

μ 8

**UNIVERSIDADE TÉCNICA DE LISBOA**

INSTITUTO SUPERIOR TÉCNICO

**Processamento Morfológico e  
Correcção Ortográfica do Português**

José Carlos Dinis Medeiros

Licenciado

Dissertação para obter o grau de Mestre em Engenharia Electrotécnica e de Computadores

Lisboa e IST, Fevereiro de 1995

Tese realizada sob a orientação de  
Professora Doutora Eng. Isabel Maria Martins Trancoso  
Professora Associada  
do Departamento de Engenharia Electrotécnica e de Computadores do

# Instituto Superior Técnico

## **Resumo**

Esta dissertação de tese trata o problema do processamento automático da morfologia e correcção ortográfica do português, descrevendo duas aplicações: o Palavroso, um analisador morfológico do português europeu, e o Correcto, um corrector ortográfico que usa o Palavroso.

O Palavroso é um analisador morfológico de utilização geral, que cobre cerca de um milhão e trezentas mil formas de português. Pode ser usado como componente morfológica e lexical de sistemas de processamento de linguagem natural para português, sendo a correcção ortográfica uma das suas aplicações possíveis.

O Correcto possui alguns aspectos inovadores em relação a outros correctores existentes para português. Inova, particularmente, no que diz respeito ao tratamento de erros que envolvam verbos com enclíticos e pelo facto de ter um analisador morfológico subjacente.

Dá-se ainda especial atenção à avaliação de desempenho de correctores ortográficos, por este ser um aspecto fundamental no desenvolvimento de aplicações. Apresenta-se uma metodologia para avaliação de correctores ortográficos, baseada em medidas de desempenho sobre determinados tipos de corpos de teste, aplicando-a a vários correctores existentes.

## **Palavras Chave**

Processamento de linguagem natural; correcção ortográfica; avaliação de correctores ortográficos; morfologia; análise morfológica.

## **Abstract**

The aim of this thesis is the study of automatic morphological processing and spelling checking for Portuguese. We describe two systems which match those goals: Palavroso, a morphological analyser for European Portuguese, and Correcto, a spelling checker and corrector based on Palavroso.

Palavroso is a general purpose morphological analyser which covers around 1 300 000 citation forms and can be used as morphological and lexical component in natural language processing systems for Portuguese. Spelling checking and correction is one of its possible applications.

The spelling checker/corrector Correcto has some innovative features when compared with other existing spelling checkers/correctors for Portuguese. Namely, the processing of errors involving verbs with enclitics and its subjacent morphological analyser.

Another topic discussed in this thesis is the evaluation of the performance of spelling correctors, a fundamental issue in applications development. The evaluation methodology presented in this thesis is based on performance measures obtained running the spelling checker over certain types of corpora and can be used to objectively compare the performance of several systems.

## **Key words**

Natural language processing; spelling checking; evaluation of spelling checkers; morphology; morphological analysis.

## **Agradecimentos**

Quero agradecer a todos os meus amigos que de alguma forma estiveram presentes durante a elaboração deste trabalho. Pela amizade e pelas listas de erros ortográficos que de vez em quando me ofereciam.

Este trabalho foi possível em grande parte devido a uma bolsa de mestrado concedida pela JNICT, no âmbito do programa CIÊNCIA, bem como ao apoio concedido pelo INESC.

Agradeço aos elementos do Grupo de Linguagem Natural do INESC a sua amizade e a colaboração no desenvolvimento do Palavroso: a eles se deve o preenchimento lexical do Palavroso.

Ao José João agradeço as conversas que tivemos, as sugestões e elementos bibliográficos; ao Dr. Manuel Costa, agradeço as conversas sobre vários aspectos da língua portuguesa e as listas de erros ortográficos; à Luzia agradeço as sugestões dadas na correcção deste texto.

Agradeço à Professora Isabel Trancoso a sua disponibilidade para orientar este trabalho, todas as críticas e o rigor exigido.

Duas pessoas merecem destaque nos meus agradecimentos: a Diana Santos, não só pelas várias revisões deste documento, mas por tudo quanto me tem ensinado, pela amizade e pelo encorajamento constante; a Nazaré, minha mulher, pelas privações a que foi obrigada para que eu pudesse estudar.

Por fim, dedico este trabalho ao meu filho, Zé Pedro; um dia ele poderá falar com os computadores.

# Índice

<b><i>Introdução</i></b> .....	<b>1</b>
<b><i>1 Morfologia</i></b> .....	<b>4</b>
2 Tipos de línguas.....	6
3 Alguns aspectos da morfologia portuguesa.....	9
4 Nomes e adjectivos.....	9
5 Verbos.....	13
6 Outras classes gramaticais.....	15
7 Morfologia e léxico no processamento automático de linguagem natural.....	15
8 Léxico, morfologia e dicionários.....	19
<b><i>9 O Palavroso, um analisador morfológico</i></b> .....	<b>22</b>
10 Arquitectura.....	23
11 Comunicação com o exterior.....	25
12 Pré-processamento.....	26
13 Pós-processamento.....	27
14 Modos de operação.....	32
15 Normalização e acentuação.....	32

16	Quantidade de informação.....	34
17	Utilização como dicionário booleano.....	35
18	Depuração.....	36
19	As componentes.....	36
20	Palavras fechadas.....	37
21	Nomes e adjectivos.....	40
22	Processamento quanto ao número	
	.....	
42		
23	Processamento quanto ao género	
	.....	
44		
24	Processamento quanto ao grau	
	.....	
46		
25	Dicionário	
	.....	
48		
26	Verbos.....	49
27	Regras produtoras	
	.....	
49		
28	Regras de transformação	
	.....	
53		
29	Regras de restrição	
	.....	
57		
30	Dicionário	
	.....	
59		
31	Funcionamento global	
	.....	
60		
32	Advérbios de modo.....	61

33	Verbos com enclíticos.....	62
34	Palavras compostas.....	66
35	Utilizações do Palavroso.....	68
36	Em ferramentas de processamento de corpos de texto.....	69
37	Em gramáticas computacionais.....	69
38	Estudos efectuados com o Palavroso.....	70
39	Alguns dados sobre a implementação do Palavroso.....	71
40	Resumindo.....	74
<b>41</b>	<b>Correcção ortográfica.....</b>	<b>75</b>
42	O problema.....	76
43	Modos de correcção.....	77
44	Cobertura lexical.....	78
45	Medidas de distância.....	79
46	Distância n-grâmica.....	79
47	Distância de edição.....	80
48	Matriz de semelhança.....	80
49	Classificação do tipo de erros.....	82
50	Erros linguísticos.....	83
51	Erros tipográficos.....	84
52	Erros de transmissão.....	86
53	Análise de erros ortográficos.....	86
54	Erros linguísticos.....	87
55	Semelhanças fonológicas	
	.....	
88		
56	Incorrecção morfológica	
	.....	
90		
57	Erros de acentuação	
	.....	
97		

58	Criação	de	regras	
.....				
98				
59	Erros tipográficos.....			101
60	Distribuição	dos	tipos	de erro
.....				
101				
61	Posição	do	erro	
.....				
104				
62	Distância	no	teclado	
.....				
106				
63	Erros de origem linguística considerados como tipográficos			
.....				
107				
64	Processamento	de	trigramas	
.....				
110				
65	Abordagens utilizadas.....			114
66	Métodos de correcção básicos.....			116
67	Corrector	de	Damerau	
.....				
116				
68	Corrector	de	Peterson	
.....				
121				
69	Chaves de semelhança e abreviaturas.....			123
70	Trigramas.....			124
71	Fonemas e trifones.....			128
72	Autómatos finitos.....			131
73	O método ideal.....			132
74	<i>Correcto,</i>	<i>um</i>	<i>corrector</i>	<i>que usa o Palavroso</i>
.....				
134				

75	Descrição global do Correcto.....	135
76	Interface com outras aplicações.....	136
77	Sessão de correcção.....	138
78	Verificação.....	140
79	Sugestões.....	141
80	Verbos com enclíticos e palavras compostas.....	142
81	Caso geral.....	145
82	Erros complexos.....	146
83	Processamento relativo aos acentos.....	147
84	Regras heurísticas.....	151
85	Maiúsculas.....	156
86	Processamento trigrâmico.....	156
87	Falta de espaço.....	161
88	Resumindo.....	162
<b>89</b>	<b><i>Parâmetros de avaliação de correctores ortográficos</i></b> .....	
<b>163</b>		
90	Velocidade de processamento.....	167
91	Correcção dos resultados.....	168
92	Dispersão das sugestões.....	169
93	Ordenação das sugestões.....	171
94	Insucesso.....	171
95	Eficiência do motor.....	174
96	Corpos de teste. Índice de correcção.....	175
97	Funcionalidade.....	177
98	Concluindo.....	178
<b>99</b>	<b><i>O Correcto e os outros</i></b> .....	
<b>180</b>		
100	Os outros.....	181

101	Preparação do teste.....	182
102	Resultados.....	183
103	Avaliação global.....	186

**Conclusão**

.....  
**187**

**Bibliografia**

.....  
**192**

**Apêndice A - Os corpos de teste**

.....  
**197**

A.1. Corpo sem erros.....197

A.2. Corpo de palavras com erros.....199

**Apêndice B - Rótulos usados no Palavroso**

.....  
**203**



## Introdução

Quando pensamos nos "computadores do ano 2000", invariavelmente imaginamos computadores capazes de raciocínios à imagem do Homem, sobretudo capazes de comunicar com este da maneira mais confortável: através de linguagem natural.

Nesta dissertação abordamos o problema da morfologia no processamento automático de linguagem natural sob dois aspectos. Por um lado, discutimos alguns problemas da morfologia e descrevemos um analisador morfológico para português europeu: o Palavroso. Por outro lado, demonstramos a utilização efectiva de uma ferramenta linguística numa aplicação de uso corrente: a correcção ortográfica.

O Palavroso é um analisador morfológico completo e robusto. Cobre mais de 1 300 000 formas do Português e é uma ferramenta linguística capaz de resolver eficazmente o problema do processamento morfológico em sistemas mais abrangentes do processamento de linguagem natural.

Como características principais deste analisador morfológico apontamos o processamento baseado em regras, a separação entre as regras morfológicas e o léxico, e, como consequência destas características, a possibilidade de dar sempre algum tipo de resposta, mesmo que a sua componente lexical não seja suficientemente representativa.

O Correcto, um corrector ortográfico que desenvolvemos e aqui descrevemos, usa o Palavroso como componente lexical e morfológica. Este corrector apresenta algumas inovações em relação a outros correctores ortográficos para o Português, possibilitando-lhe, por exemplo, o tratamento de erros que envolvem verbos com enclíticos. Esta característica só é possível devido à utilização de um analisador morfológico para processamento do léxico.

Devido à importância da avaliação de sistemas computacionais, abordamos o problema da avaliação de desempenho de correctores ortográficos. Nesse sentido, sugerimos uma metodologia de avaliação destes sistemas baseada em medidas calculadas a partir de corpos de teste. Como aplicação desta metodologia, avaliamos e comparamos vários correctores ortográficos para português actualmente comercializados.

Esta dissertação divide-se em três partes fundamentais: principia com uma abordagem à análise morfológica, passando a seguir para a correcção ortográfica e finaliza com a avaliação de correctores ortográficos.

O capítulo sobre morfologia tem dois objectivos básicos. Por um lado, introduzir alguns conceitos relacionados com a morfologia, usados ao longo da dissertação. Por outro lado, motivar o processamento automático da morfologia e enquadrá-lo no processamento automático de linguagem natural.

Segue-se um capítulo destinado à descrição do analisador morfológico Palavroso. Optou-se por fazer, primeiro, uma descrição genérica da sua arquitectura, utilização e modos de operação. Só depois as diversas componentes do sistema são descritas mais pormenorizadamente, procurando-se esclarecer as soluções encontradas para os vários problemas da análise morfológica do Português.

No terceiro capítulo, inicia-se a discussão da correcção ortográfica. Focando os vários aspectos envolvidos nesta questão, abre-se caminho para os capítulos seguintes. Descreve-se o problema da correcção ortográfica e introduzem-se os conceitos e definições usados no restante da dissertação. De seguida, classificam-se os tipos de erros ortográficos e faz-se uma análise dos erros ortográficos usualmente encontrados em textos portugueses. Esta análise fornece informação que será, depois, usada no estabelecimento de estratégias de correcção adoptadas pelo Correcto. Por fim, são apresentadas algumas abordagens conhecidas na resolução do problema da correcção ortográfica. Esta apresentação ajuda a ponderar os prós e os contras das várias estratégias conhecidas e a definir os critérios a que deve obedecer um bom algoritmo de correcção.

No quarto capítulo descreve-se o Correcto. Tal como na descrição do Palavroso, parte-se de uma descrição genérica, em que se procura dar a ideia global do método usado, para a seguir descrevê-lo mais detalhadamente, explicando em pormenor as soluções encontradas para alguns problemas específicos.

Os dois capítulos seguintes incidem sobre a avaliação de correctores ortográficos. No primeiro,

motiva-se a necessidade da criação de parâmetros objectivos de avaliação, propondo algumas medidas que podem ser usadas com este objectivo. No segundo, demonstra-se a aplicação destas medidas comparando entre si cinco correctores ortográficos para português, incluindo o Correcto.

Na conclusão procuramos inserir algumas ideias para o trabalho futuro, no seguimento da tese apresentada nesta dissertação.

