morfológicas, existindo Paradigmas de Flexão Compostos que são definidos tendo em conta a flexão de cada Unidade Morfológica usada na composição. Observem-se os seguintes exemplos:

1. a UMC c*aminho-de-ferro*, usaria as UMS *caminho*, *de* e *ferro*.
2. a UMC *marco do correio*, usaria as UMS *marco* e *correio* e a UMCont *do*, anteriormente definida com base nas UMS *de* e *o*.

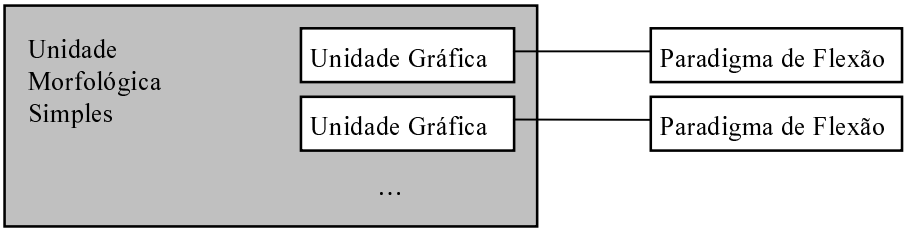

Figura 5: modelo PAROLE para a morfossintaxe

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

5. Conversão de classificações e regras do modelo PALAVROSO para o modelo PAROLE/GENELEX

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)

Quadro 10

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.

6. Outros aspectos do Projecto LE-PAROLE

Concomitantemente à criação dos léxicos, compreendendo selecção lexical com base em *corpus*, codificação morfossintáctica e codificação sintáctica, foi criado um *corpus* com 20 milhões de palavras em cada uma das línguas representadas. Estes *corpora* seguiram todos o mesmo desenho, uma percentagem aproximada quanto aos meios (jornal, 70%; livros, 20%; revista 5%; miscelânia, 5%), a mesma codificação em SGML, através de um DTD.

Um subcorpus com 250 mil palavras foi anotado no nível morfossintáctico, a partir de um conjunto de etiquetas previamente definido de modo a cobrir todas as línguas envolvidas e conferido manualmente. O subcorpus do Português foi anotado a partir do Morfolog do sistema Palavroso. Infelizmente as restrições de espaço não permitem uma exposição mais alargada dessa experiência de anotação de corpora de língua portuguesa.

7. Trabalhos em curso

Está em curso o seguimento do projecto LE-Parole para a codificação dos léxicos no nível semântico, através do projecto SIMPLE, também financiado pelo Programa *Telematics* da CE - DG XIII.

Por outro lado, encontra-se em fase de tramitação um projecto para a disponibilização de parte dos recursos através da Internet, nomeadamente um subcorpus de 3 milhões de palavras correntes. A hipótese de disponibilização dos léxicos está em estudo.

8. Conclusão

A título de conclusão, gostaríamos de sublinhar a importância de projectos de desenvolvimento de recursos linguísticos qualitativamente bem estruturados, em paralelo à investigação avançada. Aplicada ao Português, essa afirmação ganha mais força, visto ser a terceira língua ocidental mais falada, e face a necessidade de dar resposta à crescente demanda e à absoluta necessidade de manter o Português como uma língua activa na sociedade da informação.

Bibliografia

Bacelar do Nascimento, Maria Fernanda; Bettencourt, José e Wittmann, Luzia. *Lexicon Morphosyntactic Specifications. Portuguese instantiation of the EAGLES.* MLAP PAROLE report. CLUL/INESC, Lisboa, Dezembro de 1995.

Bacelar do Nascimento, Maria Fernanda; Marrafa, Palmira; Santos Pereira, Luísa Alice; Ribeiro, Ricardo Daniel; Veloso, Rita e Wittmann, Luzia. *LE-PAROLE – Do corpus à modelização da informação lexical num sistema multifunção* in Actas do XIII Encontro da Associação Portuguesa de Linguística, Lisboa, 1998. (no prelo)

Medeiros, José Carlos. *Processamento Morfológico e Correcção Ortográfica do Português*. Tese de Mestrado. IST, Lisboa, 1995.

Pinto, Edite. *Relatório de Estágio*. UBI/INESC, Lisboa, 1997.

Santos, Diana; Medeiros, José Carlos e Wittmann, Luzia. *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, 1995.