Resta referir que ambas as aplicações aqui descritas – analisador morfológico e corrector ortográfico – foram desenvolvidas no Grupo de Linguagem Natural do INESC. A parte de correcção ortográfica foi exclusivamente concebida e implementada pelo autor, enquanto que o analisador morfológico foi implementado pelo autor e concebido sob a orientação de Diana Santos. O preenchimento lexical do analisador morfológico e a programação das regras morfológicas é da responsabilidade de Rui Marques, Maria de Jesus Pereira e Anabela Barreiro, sob a orientação de Diana Santos.

# 1 Morfologia

Há autores (Florido e Silva, 1978) que restringem a morfologia à flexão. Outros, como Figueiredo e Ferreira (1974), embora não definam explicitamente o que é a morfologia, incluem neste tópico apenas a flexão das palavras, deixando a derivação para o capítulo referente ao léxico.

De facto, até fins do séc. XVIII, nem sequer aparecia nas gramáticas uma secção para a morfologia. Em vez disso, existia uma secção para a flexão e outra para a derivação. Só no século dezanove se começou a usar o termo morfologia para cobrir tanto a flexão como a derivação. Se quisermos definir morfologia por oposição à sintaxe, podemos dizer que a morfologia lida com a estrutura interna das palavras, enquanto a sintaxe trabalha com a forma como as palavras se relacionam para formar frases (Lyons, 1968).

O que para o leigo podem ser duas palavras diferentes, para o linguista podem ser formas diferentes da mesma palavra. Por exemplo, *livro* e *livros* são formas da mesma palavra (Lyons, 1968).

Na distinção de palavras, o importante é o significado lexical. Como o processo de flexão preserva o significado lexical, são consideradas formas da mesma palavra as formas que se obtêm por um processo de flexão. Por outro lado, a derivação modifica o significado, pelo que é considerado um

processo de construção de novas palavras (Vilela, 1994; Lyons, 1968; Figueiredo e Ferreira, 1974).

Na análise da estrutura interna das palavras, é costume dividi-las em partes constituintes. Chama-se "morfema" à menor parte portadora de significado em que uma palavra pode ser decomposta (Galisson e Coste, 1976; Vilela, 1994). Existem vários tipos de morfemas, que são classificados de acordo com a sua função na palavra. Destacam-se, entre outros, os morfemas de base, os morfemas derivativos e os morfemas flexivos.

Os morfemas de base, também denominados lexemas ou radicais, são aqueles que contêm o significado lexical. Se considerarmos uma família de palavras, o lexema (radical) é o elemento comum a todas elas. É a forma de partida para todas as palavras da mesma família. Por exemplo, nas palavras *comparar*, *comparação*, *comparativo* e *incomparável*, o elemento comum portador do significado lexical é *compar*, sendo este, portanto, o lexema ou radical.

Os morfemas flexivos, por oposição aos derivativos, são aqueles que não alteram o significado lexical da base. Os derivativos são usados para criar outras palavras da mesma família do morfema de base, ao passo que os flexivos são usados para criar as várias formas da mesma palavra. Estes também são chamados gramaticais, pois servem para designar o número, género, pessoa, tempo e aspecto.

Lyons (1968) distingue dois níveis de segmentação das palavras: um relacionado com a forma, o outro com a substância. Um morfema é um elemento de forma. A sua materialização (substância) ao nível fonético é um morfe<sup>1</sup>. Seguindo a sua notação, o morfema {azul} é representado fonologicamente por /z'u /<sup>2</sup> e ortograficamente por *azul*. A palavra *vendedores* é composta pelos morfemas {vende} + {dor} + {es}, que são representados fonologicamente pelos morfes /ve~d«/ + /d'oR/ + /«S/. Por outro lado, na definição de morfema nada nos diz que ele tem que estar explicitamente representado na palavra. Assim, podemos considerar que a forma *foi* do verbo ser (fonologicamente /f'õ /) é constituída pelos morfemas {s} + {eu} e não pode ser segmentada em morfes.

Por vezes, um determinado morfema não aparece sempre representado pelo mesmo morfe. Por exemplo, *corrig-ir* vs. *correc-ção*. Neste caso, as representações alternativas do mesmo morfema são chamadas alomorfes (Lyons, 1968; Vilela, 1994).

---

1 No original, o autor identificava estes elementos por *morphs*. Isto é, um *morph* consiste num morfema representado fonologicamente. Como não encontramos na bibliografia disponível nenhum termo correspondente, adoptámos esta tradução.

2 Ao longo deste texto, usamos os símbolos fonéticos de acordo com o Alfabético Fonético Internacional (AFI).

## 2 Tipos de línguas

As línguas podem ser classificadas em analíticas ou sintéticas. Dentro destas, distinguem-se as aglutinativas das fusivas (Lyons, 1968).

As línguas analíticas (ou isolantes) são aquelas em que todas as palavras são invariáveis, ou seja, cada palavra tem apenas uma forma.

Podemos definir línguas analíticas recorrendo à noção de morfema. Assim, diz-se que uma língua é analítica se a cada palavra corresponder apenas um morfema. Um exemplo comumente dado de língua isolante é o Vietnamita.

O que caracteriza as línguas sintéticas é a possibilidade das palavras poderem ser decompostas em morfemas. Casos típicos deste tipo de língua são o Turco, como exemplo quase perfeito de língua aglutinativa, e o Latim, como paradigma de língua fusiva.

Numa língua aglutinativa, existe uma correspondência biunívoca entre os morfemas e os morfes. Por outro lado, é sempre possível segmentar uma palavra nos seus morfes constituintes<sup>1</sup>.

Tomando o Turco como exemplo, o morfema {ler} significa plural, {i} significa posse (dele, dela) e {den} tem função ablativa. Sabendo que *ev* significa casa, podemos construir uma série de formas por concatenação os vários morfemas:

Forma	Descrição
<i>ev</i>	casa
<i>evler</i>	casas
<i>evi</i>	casa dele/dela
<i>evden</i>	da casa

---

<sup>1</sup> Repare-se que em Português tal não acontece, como no exemplo dado da forma *foi*, que, aliás, também é forma do verbo ir.

<i>evleri</i>	casas deles/delas
<i>evlerden</i>	das casas
<i>evinden</i> <sup>1</sup>	da casa dele/dela
<i>evlerinden</i>	das casas deles/delas

---

Partindo de um mesmo radical, usando os vários morfemas do Turco, pode-se conseguir construir milhares, ou mesmo milhões, de palavras (Oflazer, 1994).

O que caracteriza as línguas fusivas é, por um lado, a impossibilidade de dividir coerentemente uma palavra em morfes, e, por outro lado, a não existência de correspondência entre morfes e morfemas a maior parte das vezes.

Embora seja usual classificar as línguas quanto ao seu tipo (analítico, aglutinativo, ou fusivo), deve-se reconhecer que não existem tipos puros. Quando muito, podemos falar de tendências mais ou menos acentuadas. O Chinês, por exemplo, é fortemente analítico, mas tem algumas palavras compostas por mais do que um morfema.

Pode-se medir o grau de análise de uma língua relacionando o número de morfemas com o número de palavras da língua. A razão destes valores dá o grau de isolamento da língua. A língua analítica ideal tem razão unitária. Quanto mais próximo de 1 for o valor da razão, mais analítica é a língua.

O Português é uma língua predominantemente fusiva, embora se possam encontrar várias situações de aglutinação.

Por exemplo, todas as palavras invariáveis da língua contribuem para o seu afastamento do tipo fusivo. Características aglutinativas podem ser encontradas principalmente na derivação, onde é possível segmentar a palavra em morfes, havendo correspondência entre os morfes e os morfemas.

Um exemplo:

<b>Palavra</b>	<i>faladoras</i>
----------------	------------------

---

<sup>1</sup> A inserção do *n* entre o *i* e *den* é regular e obrigatória.

<b>Morfemas</b>	{fala} + {dor} + {as}
<b>Morfes</b>	/f   / + /d'oR/ + / S/

O tipo a que a língua pertence tem uma grande influência na sua morfologia. Se a língua é analítica pura não tem morfologia. Quanto mais uma língua se aproxima do tipo sintético, mais rica tende a ser a sua morfologia. Isto é, as palavras transportam consigo mais informação morfológica.

O grau de aglutinação da língua também tem influência na expressividade da sua morfologia. Quanto mais aglutinativa for a língua, mais palavras se podem criar a partir de um único radical. No Português, por exemplo, o número de palavras e formas que é possível criar a partir de um radical mede-se na ordem das centenas<sup>1</sup>. Já no Finlandês, mede-se na ordem dos milhares (Koskenniemi, 1983); no Turco, esse valor pode ascender ao milhão (Oflazer, 1994).

Quanto às dificuldades do processamento morfológico, também há diferenças entre as línguas aglutinativas e as fusivas. Nas primeiras, a dificuldade encontra-se em especificar todos os tipos de derivação e flexão que possíveis, pois o encadeamento dos morfemas pode ser bastante grande e complexo. A grande vantagem destas línguas é a possibilidade de facilmente segmentar uma palavra nos seus morfemas.

Nas línguas fusivas, o difícil é segmentar as palavras em morfemas. Como na maior parte das vezes não existe uma correspondência entre morfemas e morfes, a segmentação em morfes não é possível. Embora o encadeamento de morfemas não seja tão grande como nas línguas aglutinativas, a tarefa de especificar as regras de interligação dos morfemas pode ser bastante complexa, principalmente devido às alterações fonológicas provocadas pela ligação entre certos morfemas. Por outro lado, os morfemas gramaticais não reúnem em si só um tipo de informação. Por exemplo, nem sempre é possível separar um morfema para caracterizar o número e outro para caracterizar o género. Há situações em que um morfema reúne as duas características. Com os verbos acontece algo idêntico no caso da pessoa e do tempo.

---

<sup>1</sup> Um verbo regular tem 65 formas simples. Se considerarmos a conjugação pronominal (conjugação pronominal no sentido do verbo *vir* ligado a um pronome, como em Cuesta e Luz, 1971), facilmente se atingem as centenas de flexões para o mesmo radical. A este número ainda podemos juntar todas as palavras que podem ser derivadas de um radical, quer por prefixação, quer por sufixação.

### 3 Alguns aspectos da morfologia portuguesa

O Português, como descende do Latim, é uma língua fusiva. Aplicam-se-lhe, portanto, algumas das dificuldades acima referidas.

Para ultrapassar parte dessas dificuldades, é corrente adoptarem-se formalismos diferentes para descrever as flexões do campo verbal e as do campo nominal. No campo verbal temos as conjugações. No campo nominal temos as declinações.

Em ambos os casos, a palavra é dividida em tema e desinência. O tema é a parte da palavra que serve de base à flexão. É composto pelo radical (ou raiz) da palavra mais a vogal temática. O radical é o elemento comum a uma família de palavras (cf. início deste capítulo).

A desinência é o morfema que se junta ao tema para exprimir diversas noções a acrescentar ao significado da palavra. Nos nomes indica o número e o género. Nos verbos indica o número e a pessoa (Figueiredo e Ferreira, 1974; Vilela, 1994).

Embora o tema e a desinência sejam comuns no caso verbal e no caso nominal, há elementos que são característicos somente de determinadas classes gramaticais. Com isto alargamos a discussão às outras classes gramaticais, como os pronomes e artigos.

### 4 Nomes e adjectivos

No que diz respeito à flexão, os nomes e adjectivos podem flexionar quanto ao número, género e grau, de acordo com o quadro 1.1. (Figueiredo e Ferreira, 1974).

<b>Flexão dos nomes e adjectivos</b>	Número	Singular
		Plural
	Género	Masculino

		Feminino	
	Grau	Nomes	Aumentativo  Diminutivo
		Adjectivos	Superlativo  Comparativo  Normal

Quadro 1.1. - Flexão dos nomes e adjectivos segundo Figueiredo e Ferreira (1974). Alguns autores consideram mais graus. Por exemplo, Cunha e Cintra (1987) acrescentam o grau normal dos nomes; Barreiro et al. (1993) acrescentam aos adjectivos os graus aumentativo e diminutivo.

Geralmente, a flexão é feita por sufixação de morfemas flexivos ou desinências. Há situações em que a flexão é feita por um processo de aglutinação, outras em que se observa um processo de fusão e outras ainda de semi-aglutinação. Tomemos como exemplo a palavra *gato* e construamos algumas das suas flexões.

	Diminutivos	Aumentativos
gato	gatinho	gatão
	gatinhos	gatões
gata <sup>1</sup>	gatinha	gatona

<sup>1</sup> Segundo Cunha e Cintra (1987), *gata* obtém-se por flexão a partir de *gato*. Segundo Lyons (1968), ou Barreiro et al. (1993), corresponde a um item lexical diferente.

gatinhas

gatonas



Para os diminutivos há uma grande regularidade. Podemos considerar que as flexões foram obtidas por um processo de aglutinação de vários morfemas:

tema + morfema de diminuição + morfema do género + morfema do número

Concretizando:

Palavra	Tema	Morfema de diminuição	Morfema do género	Morfema do número
gatinho	{gat}	{inh}	{o}	{}
gatinhos	{gat}	{inh}	{o}	{s}
gatinha	{gat}	{inh}	{a}	{}
gatinhas	{gat}	{inh}	{a}	{s}

Em relação aos aumentativos, temos casos de fusão e casos de semi-aglutinação. A forma *gatão* é composta por dois morfemas: {gat} + {ão}. O primeiro morfema é o tema da palavra. O segundo, {ão}, é o morfema flexivo, que reúne em si informação relativa ao grau, número e género: é um morfema que caracteriza a flexão de aumentativo masculino singular.

De forma idêntica podemos classificar *gatões*, em que o morfema flexivo {ões} é característico de aumentativo masculino plural.

Os casos de aumentativo feminino são ligeiramente diferentes. Podemos considerar que aqui existe um processo de semi-aglutinação, pois embora não seja possível fazer a segmentação feita para o caso dos diminutivos, ainda é possível fazer alguma segmentação. Poderíamos, por exemplo, reunir a informação relativa a aumentativo feminino num morfema, considerando a existência de outro para o número:

tema + morfema de diminuição e gênero + morfema do número

Concretizando a ideia, *gatonas* seria segmentado em {gat} + {ona} + {s}, com o morfema {ona} reunindo as características de aumentativo feminino e o morfema {s} marcando o plural.

## 5 Verbos

Segundo Figueiredo e Ferreira (1974), os verbos flexionam em:

**Número** Tal como no nome e o adjetivo, o verbo pode ser singular ou plural.

**Pessoa** São três: a primeira (eu, nós), a segunda (tu, vós) e a terceira (ele/ela, eles/elas).

**Voz** Pode ser activa ou passiva.

**Modo** Indicativo, Conjuntivo, Imperativo, Condicional, Infinitivo.

**Tempo** Presente, pretérito (perfeito, imperfeito e mais-que-perfeito) e futuro (perfeito e imperfeito).

Há autores que consideram a existência de tempos derivados de outros (Cunha e Cintra, 1987). Assim, temos três tempos primitivos dos quais os outros são derivados:

<b>Tempo primitivo</b>	<b>Tempos derivados</b>
------------------------	-------------------------

Presente do indicativo	Presente do conjuntivo, imperativo, pretérito imperfeito do indicativo
Pretérito perfeito do indicativo	Pretérito mais-que-perfeito, futuro do conjuntivo e pretérito imperfeito do conjuntivo.
Infinitivo impessoal	Futuro imperfeito do indicativo, condicional, infinitivo pessoal, particípio passado, gerúndio.

Como nos nomes e adjectivos, é possível segmentar as formas dos verbos isolando os seus constituintes. Estas partes são:

<b>Tema</b>	É a parte fundamental do verbo e serve de base à flexão. Como no caso nominal, obtém-se acrescentando ao radical a vogal temática (ou retirando o <i>r</i> do infinitivo impessoal). A vogal temática determina a conjugação a que o verbo pertence.
<b>Característica</b>	Morfema que se acrescenta ao tema para caracterizar um tempo ou um modo.
<b>Desinência pessoal</b>	Indica a pessoa e o número

Exemplos:

Forma	Tema	Característica	Desinência	Conjugação
cantavas	canta	av	as	1ª conjugação
vendereis	vende	r	eis	2ª conjugação

partíssemos      parti              isse              mos              3ª conjugação

---

## 6 Outras classes gramaticais

Algumas classes gramaticais são invariáveis. É o caso das preposições, dos advérbios, das conjunções e das interjeições, em que cada forma é independente das outras. Outras classes admitem variações nas suas formas para representar determinadas características. Os artigos, por exemplo, variam quanto ao género e quanto ao número.

Nos casos que admitem variações, estas referem-se, em geral, apenas ao número e ao género. Contudo, existem casos particulares, como os pronomes pessoais e os pronomes possessivos, que também possuem características que os identificam com a pessoa. Por exemplo,

<b>Pessoa</b>	<b>Pessoal</b>	<b>Possessivo</b>
1ª pessoa	<i>eu, nós</i>	<i>meu, nosso</i>
2ª pessoa	<i>tu, vós</i>	<i>teu, vosso</i>
3ª pessoa	<i>ele, eles</i>	<i>dele, deles</i>

---

## 7 Morfologia e léxico no processamento automático de linguagem natural

A linguagem natural é um meio de transmissão de informação. Para que essa transmissão efectivamente ocorra, o receptor tem que interpretar as frases que lê ou ouve, atribuindo-lhe um

determinado valor semântico. Para os humanos, esse processo pode ser efectuado sem quaisquer conhecimentos gramaticais explícitos da língua usada. Basta lembrarmos que as crianças também compreendem a língua (ou línguas) que falam. No processamento automático de linguagem natural, não é possível fazer uma análise semântica da informação sem recorrer aos aspectos formais da língua. Por outras palavras, é necessário recorrer à análise sintáctica da frase, que por sua vez tem que recorrer à informação morfológica das palavras que a constituem.

Dependendo do tipo de aplicação e dos objectivos impostos, em qualquer sistema de processamento automático de linguagem natural, podemos identificar três níveis de manipulação de informação que são geralmente usados.

Nível	Descrição
1 - Nível lexical e morfológico	A informação é limitada à palavra, sem entrar em linha de conta com o contexto em que é usada.
2 - Nível sintáctico	A informação diz respeito à relação que existe entre os vários constituintes de uma frase, num sentido meramente formal.
3 - Nível semântico e pragmático	A informação é sobre o significado da associação das palavras e o significado da frase no seu conjunto.

Quadro 1.2. - Níveis de manipulação de informação.

Estes níveis são hierárquicos: não é concebível que se manipule (use e produza) informação semântica sem recorrer à sintaxe. De forma idêntica, não é possível trabalhar com sintaxe sem recorrer à morfologia.

Em última análise, o processamento de linguagem natural envolve sempre conhecimento ao nível 3. No entanto, para simplificar o processamento, fazem-se aproximações mais ou menos rudimentares, conforme o nível de sofisticação requerido. Por exemplo, há situações na correcção ortográfica em que é necessário recorrer, no mínimo, a uma análise sintáctica para detectar determinados erros. No entanto, são poucos os correctores ortográficos que o fazem. Só pontualmente aparecem correctores a este nível, geralmente em ambientes académicos. Na maioria dos casos, podemos dizer que **correcção ortográfica = léxico + estatística**.

Qualquer sistema de processamento automático de linguagem natural usa, no mínimo, informação

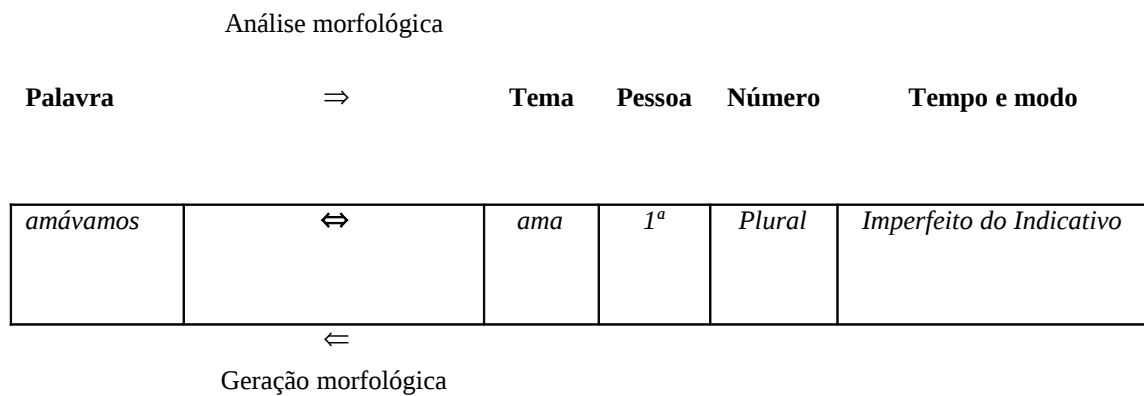
lexicológica.

A informação lexicológica compreende **léxico + morfologia + semântica das palavras**. No entanto, estas três componentes nem sempre são usadas. Se o objectivo é a análise sintáctica de uma frase, a semântica das palavras pode não ser importante, usando-se apenas a informação morfológica.

Não é normal o léxico de uma aplicação conter todas as formas de palavras que podem ser obtidas por um processo de flexão. O mais natural é haver uma componente no sistema que faça processamento morfológico. Este processamento pode envolver a análise ou geração morfológica.

A análise morfológica consiste no processo de, dada qualquer palavra da língua, determinar todas as suas características morfológicas. Inversamente, a geração morfológica consiste em, dado um conjunto de características morfológicas, construir a palavra que as possui.

Tomemos como exemplo a palavra *amávamos*. Podemos esquematizar os dois processos da seguinte forma:



O tipo de informação usado em ambos os processos é essencialmente o mesmo, pois baseia-se na decomposição da palavra nos seus morfemas, como ilustrada na figura 1.3 (Figueiredo e Ferreira, 1974):

μ §

Figura 1.3 - Decomposição nos seus morfemas da forma verbal *amávamos*. Nesta forma, o acento agudo é colocado para manter o acento tónico na vogal temática, por analogia com a forma latina, *amabamus*.

Há sistemas que envolvem a criação automática de mensagens. Podemos mencionar os sistemas que dialogam interactivamente com o utilizador e os geradores automáticos de relatórios técnicos

(Kerpedjiev, 1992; Reiter et al., 1992; Rösner & Stede, 1994), em particular, os geradores de cartas comerciais (Coch, 1994).

Este tipo de aplicações necessita de geração morfológica, embora a estrutura sintáctica das frases possa ser bastante rígida e estabelecida à priori.

Um exemplo de utilização dos dois tipos de processamento morfológico poderá ser o da tradução automática.

A figura 1.4. representa um processo de tradução automática, em que a frase "O gatinho comeu os ratos." é traduzida para a frase inglesa "The little cat ate the mice."

*o gatinho comeu os ratos.*



<b>Análise Morfológica</b>	
<i>o</i>	["o", art, def, M, S]
<i>gatinho</i>	["gato", nome, M, S, Dim]
<i>comeu</i>	["comer", verbo, 3, S, pret. perf.]
<i>os</i>	["o", art, def, M, P]
<i>ratos</i>	["rato", nome, M, P]
<i>.</i>	[".", pont]



**Regras de tradução**

⇓

<b>Geração Morfológica</b>	
["the", art, def, S]	<i>the</i>
["little", adj, S]	<i>little</i>
["cat", nome, S]	<i>cat</i>
["eat", verbo, 3, S, simp. past]	<i>ate</i>
["the", art, def, P]	<i>the</i>
["mouse", nome, P]	<i>mice</i>
[".", pont]	.

⇓

*the little cat ate the mice.*

Figura 1.4. - Exemplo de análise e geração morfológica na tradução automática.

No presente trabalho estamos particularmente interessados na análise morfológica do português, motivados por duas razões. Primeiro, hoje em dia dá-se mais ênfase à obtenção automática de informação do que à geração automática de texto. A segunda razão está relacionada ao tipo de aplicação que iremos descrever: a correcção ortográfica.

## 8 Léxico, morfologia e dicionários

"Dicionário" e "léxico" têm significados equivalentes (Costa e Melo, 1989). Contudo, no seu uso

corrente, associa-se ao termo léxico todo o conjunto de palavras que compõem uma língua. Dicionário é identificado como uma colecção dos vocábulos de uma língua ordenados por ordem alfabética, com a respectiva descrição semântica e/ou morfossintáctica.

Sob o ponto de vista do processamento de linguagem natural, deve-se distinguir léxicos computacionais de dicionários electrónicos<sup>1</sup>.

Os dicionários electrónicos são realizações computacionais dos tradicionais dicionários de utilização geral, com a mais-valia da rapidez de acesso aos vocábulos e algumas ferramentas que auxiliem a sua consulta. Aurélio Electrónico (Ferreira, 1993; Ferreira, 1986) é um exemplo de dicionário electrónico interactivo de português, com facilidades de hipertexto. Assim, é possível ao utilizador consultar facilmente várias entradas que de alguma forma estejam relacionadas entre si.

Os léxicos computacionais são construídos com o objectivo de dar suporte a uma determinada aplicação específica de linguagem natural (Nijholt, 1992), ou determinado tipo aplicações. Por esta razão, as entradas dos léxicos computacionais geralmente não têm tanta informação como os dicionários electrónicos. Normalmente contêm apenas a informação mínima indispensável ao tipo de processamento a que se destinam.

Um dicionário electrónico, ou léxico computacional, não contém necessariamente toda a informação de forma explícita. Alguma parte da informação pode estar implicitamente representada sob a forma de regras.

Só em aplicações muito simplificadas de processamento de linguagem natural é que os seus dicionários contêm toda a informação representada de forma explícita, estática. Actualmente, é bastante comum o dicionário da aplicação ser constituído por uma componente de análise morfológica ligada a um banco de dados lexical, que contém informação relativa a apenas algumas palavras duma mesma família. A restante informação é deduzida através da análise morfológica e completada, ou confirmada, pela que se encontra no banco de dados lexical. Suponhamos, por exemplo, que pretendíamos obter informação lexicológica acerca da palavra *dormirei*. Esta começaria por ser analisada morfológicamente e o analisador morfológico concluiria que se tratava de uma forma do verbo *dormir*. Então, usaria *dormir* como chave de acesso ao banco de dados lexical, confirmando a sua existência e obtendo informação adicional relativa a este verbo.

Criar manualmente um léxico computacional para uma determinada aplicação é uma tarefa bastante dispendiosa e difícil de realizar. Hoje em dia, começam a aparecer sistemas que constroem automaticamente léxicos computacionais a partir de dicionários electrónicos e também a partir de texto corrente. Mais do que isso, há sistemas capazes de, automaticamente, extrair de dicionários electrónicos informação sintáctica e semântica, permitindo armazenar essa informação de forma estruturada e sistemática (Ageno et al. 1992; Dolan et al. 1993).

---

1 Geralmente denotados por MRD - *Machine Readable Dictionary*.

A partir de agora, usaremos sempre o termo "dicionário" para nos referirmos tanto a dicionários electrónicos como a léxicos computacionais, adoptando o termo completo apenas em caso de ambiguidade.

## 9 O Palavroso, um analisador morfológico

Quando uma pessoa é confrontada com a necessidade de dar uma descrição morfológica de uma palavra, a maior parte das vezes é capaz de o fazer, pelo menos parcialmente, mesmo que não conheça a palavra, quer recorrendo a ideias genéricas e regras empíricas, quer resolvendo o problema por analogia com situações idênticas. No português há várias regras de flexão e derivação que dizem muito acerca da palavra. Por exemplo, se uma palavra termina em *mente*, muito possivelmente tratar-se-á de um advérbio de modo. Geralmente os nomes e adjectivos terminados na letra *o* são masculinos, enquanto os femininos terminam na letra *a*. Se um nome termina em *s*, é provável que seja plural. Se terminar em *as*, tratar-se-á de um feminino plural. Se a terminação for *íamos*, *ávamos*, *asse*, ou *este*, o mais provável é tratar-se de uma forma de um verbo.

Por outro lado, quando se lida com irregularidades, o mais natural é tirar partido de semelhanças com casos conhecidos. A palavra *contenha*, por exemplo, por analogia com *tenha*, pode imediatamente ser identificada como uma forma do verbo *conter*. Exagerando um pouco esta ideia, se aparecesse a palavra *ugapongatenha*, embora não soubéssemos o seu significado, seríamos levados a dizer que se tratava de uma forma do verbo *ugapongater*, ou um adjectivo feminino como *nortenha*.

Um analisador morfológico automático terá que, na sua essência, funcionar com critérios semelhan-

tes.

Os processos mais comuns de fazer análise morfológica são a abordagem paradigmática, que funciona por analogia (Pentheroudakis & Higinbotham, 1991; Dumitrescu, 1992; Andrade et al., 1993; Lima e Kipper, 1993), e a abordagem por regras morfológicas de flexão e derivação, como em Koskenniemi (1983). A escolha de uma tendência ou outra está intimamente relacionada com a língua que se pretende descrever e com a representação que se tem, ou quer ter, do léxico, tal como acontece em Dumitrescu (1992) e Lima e Kipper (1993).

O Palavroso é um analisador morfológico que baseia o seu processamento fundamentalmente em regras. A escolha desta abordagem foi motivada pela sua primeira aplicação: gramática sem dicionário (Santos et al., 1992) e também pelo desejo de tratar o problema de palavras desconhecidas.

Em Santos et al. (1992) vem completamente descrita a arquitectura inicial do sistema. Entretanto, o Palavroso foi evoluindo em vários aspectos. O seu dicionário foi melhorado. Foi melhorada a sua funcionalidade. As regras foram aumentadas e depuradas. Em Medeiros et al. (1993) é encontra-se uma nova descrição do sistema conforme se encontrava naquela data, além de um relatório sobre o preenchimento lexical do Palavroso. Em Barreiro et al. (1993) estão completamente descritas todas as opções linguísticas tomadas no seu desenvolvimento.

Passamos, então, a descrever o analisador de acordo com a sua actual configuração e funcionalidade. Primeiro, como foi já referido, faremos uma descrição genérica. Depois, virá a descrição pormenorizada de cada uma das partes que compõem o sistema.

## 10 Arquitectura

Há duas características fundamentais do Palavroso que importa evidenciar:

1. Baseia o seu processamento fundamentalmente em regras.
2. Existe uma separação clara entre o léxico e as regras morfológicas.

A grande vantagem desta abordagem reside na possibilidade de ter um analisador morfológico que, mesmo que não tenha o léxico suficientemente preenchido, pode dar sempre algum tipo de resposta. No extremo, pode funcionar sem a parte lexical e, mesmo assim, fazer análises correctas.

Há formalismos para representar a morfologia de uma língua que, embora sejam bastante poderosos

e expressivos, falham neste aspecto. É o caso do formalismo de representação de dois níveis (Koskenniemi, 1983), em que as regras morfológicas e as entradas lexicais são colocadas em conjunto. Se a palavra em análise não estiver representada no léxico do analisador morfológico, então o sistema não será capaz de dar uma resposta.

Quando uma palavra é submetida a análise morfológica, são-lhe primeiro aplicadas as várias regras do Palavroso. Essas regras produzem análises hipotéticas que, depois, são verificadas no seu dicionário. Umas estarão correctas, outras não. Em função do modo de utilização, o analisador poderá dar como resposta todas as análises conseguidas ou apenas aquelas que forem confirmadas pelo dicionário.

Tomemos como exemplo a palavra *batas*. Ao ser submetida a análise morfológica, as regras morfológicas produziriam as seguintes hipóteses de análises:

Verbo *batar*, 2ª pessoa do singular do presente do indicativo

Verbo *bater*, 2ª pessoa do singular do presente do conjuntivo

Verbo *batir*, 2ª pessoa do singular do presente do conjuntivo

Nome *bata*, feminino, plural

Adjectivo *bato*, feminino, plural

Na segunda fase do processamento, o analisador verifica se as formas encontradas existem em dicionário. Neste exemplo, encontraria apenas *bater* e *bata* (fig. 2.1).

μ §

Figura 2.1 - Fases do processamento do Palavroso

As regras morfológicas foram agrupadas em quatro classes distintas, escolhidas principalmente em função da categoria gramatical a que se destinam. Esses grupos são:

1. Classes fechadas
2. Advérbios

3. Nomes e adjectivos
4. Verbos

De forma idêntica, o dicionário encontra-se dividido em quatro partes correspondentes. Diferem entre si pelo tipo de entradas que contêm, pela informação morfológica de cada entrada e pelo formato da informação.

Conceptualmente, os quatro tipos de regras são aplicados em paralelo, com acessos independentes ao dicionário (fig. 2.2). Na prática, trata-se de processamento sequencial e, como será visto mais adiante, existem algumas opções de ordenação que tornam o sistema mais eficiente.

μ §

Figura 2.2 - Divisão das regras e do dicionário do Palavroso.

Nas secções seguintes, iremos descrever pormenorizadamente cada uma das componentes do analisador morfológico Palavroso.

## 11 Comunicação com o exterior

A maneira como o analisador morfológico comunica com o exterior diz-nos muito a seu respeito. É a face visível do sistema. O esquema da figura 2.2 representa aquilo que podemos chamar o núcleo do analisador morfológico, por fazer a análise morfológica propriamente dita. É a essência do Palavroso.

Antes de uma palavra ser analisada, pode haver, e geralmente há, algum tipo de pré-processamento. De forma idêntica, o resultado do Palavroso pode estar sujeito a algum tipo de transformação: o pós-processamento.

Estas duas fases são da responsabilidade da aplicação que usa os serviços do Palavroso. A opção de não as considerar como parte integrante do analisador morfológico deve-se aos diferentes tipos de processamento que se efectuam nestas fases, em função das necessidades de cada aplicação.

Resumindo, o Palavroso não é uma aplicação autónoma. É um módulo que pode ser usado por outras aplicações que requeiram serviços lexicais e de análise morfológica. É da responsabilidade destas aplicações a preparação dos dados de entrada do Palavroso, assim como a manipulação dos

seus resultados.

Para facilitar a integração do Palavroso noutras aplicações, foram criadas algumas rotinas de pré e pós-processamento, tendo em vista determinados tipos de operação mais usuais, como a de verificar se uma palavra existe em dicionário.

## 12 Pré-processamento

O pré-processamento pode ser mais ou menos elaborado, dependendo da forma em que se encontrem os dados que vão ser objecto de análise. Se os dados consistirem em texto normal, por exemplo, é necessário separar as palavras dos sinais de pontuação. Se o texto estiver formatado, é necessário saber distinguir os comandos, ou códigos de formatação, para que o analisador morfológico receba apenas palavras válidas.

Uma fase de pré-processamento que quase sempre existe é a da normalização das palavras. Por essa razão foi construída uma rotina que o fizesse de forma genérica. A normalização consiste em dois processos: a passagem das letras maiúsculas a minúsculas e a conversão dos caracteres acentuados para o formato usado pelo Palavroso.

Quando se passa as letras para minúsculas, essa passagem é classificada com um código que caracteriza a palavra inicial. Desta forma, por um lado, a palavra respeitará o formato de entrada. Por outro, as suas particularidades serão levadas em consideração no processamento. Será tido em conta, por exemplo, que uma palavra com a primeira letra maiúscula poderá ser um nome próprio.

Há três situações possíveis para os casos em que a palavra aparece com letras maiúsculas:

1. Só a primeira letra é maiúscula
2. Todas as letras são maiúsculas
3. Há mistura de letras maiúsculas e minúsculas

Nos dois primeiros casos, as maiúsculas são passadas a minúsculas e é efectuada a caracterização do caso, registando-a. No terceiro, a palavra é passada à análise morfológica tal como se encontra, apenas se registando a caracterização deste caso.

A conversão dos caracteres acentuados é feita de acordo com uma tabela de conversão que o utilizador pode manipular (tabcars.tab). Desta forma, é possível fazer conversão de caracteres acentuados seja qual for a plataforma computacional usada. Em particular, na utilização do Palavroso em MS-DOS, este poderá manipular os acentos em mais do que uma página de código (codepage).

O formato interno dos caracteres acentuados é bastante simples. Cada carácter acentuado é decomposto na letra mais o acento, com este último a seguir à vogal (cf. quadro 2.3). O ç é decomposto em *c^*.

<b>Carácter acentuado</b>	à	á	é	í	ó	ú	â	ê	ô	ã	õ	ç
<b>Formato interno</b>	a`	a'	e'	i'	o'	u'	a^	e^	o^	a~	o~	c^

Quadro 2.3 - Representação interna dos caracteres acentuados. Para as maiúsculas a representação é idêntica.

A tabela de conversão manipulada pelo utilizador consiste no quadro 2.3 transposto. Cada linha tem uma conversão, em que o carácter acentuado aparece separado do formato interno por um espaço, ou mais. Por exemplo:

à      a`  
 á      a'  
 ã      a~

Outras tarefas do pré-processamento poderão ser a escolha do modo adequado de operação do Palavroso, ou a escolha de opções de funcionamento. Estas opções reflectir-se-ão no modo como o analisador funciona e, conseqüentemente, no tipo de resultados que fornecerá. Usando novamente o exemplo de verificar se uma palavra existe ou não no dicionário, verificamos que aqui não interessa fazer uma análise morfológica completa. Basta encontrar uma análise possível para se poder terminar o processamento morfológico com uma resposta positiva.

### 13 Pós-processamento

Na fase de pós-processamento, o resultado fornecido pelo analisador morfológico é manipulado de acordo com os interesses do utilizador. Este tanto pode querer uma descrição completa e verbalizada do resultado, como pode querer um simples código que de alguma forma caracterize a palavra analisada. O primeiro caso é característico de uma utilização interactiva do Palavroso. O segundo é característico de um processo de anotação automática de texto.

A anotação automática de texto não é efectuada sempre da mesma maneira. Uma vez pode-se querer anotar apenas a categoria gramatical das palavras. Outras vezes pretende-se uma anotação mais fina, com mais informação morfológica. Além disso, certos utilizadores gostam de etiquetas concisas e simples, outros preferem etiquetas mais palavrosas.

Dada a variedade de tipos de anotações possíveis, houve a preocupação de possibilitar ao utilizador a especificação das etiquetas a usar na anotação automática de texto. Além disso, estas etiquetas poderiam ser usadas na interface com outras aplicações.

Foi construído um módulo de pós-processamento que transforma o resultado do Palavroso em etiquetas, de acordo com uma especificação dada pelo utilizador. Essas etiquetas podem ser usadas para rotular texto, ou como dados de outras aplicações, como uma gramática, por exemplo. Em Fonseca (1993) vem descrita uma gramática, para uso restrito, que usa o Palavroso nesta base. Também em Medeiros (1992) vêm descritas aplicações que usam o Palavroso com esta interface.

O utilizador especifica de forma tabular as etiquetas que pretende usar. Para esse efeito usa duas tabelas (output.tab e codigos.tab) que constrói usando um vulgar editor de texto que manipule documentos em formato ASCII.

A tabela output.tab é a mais importante das duas, já que não é possível especificar as etiquetas sem ela. É composta por oito colunas. As primeiras sete estão relacionadas com a informação que a resposta do Palavroso contém, enquanto a oitava é usada para especificar o formato e o conteúdo da etiqueta. Nesta coluna é possível usar variáveis pré-definidas, referentes aos valores das primeiras sete colunas. A cada coluna está, portanto, associada uma variável.

As classificações do Palavroso que tenham a informação especificada nas primeiras sete colunas de uma determinada linha da tabela, são rotuladas com a etiqueta especificada na coluna oito.

O quadro 2.4 mostra, para cada coluna, a identificação da variável, a descrição da coluna e o que pode ser colocado na especificação. Em qualquer caso pode ser colocado o carácter '\*'. Desta forma, indica-se que a variável da coluna correspondente pode tomar qualquer valor.

Coluna	Variável	Descrição	Especificação possível
1	PAL	Palavra a ser analisada	<cadcars>, *<cadcars>, <cadcars>*
2	ORI	Origem da palavra. Por exemplo, no caso dos verbos tem o infinitivo impessoal.	Como na coluna 1
3	TEM	Tempo, no caso de se tratar de um verbo. Grau no caso de ser nome ou adjectivo. Ou outras características para outras classes gramaticais.	pi - Presente do indicativo pc - Presente do conjuntivo, etc. <sup>1</sup>
4	CLA	Classe gramatical	verbo, nome, adj, etc. <sup>2</sup>
5	NUM	Número	S - Singular P - Plural I - Invariável
6	PES	Pessoa, no caso de se tratar de um verbo	1 - Primeira 2 - Segunda 3 - Terceira
7	GEN	Género, nos casos em que tal tem sentido	M - Masculino F - Feminino I - Invariável

Quadro 2.4 - Descrição dos campos da tabela output.tab. <cadcars> representa uma cadeia de caracteres.

Quando se especifica a etiqueta, podem-se usar as variáveis para representar o conteúdo da coluna que lhe corresponde. A especificação da etiqueta vem separada das outras colunas pelo sinal ':', para melhor legibilidade. Na figura 2.5 mostram-se alguns exemplos de especificação de etiquetas.

Nesta figura, a segunda especificação aplica-se a qualquer classificação morfológica cuja classe gramatical seja verbo. A etiqueta resultante contém entre parêntesis o infinitivo impessoal do

1 Os códigos dos tempos dos verbos vêm no quadro B.5 do apêndice B. A identificação dos graus vem no quadro B.6 do mesmo apêndice.

2 As classes gramaticais que são reconhecidas pelo Palavroso vêm no quadro B.1 do apêndice B.

verbo. Por exemplo, a palavra *cantamos* será etiquetada com:

verbo(cantar)\_1P\_pi

PAL	ORI	TEM	CLA	NUM	PES	GEN	Etiqueta
*	*ter	*	verbo	*	*	*	: verboter(\$ORI)
*	*	*	verbo	*	*	*	: verbo(\$ORI)_\$PES\$NUM_\$STEM
*	*	Sup	adj	S	*	M	: adjsup(\$ORI)
*	*	*	nome	*	*	*	: nome(\$ORI)_\$STEM

Figura 2.5 - Exemplos de especificação de etiquetas. A identificação das colunas não faz parte duma tabela normal, aparecendo aqui por uma questão de legibilidade.

A primeira especificação é idêntica, aplicando-se a todas as formas de verbos cujo infinitivo impessoal termina em *ter*. Por exemplo, *conter* ficará com etiqueta *verboter(conter)*.

A terceira especificação aplica-se somente a adjectivos que sejam masculinos singulares e se encontrem no grau superlativo. Por exemplo, *belíssimo* ficaria com a etiqueta *adjsup(belo)*.

A última especificação aplica-se a todas as classificações de nome. Por exemplo, *mesinha* origina a etiqueta *nome(mesa)\_Dim*.

No caso de várias especificações se aplicarem à mesma classificação morfológica, só a primeira é considerada.

A outra tabela (*codigos.tab*), que complementa a funcionalidade desta, aplica-se para classificar uma palavra no seu conjunto. Por exemplo, a palavra *canto* é classificada pelo Palavroso como nome e como verbo. Ao aplicar as especificações da tabela *output.tab*, serão obtidas duas etiquetas, uma para cada classificação. A tabela *codigos.tab* permite conciliar as duas classificações numa só.

Vamos supor que as etiquetas resultantes de *output.tab* só têm informação relativa à classe gramatical. Então, a palavra *canto* originaria, por exemplo, as etiquetas *nome verbo*. Através da tabela *codigos.tab* seria possível converter este resultado em *nomverb*, ou outro rótulo parecido.

A sintaxe desta tabela é extremamente simples. Cada linha tem uma conversão dividida em duas partes. Na primeira, aparecem as etiquetas que se esperam como resultado de *output.tab*, enquanto que na segunda aparece a etiqueta que deve ser dada como resultado. Na primeira, pode aparecer

um asterisco a substituir uma etiqueta, dizendo que aceita qualquer coisa.

Exemplos de especificações de conversão são:

nome verbo : nomverb  
verbo verbo : verbo  
nome adj : nomadj  
verbo \* : verbamb  
\* : \*

As três primeiras especificações têm significados óbvios. A quarta especificação aplica-se a todas as situações em que apareça pelo menos uma classificação de verbo. Por exemplo, verbo adj seria convertido em verbamb.

Este tipo de especificação pode ser alargado a mais etiquetas. Por exemplo, poderíamos ter

nome verbo \* : nomverbamb

Neste caso, a especificação aplicar-se-ia a qualquer resultado da classificação do Palavroso que tivesse pelo menos uma classificação de nome e uma classificação de verbo, podendo aparecer quaisquer outras classificações. Esta especificação é menos restritiva que a primeira do exemplo, uma vez que aquela só se aplica a classificações consistindo exactamente num nome e num verbo.

A última especificação indica, simplesmente, que tudo o que aparecer deve ser dado como resultado, sem sofrer qualquer alteração.

Os verbos com enclíticos representam uma situação específica que merece destaque nesta descrição. Nestes casos, o Palavroso dá uma, ou mais, classificações ao verbo, seguindo-se uma lista de classificações relativa aos enclíticos, fazendo corresponder uma classificação por cada. Desta forma, as especificações aplicam-se como se se tratasse de uma palavra simples, com classificação ambígua. Por exemplo, se a especificação fosse

verbo clítico : verbclit

aplicar-se-ia a todos os verbos com apenas um enclítico. Por exemplo, *chama-se*.

Tal como nas especificações da tabela anterior (output.tab), é usada a primeira especificação que se aplica a uma determinada lista de classificações. A ordem por que aparecem as etiquetas dentro da mesma especificação não é importante.

É possível encontrar outros exemplos da utilização das duas tabelas e seu processamento em Santos et al. (1992).

## 14 Modos de operação

O Palavroso pode funcionar sob várias condições, a que chamamos modos de operação. Pode fornecer, por exemplo, todo o tipo de classificações que consegue fazer, ou dar apenas classificações que tenham sido confirmadas por acesso ao dicionário.

Estes modos de operação são estabelecidos activando ou desactivando certas opções. Convém fazer aqui a distinção entre as opções do Palavroso e as opções da aplicação que usa o Palavroso. São duas coisas distintas. Por exemplo, quando o Palavroso é usado em correcção ortográfica, o utilizador não tem controlo sobre o seu modo de operação. É o próprio corrector que estabelece as opções automaticamente, de acordo com o processamento adequado.

Existem seis opções que passamos a descrever. Para isso, vamos socorrer-nos da interface dada por uma aplicação que permite ao utilizador trabalhar interactivamente com o Palavroso. Para informação mais detalhada sobre esta aplicação consultar Medeiros (1992).

No seu modo normal de funcionamento, se nenhuma opção tiver sido activada, o Palavroso dá os resultados da seguinte forma<sup>1</sup>:

*bonezinho*

bonezinho:

nome, masculino, singular, Diminutivo (bone', 100%)

verbo bonezinhar, 1. pessoa do singular do Presente Ind. (0%)

Este exemplo, embora pequeno, apresenta duas particularidades. A primeira é o acento em *boné*, que aparece no formato internamente usado pelo Palavroso. A segunda particularidade é o aparecimento de uma classificação morfológica inesperada. Contudo, repare-se que na primeira classificação aparece o valor 100%, enquanto na segunda o valor homólogo é 0%. Na actual configuração do sistema, isso quer dizer que a primeira classificação foi confirmada pela informação disponível no dicionário do Palavroso, enquanto na segunda classificação tal não aconteceu. É, portanto, apenas fruto da aplicação das regras do Palavroso.

## 15 Normalização e acentuação

A primeira particularidade motiva a existência das opções de acentuação e normalização, a que chamamos **opção a**. Esta última serve apenas para o Palavroso saber se está a lidar com uma palavra que já foi normalizada ou não. No caso de não estar normalizada poderá ter que efectuar essa normalização internamente.

---

1 As palavras submetidas a análise aparecem em itálico e a resposta do Palavroso aparece em texto normal.

A opção de acentuação, quando activada, faz com que as palavras acentuadas não venham no formato interno do Palavroso, mas no formato usado no exterior, como no exemplo que se segue.

*comunicações*  
comunicações:  
nome, feminino, plural (comunicação, 100%)  
adjectivo, feminino, plural (comunicação, 0%)

## 16 Quantidade de informação

A segunda particularidade chama a atenção para a quantidade de informação que o Palavroso pode dar como resposta. Se nada for especificado, o Palavroso dá o maior número possível de classificações. Há, contudo, situações em tal não é desejável. Para controlar este tipo de resposta existem duas opções, que identificamos por **opção u** e **opção v**. A ideia genérica das duas é a mesma, só que a **opção v** é mais restritiva.

Se alguma destas opções estiver activada, o Palavroso dá como resposta apenas as classificações que são confirmadas pelo dicionário. Por exemplo, funcionando em modo normal, a palavra *bata* dá origem à classificação:

*batas*  
batas:  
nome, feminino, plural (bata, 100%)  
verbo bater, 2. pessoa do singular do Presente Conj. (100%)  
adjectivo, feminino, plural (bato, 0%)  
verbo batar, 2. pessoa do singular do Presente Ind. (0%)  
verbo batir, 2. pessoa do singular do Presente Conj. (0%)

Se uma das duas opções estivesse activada, o resultado do Palavroso, para a mesma palavra, seria:

*batas*  
batas:  
nome, feminino, plural (bata, 100%)  
verbo bater, 2. pessoa do singular do Presente Conj. (100%)

A diferença entre as duas opções reside no comportamento do Palavroso, quando nenhuma das classificações produzidas pelas regras é confirmada pela informação existente no seu dicionário. Imaginemos que pretendíamos analisar a palavra *escloras*. Se estivesse activada a **opção u**, o resultado seria:

*escloras*

escloras:

nome, genero indeterminado, plural (esclora, 0%)

adjectivo, feminino, plural (escloro, 0%)

adjectivo, feminino, plural (esclor, 0%)

verbo esclorar, 2. pessoa do singular do Presente Ind. (0%)

verbo esclorir, 2. pessoa do singular do Presente Conj. (0%)

Se, para a mesma palavra, estivesse activada a **opção v**, o resultado seria nulo:

*escloras*

escloras: NULO

Resumindo, com a **opção u**, o analisador morfológico dá sempre alguma resposta. Se as classificações forem confirmadas pela informação do dicionário, fornece apenas estas. Se nenhuma tiver sido validada pelo dicionário, dá todas as aproximações que conseguir calcular.

Com a **opção v**, o Palavroso só dá classificações que sejam confirmadas pelo dicionário. Se nenhuma o for não dá qualquer resposta.

Esta opção é útil quando se usa o Palavroso como dicionário de outra aplicação que apenas necessita saber se uma palavra existe ou não no dicionário (cf. secção 2.3.3).

## 17 Utilização como dicionário booleano

Quando se usa um dicionário apenas para saber se uma determinada palavra existe ou não, não é necessário que o analisador morfológico dê todas as classificações morfológicas possíveis. Basta encontrar uma classificação confirmada pela informação existente em dicionário, para de imediato concluir que essa palavra existe. Nessa situação, o processo de análise morfológica pode ser interrompido, e é conveniente fazê-lo, por razões de eficiência.

Para funcionar neste modo de processamento é necessário estarem activadas duas opções: a **opção d**, para o processamento parar assim que seja encontrada uma análise válida; a **opção v**, para que sejam dadas apenas classificações que sejam confirmadas pelo dicionário.

Desta forma, pode-se economizar bastante tempo de processamento, principalmente se a maioria das palavras analisadas existir no dicionário. Se tal não acontecer, será sempre necessário calcular todas as classificações possíveis, para poder concluir que uma palavra não existe no dicionário.

## 18 Depuração

A *opção I*, uma vez activada, faz com que o Palavroso escreva num ficheiro (palav.log) uma descrição de todas as operações efectuadas. Desta forma, é possível analisar todo o processamento efectuado à posteriori e, eventualmente, fazer correcções às regras do Palavroso.

Embora esta opção possa ser activada a partir de qualquer aplicação que use o Palavroso, o normal é isso ser feito apenas em aplicações interactivas, usadas no estudo, desenvolvimento e teste do Palavroso.

Eventualmente pode ser usado para auxiliar a compreensão dos processos internos do Palavroso.

## 19 As componentes

Para além dos módulos de pré e pós-processamento acima descritos, que na realidade não fazem parte do Palavroso propriamente dito, podemos enumerar seis componentes fundamentais:

1. Componente responsável pelas palavras fechadas<sup>1</sup>
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 duas últimas componentes representam casos especiais, que poderiam ser entendidos como casos de pré-processamento interno do Palavroso. Ambas aplicam apenas algumas regras específicas, acabando por usar um ou vários dos módulos restantes.

As várias componentes são processadas na sequência que se segue:

0. Iniciar variáveis.
1. Verificar se a palavra contém algum hífen. Se contiver, será efectuado o processamento

<sup>1</sup> Neste texto, por palavras fechadas referimo-nos a palavras pertencentes a classes fechadas, isto é, artigos, preposições, conjunções, advérbios (não terminados em *mente*), interjeições e pronomes.

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 por 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 de advérbios de modo.
6. Ordenar os resultados e formatá-los de acordo com a palavra analisada. Por exemplo, passar os resultados a maiúsculas e colocar os acentos como na palavra submetida à análise.

O programa vai processando sucessivamente cada um dos módulos, acrescentando todas as classificações morfológicas possíveis a uma lista de classificações, que será o produto final do analisador morfológico.

Dedicaremos as secções seguintes à descrição mais pormenorizada de cada um destes passos.

## 20 Palavras fechadas

Esta componente é bastante simples. Consiste em pouco mais do que verificar se a palavra em análise faz parte de uma lista de palavras fechadas.

Existem palavras pertencentes a classes fechadas que são homógrafas com outras pertencentes a classes abertas. Por exemplo, *como*, pertence a uma classe fechada (conjunção), mas também pode ser uma forma do verbo *comer*. Assim, dividimos a lista de palavras fechadas em duas: uma contém palavras que pertencem apenas a classes fechadas, como *a*, *com*, *comigo*, *depressa*, *ele*, *sim*, *vosso*. A outra contém palavras que são homógrafas com outras pertencentes a classes abertas, como *consoante*, *sua*, *quer*.

O motivo desta divisão está relacionado com o tipo de processamento necessário num e noutro caso. Se a palavra em análise for encontrada na lista das palavras pertencentes apenas a classes fechadas, não será necessário prosseguir a análise, pois temos a certeza de ter encontrado todas as classificações possíveis. Se for encontrada na outra lista, apenas acrescenta a sua classificação morfológica à lista de classificações possíveis, prosseguindo o processamento normal.

Neste módulo, para além das palavras fechadas, são verificados outros tipos de palavras. Referimo-

nos a numerais, abreviaturas, nomes próprios, siglas e acrónimos (se a palavra tiver maiúsculas).

É discutível se os nomes próprios, siglas e acrónimos devem fazer parte do léxico de um analisador morfológico. Por um lado, são tantos e tão variados que é difícil reunir uma lista suficientemente representativa. Por outro lado, não é invulgar os dicionários, gramáticas e prontuários, trazerem listas de siglas e abreviaturas mais conhecidas e usadas, adjectivos gentílicos, nomes próprios e nomes geográficos (Costa e Melo, 1989; Vilela, 1990).

Como a palavra a ser analisada foi previamente normalizada, ao verificar se se encontra presente na lista de nomes próprios e siglas, é temporariamente transformada em maiúsculas (toda, ou apenas a primeira letra, conforme se encontre a palavra antes do processo de normalização).

Se a palavra for um nome próprio, sigla ou abreviatura, a análise morfológica parará. Nestes casos, considera-se que não pode ter outra classificação morfológica.

Os numerais ordinais e os múltiplos flexionam quanto ao número e quanto ao género. Por isso, como apenas se encontram listados os masculinos singulares, por vezes é necessário detectar as terminações características da flexão do número e do género no numeral, construir o correspondente masculino singular e usar esta forma para procurar na lista dos numerais.

O léxico processado no módulo das palavras fechadas, mais a sua informação morfológica, é representado sob forma tabular, em que cada linha da tabela contém informação relativa a uma palavra e cada coluna corresponde a um tipo de informação.

Estas tabelas têm seis colunas com o seguinte significado:

<b>Coluna</b>	<b>Descrição</b>
1	Entrada lexical
2	Número
3	Género
4	Classe gramatical
5	Subclasse gramatical

A subclasse gramatical serve para fazer algumas distinções entre palavras da mesma classe. Por exemplo, *terceiro* pertence à classe dos numerais, subclasse dos números ordinais; *sua* pertence à classe dos pronomes possessivos, terceira pessoa. Este campo nem sempre é necessário e nem sempre é usado. Nos casos em que não existe uma subclasse, este campo é preenchido com a classe gramatical.

O campo relativo à probabilidade, por agora, é preenchido invariavelmente com o valor 1.0. Destina-se, futuramente, a conter informação probabilística sobre as entradas lexicais. Isso poderá permitir ordenar as classificações do Palavroso por ordem decrescente de probabilidade, ou ainda, organizar a informação segundo uma motivação estatística, com o objectivo de melhorar os tempos de acesso à informação, conforme descrito em Medeiros (1994).

Alguns exemplos de entradas deste tipo de tabelas são:

<b>Entrada</b>	<b>Número</b>	<b>Género</b>	<b>Classe</b>	<b>Subclasse</b>	<b>Probabilidade</b>
a	S	F	art	def	1.0
cerca	I	I	adv	adv	1.0
sua	S	F	poss	3	1.0
terceiro	S	M	num	ord	1.0

O quadro B.1 do apêndice B contém todas as classes e subclasses correntemente usadas no Palavroso.



matemática).

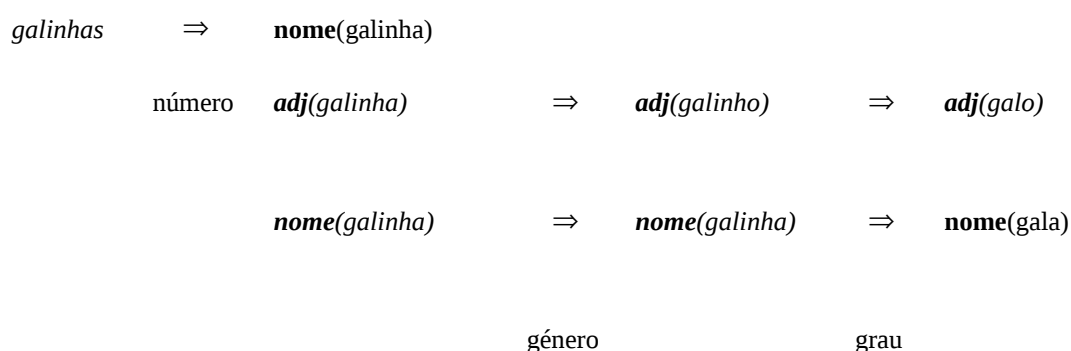
Este facto tem duas implicações a nível de processamento:

1. É necessário em cada fase do processamento verificar se as origens encontradas até ao momento existem no dicionário.
2. Por cada origem encontrada no dicionário, deve ser criada uma imagem desta origem, para prosseguir a análise.

Há, ainda, um problema motivado pelo facto de estarmos a lidar com duas classes gramaticais diferentes: o processamento relativo ao género é diferente no caso da palavra ser um nome ou ser um adjectivo. No caso dos nomes, as regras limitam-se a dizer qual o género da palavra. No caso dos adjectivos, se é encontrada uma forma feminina, esta é transformada no masculino correspondente.

Esta separação de processamento é feita porque, geralmente, os adjectivos têm os dois géneros, verificando-se que a passagem de um para outro é relativamente regular. No caso dos nomes, na maior parte dos casos não existem os dois géneros (vd. nota 6, secção 1.2). Na verdade, verifica-se que só os nomes que designam seres animados flexionam no género (Figueiredo e Ferreira, 1974). No Palavroso optou-se por considerar que o género é uma propriedade intrínseca dos nomes, não sujeita a flexão (cf. Barreiro et al., 1993).

Tomando novamente como exemplo a palavra *galinhas*, obtemos



O processamento relativo ao número é igual para nomes e adjectivos, daí só ser criado o ramo para adjectivo após o processamento do número. Na fase das regras relativas ao número verifica-se o resultado no dicionário. Como o nome *galinha* é encontrado, este ramo pára e é criada a sua imagem para prosseguir o processamento, que acaba por dar mais uma origem (*gala*). O ramo de adjectivo é analisado nas fases seguintes, mas não produz qualquer origem dicionarizada.

No caso de *matemática*, o processamento é idêntico. Contudo, no final, uma das origens é eliminada, pois não produziu nada de novo.

*matemática*      ⇒      **nome**(matemática)

número    **adj**(matemática)                      ⇒      **adj**(matemático)

**nome**(matemática)                      ⇒      **nome**(matemática)      eliminada

gênero

## 22 Processamento quanto ao número

No que diz respeito ao número, existe um conjunto de regras genéricas, responsável pelo processamento da maioria dos casos, e existe um conjunto de regras específicas para o tratamento de plurais de palavras terminadas em *ão*.

Para estas palavras, a regra mais produtiva de formação do plural transforma *ão* em *ões*. Mas, existem outras regras igualmente válidas, embora não tão produtivas. Por exemplo, transformar *ão* em *ãos* ou *ães* (*mão* -> *mãos*; *pão* -> *pães*). Considerámos, portanto, a conversão de *ão* em *ões* a regra geral, tratando os outros casos como exceções.

Assim, se uma palavra termina em *ões*, *ães* ou *ãos*, consulta-se uma tabela de exceções que associa a cada palavra o seu (ou seus, no caso de admitir mais do que um) plural. Se a palavra não se encontrar nesta tabela e se terminar em *ões*, então a terminação é modificada para *ão*. Senão (se não existir na tabela e terminar em *ães* ou *ãos*), considera-se que se trata de uma transformação pertencente ao caso geral e a terminação não é modificada para *ão*.

O quadro 2.7 contém um exemplo de entradas desta tabela.

Plural	Singular
tabeliães	tabelião
verãos	verão

verões          verão

---

Quadro 2.7 - Exemplo de entradas de uma tabela de exceções de pluralização de nomes e adjectivos terminados em *ão*.

Se aparecessem as palavras *verões* ou *verãos*, saberíamos que se tratavam de duas formas do plural de *verão*. De forma idêntica, *tabeliães* seria aceite como sendo o plural de *tabelião*. No entanto, *tabeliões* ou *tabeliãos*, nunca poderiam ser aceites como formas plurais de *tabelião*, pois não existe nenhuma entrada que o estabeleça.

O caso geral do processamento quanto ao número é baseado numa tabela que contém terminações características de plurais associadas à respectiva terminação singular, de acordo com o exemplo do quadro 2.8. Se uma palavra termina de acordo com alguma das entradas da tabela, então é reescrita substituindo a terminação plural pela singular.

Plural	Singular	Número
as	a	P
os	o	P
is	i	P
teis	til	P

---

Quadro 2.8 - Exemplo de entradas de regras de pluralização de nomes e adjectivos.

A primeira coluna contém as terminações do plural e a segunda coluna a terminação singular associada à primeira. A terceira coluna contém um parâmetro que classifica o tipo de entrada. Se estivermos a fazer análise morfológica sem dicionário, pode interessar ter entradas que simplesmente classifiquem o tipo de terminação quanto ao número. Por exemplo, poder-se-ia ter a entrada:

para classificar palavras terminadas em *is* (*lápiz*) como invariáveis.

A terceira e quarta entradas podem aplicar-se à mesma palavra. Por exemplo, consultando esta tabela com a palavra *úteis*, são produzidas duas origens possíveis: *útei* (terceira regra) e *útil* (quarta regra).

Após criadas todas as origens possíveis, consulta-se o dicionário dos nomes para verificar se alguma corresponde a uma palavra do léxico.

## 23 Processamento quanto ao género

Tal como no número, as palavras terminadas em *ão* têm um tratamento especial no que diz respeito ao género, embora mais simplificado.

Existe uma lista de palavras com o feminino de adjetivos terminados em *ão*. Se a palavra em análise se encontrar nessa lista é transformada no masculino, passando a sua terminação para *ão*. Se a palavra não se encontrar na lista é tratada como qualquer outra palavra.

O caso geral de processamento quanto ao género é baseado numa tabela que contém pares de terminações, associando uma terminação característica de palavras femininas à correspondente terminação masculina. O quadro 2.9 apresenta alguns exemplos que podem ser encontrados numa tabela deste tipo.

Na primeira coluna, encontram-se as terminações características do masculino e, na segunda coluna, as terminações femininas associadas às primeiras. Assim, seguindo as regras do quadro 3.9, uma palavra terminada em *a* será considerada o feminino de uma palavra terminada em *o* (terceira regra). Por exemplo, *amarela/amarelo*.

Masculino	Feminino
or	ora
o	a
.	oa

.	ã
eiro	.
ista	ista

---

Quadro 2.9 - Exemplos de regras do género.

As entradas em que numa das colunas aparece um ponto servem unicamente para identificar terminações características de um dos géneros. Por exemplo, a terminação *ã* é característica de palavras femininas (*capitã; irmã; beirã*), enquanto que *eiro* é terminação característica de palavras masculinas (*barqueiro; carteiro*).

Quando aparecem as mesmas terminações em ambas as colunas, isto significa que a palavra é invariável quanto ao género. Por exemplo, *o xadrezista vs. a xadrezista; o alpinista vs. a alpinista*.

As regras em que se modifica a terminação só se aplicam aos adjetivos. As outras aplicam-se aos nomes exclusivamente para os qualificar quanto ao género. Este processamento só interessa no caso dos aumentativos e diminutivos ou quando os nomes não se encontram no dicionário.

## 24 Processamento quanto ao grau

O processamento quanto ao grau é baseado fundamentalmente em regras. Para cobrir determinados casos em que o radical do nome ou adjetivo é alterado devido à inclusão de informação relativa ao grau, existe uma tabela de exceções que relaciona as palavras com a sua flexão num determinado grau. O quadro 2.10 mostra alguns exemplos que podem aparecer numa tabela deste tipo.

A terceira coluna do quadro especifica qual o tipo de grau a que se refere a relação. Neste exemplo, três das entradas referem-se a exceções de superlativos e a outra é uma exceção de diminutivo.

<b>Forma flexionada</b>	<b>Palavra</b>	<b>Grau</b>
amicíssimo	amigo	Sup
bonzinho	bom	Dim
ótimo	bom	Sup
paupérrimo	pobre	Sup

Quadro 2.10 - Exemplos de exceções dos graus dos nomes e adjetivos

Se uma palavra não for encontrada na tabela de exceções, então aplicam-se-lhe as regras normais de processamento dos graus dos nomes e adjetivos. Essas regras têm um formato idêntico à tabela de exceções, com a diferença de se aplicarem a terminações e não a palavras. Também lhe foram acrescentados dois campos para acrescentar alguma informação relativa à acentuação da palavra, já que acontece com bastante frequência o acento tónico da palavra mudar com a flexão. Por exemplo, *café* vs. *cafezinho*.

<b>Terminação da flexão</b>	<b>Terminação da palavra</b>	<b>Grau</b>	<b>Verificar acento</b>	<b>Acentuar</b>
-----------------------------	------------------------------	-------------	-------------------------	-----------------

inha	a	Dim	1	1
inho	o	Dim	1	1
ona	a	Aum	0	1
rzinho	r	Dim	0	1

Quadro 2.11 - Exemplo de regras relativas ao grau dos nomes e adjectivos

De acordo com a primeira regra, se uma palavra terminar em *inha*, então trata-se de um diminutivo de uma palavra que se obtém substituindo *inha* por *a*. A penúltima coluna indica que a regra não pode ser aplicada a palavras que tenham acentos gráficos; a última coluna indica que se a mudança de terminações não conduzir a uma origem existente no dicionário, então deve experimentar-se colocar acentos gráficos. Por exemplo, a palavra *monogaminho* é considerada, pela segunda regra, o diminutivo de *monógamo*. Repare-se que a aplicação simples da primeira parte da regra conduziria a diminutivo de *monogamo*. Como tem a opção de acentuar, experimenta colocar criteriosamente acentos agudos e circunflexos, até encontrar uma palavra que exista no dicionário.

Há situações em que o grau acrescenta à palavra informação relativa ao género. Por exemplo, as palavras *inteligente* e *azul* são invariáveis quanto ao género. No entanto, *intelligentíssima* é um adjectivo feminino e *azulinho* é um nome ou adjectivo masculino. Nestas situações, as regras de flexão têm preferência relativamente à informação que existe no dicionário. Compare-se os seguintes resultados obtidos com o Palavroso:

*azul*

azul:

adjectivo, genero indeterminado, singular (azul, 100%)  
nome, masculino, singular (azul, 100%)

*azulinho*

azulinho:

adjectivo, masculino, singular, diminutivo (azul, 100%)  
nome, masculino, singular, diminutivo (azul, 100%)

*azulinha*

azulinha:

adjectivo, feminino, singular, diminutivo (azul, 100%)

Como adjectivo, *azul* é invariável quanto ao género, mas *azulinho* é masculino. A diferença entre *azulinha* e *azulinho* deve-se ao facto das regras de flexão quanto ao género só se

aplicarem aos adjectivos.

## 25 Dicionário

O dicionário dos nomes simples é composto por cinco campos. Para além da probabilidade e da palavra propriamente dita, aparecem três campos que descrevem morfologicamente a palavra: um para o número, outro para o género e outro para a classe gramatical. As etiquetas usadas nos campos do género e do número são já conhecidas. Das usadas no campo relativo à classe gramatical só deixa dúvidas nomadj, que classifica uma palavra como podendo ser um nome ou um adjectivo.

O quadro 2.12 contém alguns exemplos de entradas deste dicionário.

<b>Palavra</b>	<b>Número</b>	<b>Género</b>	<b>Classe</b>	<b>Probabilidade</b>
azul	S	I	adj	1.0
azul	M	S	nome	1.0
matemática	S	F	nome	1.0
matemático	S	M	nomadj	1.0

Quadro 2.12 - Extracto do dicionário dos nomes.

## 26 Verbos

Há vários tipos de regras para processar as flexões dos verbos. Uma são produtoras, enquanto que outras são restritivas. As regras produtoras são aquelas que criam classificações morfológicas possíveis para a palavra que se encontra a ser analisada. As regras restritivas são responsáveis pela exclusão de algumas classificações que à partida se sabe não serem confirmadas pela informação do dicionário.

Dentro das regras produtoras, poderemos ainda considerar um subgrupo a que chamaríamos regras de transformação, pois o seu objectivo é apenas modificar determinada informação, sem produzir novas classificações, nem eliminar as já existentes.

## 27 Regras produtoras

São as regras que dão a primeira informação acerca da classificação das palavras, no caso de serem potenciais formas de verbos. Há regras que descrevem a flexão dos verbos regulares e existe

uma lista de formas irregulares com a respectiva classificação morfológica.

Toda a informação se encontra sob a forma tabular, embora com diferentes campos e diferentes interpretações.

A lista das formas irregulares é composta por entradas com seis campos. O quadro 2.13 descreve e dá exemplos de entradas desta lista.

O primeiro campo (Forma) contém a forma verbal. Os quatro campos seguintes contêm a informação morfológica que lhe está associada e o último campo é reservado para o processamento probabilístico. Há situações em que não faz sentido falar em pessoa e número, como é o caso do gerúndio, por exemplo. Nestes casos, os respectivos campos são preenchidos com '.' (ponto). As abreviaturas dos tempos encontram-se no quadro B.5 do Apêndice B.

Se uma palavra coincidir com alguma daquelas formas, é-lhe atribuída a classificação morfológica que se encontra na linha correspondente, continuando o processamento na procura de outras regras aplicáveis à palavra em análise.

Forma	Pessoa	Número	Tempo	Verbo	Probabilidade
dão	3	P	pi	dar	1.0
leio	1	S	pi	ler	1.0
*feito	.	.	p	*fazer	1.0
*feita	.	.	p	*fazer	1.0
*eio	1	S	pi	*ear	1.0
*eio	1	S	pi	*iar	1.0

Quadro 2.13 - Exemplos de entradas na lista de formas irregulares de verbos.

Há que salientar a utilização de terminações como entradas desta tabela. São as formas que vêm com um asterisco no início. Assim, *\*feito* aplica-se tanto a *feito* como a *refeito*, ficando com a

classificação de particípio passado dos verbos *fazer* e *refazer*, respectivamente. Palavras terminadas em *eo* são potenciais formas de verbos terminados em *ear* ou *iar*. Por exemplo, *ceio* é uma forma do verbo *cear*; *odeio* é uma forma do verbo *odiar*.

Enquanto as formas que aparecem completas não carecem de qualquer verificação no dicionário, as entradas que aparecem com terminações necessitam que se verifique se a regra foi correctamente aplicada, pois, eventualmente, as regras podem originar verbos não existentes. Do processo de verificação, aproveitam-se para criar classificações morfológicas apenas as entradas que originaram verbos existentes no dicionário. Por exemplo, a forma verbal *leio*, é classificada imediatamente como forma do verbo *ler*. No entanto, também se lhe aplicam duas terminações, interpretando-a como forma do verbo *lear* e como forma do verbo *liar*. Verificando estes dois verbos no dicionário, constata-se que *liar* existe mas *lear* não existe. Como resultado final, são acrescentadas duas classificações morfológicas à lista de classificações. Uma correspondente ao verbo *ler* e outra correspondente ao verbo *liar*.

No entanto, a maior fonte de classificações no campo verbal são as regras dos verbos regulares.

Estas regras procuram detectar terminações características de flexões de verbos. Em vez de procurar a desinência e a característica da forma verbal separadamente, procura padrões do tipo **característica + desinência pessoal**. Se alguma palavra analisada tiver uma terminação que se ajuste a alguma regra, é acrescentada uma classificação à lista de classificações, de acordo com a informação morfológica associada a essa regra.

As regras dos verbos regulares têm treze campos (colunas), que se dividem em quatro partes com significados distintos, como ilustrado no quadro 2.14.

Padrão				Informação morfológica				Aplicabilidade				Probabilidade
ar	er	ir	or	Terminação	Tempo	Pessoa	Número	ar	er	ir	or	Probabilidade
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



## 28 Regras de transformação

As regras de transformação são usadas para fazer alterações aos radicais dos verbos. Existem dois tipos de regras de transformação: umas são responsáveis por determinadas conversões regulares, relacionadas com regras ortográficas; outras, mais complexas, são usadas para analisar formas de verbos irregulares a partir das regras dos verbos regulares.

As conversões regulares aplicam-se a casos bem conhecidos, que envolvem modificações ortográficas previsíveis, correspondendo sempre a casos em que diferentes letras gráficas correspondem ao mesmo som. Por exemplo, dada a forma *embarquemos*, as regras dos verbos regulares classificá-la-iam como primeira pessoa do plural do presente do conjuntivo do verbo *embarquar*. Neste caso, *qu* é convertido em *c* para dar o infinitivo ortograficamente correcto. Outro exemplo poderia ser o verbo *caçar*, que no presente do conjuntivo tem o seu radical transformado em *cac* (*cace, caces, cacemos, caceis, cácem*).

Para mais pormenores sobre este conjunto de regras, consultar Santos et al. (1992), secção 4.5.12.

Após alguma análise, verificámos que os verbos irregulares afinal o não são tanto como pareciam. Embora haja irregularidades absolutas, como as do verbo *ser*: *sou, és, é, fui, foste*, etc., a maioria dos verbos irregulares apresentam alterações nos radicais em certos tempos, mas a flexão das formas dentro desse tempo é perfeitamente regular. Tomemos como exemplo o verbo *estar* e observemos a sua conjugação em vários tempos.

Tempos	Presente Indicativo	Imperfeito Indicativo	Presente Conjuntivo	Imperativo
derivados do Presente do Indicativo	estou estás está estamos estais estão	estava estavas estava estávamos estáveis estavam	esteja estejas esteja estejamos estejais estejam	está estai

Tempos	<b>Pretérito Perfeito</b>	<b>Mais que Perfeito</b>	<b>Imperfeito Conjuntivo</b>	<b>Futuro Conjuntivo</b>
derivados do Pretérito Perfeito	estive estiveste esteve estivemos estivestes estiveram	estivera estiveras estivera estivéramos estivéreis estiveram	estivesse estivesses estivesse estivéssemos estivésseis estivessem	estiver estiveres estiver estivermos estiverdes estiverem

Tempos	<b>Futuro Indicativo</b>	<b>Condicional</b>	<b>Infinitivo Pessoal</b>	<b>Particípio</b>
derivados do Infinitivo Impessoal	estarei estarás estará estaremos estareis estarão	estaria estarias estaria estariamos estarieis estariam	estar estares estar estarmos estardes estarem	estado <hr/> <b>Gerúndio</b> estando <b>Infinitivo impessoal</b> estar <hr/>

Nos tempos derivados do presente existem algumas formas completamente irregulares (*estou, estás, está*). O imperfeito do indicativo é completamente regular e, se o verbo tivesse radical *estej* (*estejer* ou *estejir*), o presente do conjuntivo também seria completamente regular.

Nos tempos derivados do pretérito perfeito, exceptuando duas formas deste mesmo tempo (*estive, esteve*), a conjugação seria regular se se tratasse de um verbo da segunda conjugação (*estiver*).

Nos tempos derivados do infinitivo impessoal, todas as formas obedecem a um padrão de verbo regular.

Vejam os que acontece com o verbo *fazer*:

Tempos	<b>Presente Indicativo</b>	<b>Imperfeito Indicativo</b>	<b>Presente Conjuntivo</b>	<b>Imperativo</b>
do Presente Indicativo	faço fazes faz	fazia fazias fazia	faça faças faça	faz fazei

Tempos	<b>Pretérito Perfeito</b>	<b>Mais que Perfeito</b>	<b>Imperfeito Conjuntivo</b>	<b>Futuro Conjuntivo</b>
do Pretérito	fiz	fizera	fizesse	fizer

Perfeito	fizeste fez	fizeras fizera	fizesses fizesse	fizeres fizer
----------	----------------	-------------------	---------------------	------------------

Tempos	<b>Futuro Indicativo</b>	<b>Condicional</b>	<b>Infinitivo Pessoal</b>	<b>Particípio</b>
do Infinitivo Impessoal	farei farás fará	faria farias faria	fazer fazeres fazer	feito <hr/> <b>Gerúndio</b> fazendo <hr/>

De forma idêntica ao verbo *estar*, salvo algumas irregularidades puras, temos uma alteração no presente do conjuntivo, em que o radical passa a ser *faç*. Nos tempos derivados do pretérito perfeito, o radical passa a ser *fiz*, enquanto nos tempos derivados do infinitivo impessoal aparecem dois tempos com o mesmo radical, *f* (seria o verbo *far*).

Nestes exemplos, verificamos que as alterações de radicais limitam-se à mesma família de tempos derivados. Por exemplo, o "pseudo-radical" *fiz* só ocorre nos tempos derivados do pretérito perfeito e nunca em nenhum dos outros. De forma idêntica para os outros pseudo-radicais: *f* e *faç*.

A implicação prática destas observações é a possibilidade de aplicar as regras dos verbos regulares a formas de verbos irregulares, aplicando-lhes depois uma regra do género:

*Se o radical do verbo for **fiz** e se o tempo do verbo for derivado do pretérito perfeito, então transformar esse radical em **faz**.*

Assim, poupamos uma extensa lista de formas de verbos irregulares.

Para implantar este processo, as regras dos verbos regulares foram divididas por famílias de tempos derivados, associando a cada uma destas famílias a sua própria tabela de conversões de radicais. Assim, as transformações de radicais são sempre efectuadas no âmbito de uma determinada família de tempos.

O formato das regras está representado no quadro 2.15:

	<b>Verbo Analisado</b>	<b>Conversão</b>	<b>Pessoa</b>	<b>Número</b>	<b>Tempo</b>
	ouçer	ouvir	*	*	pc
<i>Presente</i>	ouçer	ouvir	1	S	pi
	peçer	pedir	*	*	pc

	peçer	pedir	1	S	pi
	*disser	*dizer	*	*	*
<i>Pretérito</i>	estiver	estar	*	*	*
	*fizer	*fazer	*	*	*
	trar	trazer	*	*	fi
<i>Infinitivo</i>	trar	trazer	*	*	c
	*far	*fazer	*	*	fi
	*far	*fazer	*	*	c

Quadro 2.15 - Formato e exemplos de regras de conversão dos radicais dos verbos.

A forma analisada é a que resulta da aplicação de uma regra dos verbos regulares. Por exemplo, a forma *ouças* é entendida pelas regras dos verbos regulares como uma forma do presente do conjuntivo do verbo *ouçer* (nesta fase ainda não foram aplicadas as regras de conversões regulares). O campo do quadro 2.15 rotulado com "Conversão" contém o novo radical que o verbo deve ter, preservando a sua classificação morfológica. Os três campos seguintes – pessoa, número e tempo – servem para restringir a aplicação deste tipo de regras a um conjunto específico, evitando assim possíveis análises erradas. Por exemplo, a primeira regra aplica-se só às formas do presente do conjuntivo (qualquer pessoa e número), enquanto que a segunda regra aplica-se apenas à primeira pessoa do singular do presente do indicativo. A forma *ouçes* (ao contrário de *ouço*) após ter sido obtido o radical *ouçer*, não seria convertida para *ouvir*.

As regras cujos verbos aparecem precedidos de asterisco aplicam-se a todos os verbos que

terminem daquela forma. Por exemplo, \**disser* aplica-se a todas as análises terminadas em *disser* (*condisser*, *bendisser*, *predisser*, etc.) e, ainda, a qualquer tempo da família de tempos derivados do pretérito perfeito.

Se um radical não sofrer nenhuma destas conversões, é sujeito a uma verificação inversa, testando-se a possibilidade de ser o resultado de uma possível conversão. Se ocorrer essa situação, elimina-se a classificação correspondente ao radical em questão.

Este processamento destina-se a evitar que uma forma com flexão regular seja aceite como correcta, quando existe uma forma irregular para a mesma situação. Imaginemos, por exemplo, que se pretendia analisar a palavra *trazerei*. Aplicando as regras dos verbos regulares, esta palavra seria reconhecida como uma forma do verbo *trazer*, não estando sujeita a qualquer conversão do radical. Fazendo a verificação inversa constatar-se-ia que, na família dos tempos derivados do infinitivo (na qual se inclui o futuro do indicativo), *trazer* é resultado da conversão de *trar*. Isto é, existe uma regra que converte o verbo *trar* no verbo *trazer*. Quer isto dizer que, no caso particular do futuro do indicativo, o verbo *trazer* tem formas irregulares. Daí a necessidade de eliminar a classificação regular obtida.

## 29 Regras de restrição

Há análises que, à priori, se pode saber não estarem correctas, principalmente pela forma como termina o infinitivo impessoal do verbo (ou o tema +r). Por exemplo, não existe nenhum verbo em português que termine em *eer*, *fer*, *barater*, *baratir*, ou *ponhar*.

Além disso, sabe-se que com a terminação *der* só existem verbos terminados em *eder*, *nder*, *oder* e *rder*. Desta forma, se a palavra *fada* fosse analisada pelo Palavroso, as regras dos verbos regulares analisá-la-iam como:

<i>fadar</i>	1S do Presente Indicativo
<i>fadir</i>	1S, 3S do Presente do Conjuntivo
<i>fader</i>	1S, 3S do Presente do Conjuntivo

A última análise, *fader*, termina em *der*, mas não é precedida de *e*, *n*, *o* ou *r*, donde se conclui que este verbo não existe. Esta classificação é retirada da lista de classificações.

Estas restrições são impostas por recurso a listas de terminações de radicais impossíveis para verbos e listas contendo as únicas terminações de radicais possíveis num determinado contexto. As listas têm todas o mesmo formato, variando apenas a sua interpretação. O formato está ilustrado no quadro 2.16.

Terminação	er	ir	ar
0ord	1	1	1
nceir	0	0	1
i	1	1	0

Quadro 2.16 - Formato e exemplos de regras de restrições dos radicais dos verbos.

Estas regras são compostas por quatro campos. O primeiro contém a terminação do radical. Os outros três são campos booleanos indicando a que conjugação se aplica a restrição. Por exemplo, a segunda regra só se aplica a verbos da primeira conjugação: é uma restrição para verbos terminados em *nceirar*.

Quando a terminação vem precedida de 0 (zero), significa que não se trata de uma terminação, mas sim de uma palavra completa. Por exemplo, assumindo que as entradas do quadro 2.16 são terminações impossíveis, a primeira regra diz que não existem os verbos *order*, *ordir* nem *ordar*.

Existem três destas listas de terminações, que identificamos aqui por Impossiv, Possiv1 e Possiv2.

A lista Impossiv contém terminações de radicais de verbos impossíveis de ocorrer em português. Se as regras dos verbos regulares produzirem algum radical cuja terminação esteja presente nesta lista, a classificação correspondente será eliminada.

A lista Possiv1 contém terminações sobre as quais nos podemos pronunciar pela afirmativa. A lista Possiv2 contém as únicas terminações possíveis no contexto dado por Possiv1.

Primeiro é consultada a lista Possiv1, verificando-se se a terminação do verbo se encontra lá. Se não for encontrada, o Palavroso não pode fazer restrições a esta forma particular. Se for encontrada, então vai-se verificar na lista Possiv2 se a terminação do verbo lá se encontra. Se existir, o verbo é possível. Se não existir, conclui-se que o verbo não existe, eliminando-o, então, da lista de classificações.

Retomemos novamente o exemplo de *fada* e do verbo *fader*.

Ao consultar a lista Possiv1, encontra-se lá a terminação *der* (*d* aplicada à segunda conjugação). Passa-se então a consultar a lista Possiv2. Aí as terminações são mais compridas, e, não só não encontramos a terminação *der*, como não encontramos terminações *ader* nem *fader*. Conclui-se então que o verbo *fader* não existe.

Suponhamos agora que se estava a verificar o verbo *elucidar*. Consultando a lista Possiv1 encontra-se a terminação *cidar*, logo é possível que haja alguma restrição ao verbo em questão. No entanto, ao consultar a lista Possiv2 encontra-se lá a terminação *ucidar*, que se aplica ao verbo corrente e, portanto, não podemos dizer que não existe.

Outro processo de restrição de radicais dos verbos consiste em eliminar todas as classificações cujo radical do verbo tenha um acento gráfico e não seja o verbo *pôr*.

## 30 Dicionário

O formato do dicionário dos verbos é o mesmo tanto para o caso simples como para os verbos formados por mais do que um lexema. Cada entrada consiste apenas numa palavra e informação probabilística relativa a essa palavra.

A figura 2.17 contém alguns extractos de cada um destes dicionários.

	<b>Palavra</b>	<b>Probabilidade</b>
<b>Verbos simples</b>	abandar	1.0
	coleccionar	1.0
	descansar	1.0

	<b>Palavra</b>	<b>Probabilidade</b>
<b>Verbos compostos</b>	ab-rogar	1.0
	bem-fazer	1.0
	contra-ordenar	1.0
	co-herdar	1.0

Quadro 2.17 - Extractos dos dicionários de verbos simples e compostos.

### 31 Funcionamento global

Em relação aos verbos, explicámos como funcionava cada parte do processamento, mas não descrevemos como se ligam as partes entre si.

Primeiro, faz-se o processamento relativo aos verbos irregulares, começando-se por verificar se a palavra em análise é uma forma irregular de um verbo.

Depois, o processamento é dividido em três partes, uma por cada família de tempos derivados. Para cada família de tempos, começa-se por aplicar as regras dos verbos regulares, criando todas as classificações morfológicas possíveis. O passo seguinte consiste em verificar cada uma destas classificações.

Assim, para cada classificação obtida por aplicação das regras dos verbos regulares, efectuem-se as seguintes operações (por esta ordem):

1. Faz-se a conversão do radical, se alguma regra de transformação se lhe aplicar.
2. Eliminam-se as classificações cujos radicais tenham acentos gráficos.
3. Eliminam-se as classificações obtidas pelas regras regulares, mas que existam na lista de formas irregulares de verbos. Por exemplo, a forma verbal *veste* é analisada pelas regras dos verbos regulares como uma forma do verbos *vestir*, e, também, como pretérito perfeito de *ver*. Esta classificação é eliminada ao ser encontrada na lista das formas irregulares a forma correcta: *viste*.
4. São efectuadas as conversões regulares, relacionadas com a ortografia.
5. É efectuada a restrição dos radicais, usando as listas Impossiv, Possiv1 e Possiv2.
6. Consulta-se o dicionário dos verbos para verificar se o verbo calculado lá se encontra. Esta-

belece-se a sua probabilidade em 100% ou 0% conforme exista, ou não.

## **32 Advérbios de modo**

Este módulo destina-se a verificar se uma palavra é de um advérbio terminado em *mente*. Todas as palavras com esta terminação são fortes candidatas à classificação de advérbio. Existem, contudo, algumas palavras que terminam em *mente* e não são advérbios. É o caso de *semente* e *comente*.

Por esta razão, apenas se verifica se uma palavra terminada em *mente* é um advérbio, no caso de ainda não ter sido classificada como nome, adjectivo ou verbo.

Palavra	Probabilidade
absolutamente	1.0
briosamente	1.0
ingenuamente	1.0

Quadro 2.18 - Extracto do dicionário dos advérbios terminados em *mente*.

Se a palavra não tiver nenhuma daquelas classificações, então verifica-se a sua existência numa lista de advérbios de modo (quadro 2.18). Se lá existir, a palavra é classificada como advérbio. Senão, não deixa de ser um potencial advérbio. A diferença reside no valor do campo reservado à probabilidade de a classificação estar correcta, pois é inferior quando não existe na lista de advérbios. Na actual forma de processamento, a probabilidade é zero, se não for encontrada na lista e 100% se lá se encontrar. Por exemplo,

*altamente*

altamente:

adv, (altamente, 100%)

*baixamente*

baixamente:

adv, (baixamente, 0%)

### 33 Verbos com enclíticos

O processamento dos verbos com enclíticos consiste, fundamentalmente, em identificar as várias partes que compõem a palavra, verificando ao mesmo tempo se se encontram de acordo com as regras gramaticais. Por exemplo, não deve aceitar formas como *canto-se*, *falaria-lhe*, *trazer-se-á*, ou *deu-o-me*.

Os elementos que podem ser ligados à forma verbal através de hífen são o marcador de reflexividade, o marcador de objecto e o objecto indirecto (Santos, 1989). Por exemplo:



lavo-me lavas-te lava-se	Reflexividade
dou-te dei-lhe	Objecto indirecto
cantou-a chama-o	Objecto directo

No caso do verbo estar no futuro do indicativo ou no condicional, em vez de se colocar os pronomes no final, colocam-se entre o radical do verbo e a terminação característica do futuro ou condicional. Por exemplo,

*lavar-me-ei* e não *lavarei-me*  
*dar-te-ia* em vez de *daria-te*

No caso de o objecto directo e o objecto indirecto estarem ligados ao verbo por hífen, estes contraem-se formando uma partícula só. Por exemplo,

*cantou-ta* (te + a)

*lavou-ma* (me + a)

*dar-lho-ei* (lhe + o)

Em determinadas circunstâncias, a ligação de um enclítico ao verbo introduz alterações a nível ortográfico e fonético. É o caso das formas no infinitivo impessoal, quando seguidas por um objecto pronominal. Por exemplo,

*cantá-la, chamá-lo, vendê-la*

Por vezes o pronome *se* tem uma função apassivante, quando torna o

sujeito indeterminado (*deu-se um acidente...*). Nestes casos, ainda podem vir ligados ao verbo tanto o objecto como o objecto indirecto. Por exemplo,

*deu-se-lhas, encontrar-se-lhos-iam*

No Palavroso, existem duas linhas condutoras no processamento de verbos com enclíticos. Uma de análise e identificação, outra de verificação. Por um lado, é necessário reconhecer os enclíticos e a forma verbal. Por outro, é necessário garantir que uma forma inválida nunca seja aceite como correcta.

Quanto ao processo de análise, efectua-se as seguintes operações:

1. Reconhecer a terminação verbal e o pronome que a segue, no caso de este ter introduzido alterações a nível ortográfico. Por exemplo, *cantá-la* deve ser transformada em *cantar-a*. Estas conversões são feitas por recurso a regras que estabelecem as transformações possíveis, como as do quadro 2.19.

Nem sempre é possível fazer este tipo de conversões de forma não ambígua. Por exemplo, *pô-los* pode ser transformado em *pôr-os* ou em *pôs-os*. Para tratar estes casos, são aplicadas todas as regras de transformação possíveis, criando um novo radical por cada regra que seja aplicada. Por vezes alguns dos radicais assim criados são eliminados, pois não são válidos. Por exemplo, o verbo *fá-lo*, é convertido em três formas: *faz-o*, *fas-o* e *far-o*. Só a forma *faz* é válida, pelo que as outras conversões são eliminadas.

No caso do verbo *pôr*, são dadas as duas formas como resultado.

Sequência existente	Nova sequência
á-l	ar-
ê-l	er-
o-n	os-
õe-n	ões-

---

Quadro 2.19 - Exemplos de entradas numa tabela de regras de conversão dos enclíticos.

2. Reconhecer contracções de pronomes e isolá-los. Por exemplo, *ma* deve ser convertido em *me-a*. Existe uma tabela com este tipo de conversões.
3. Colocar junto ao radical do verbo a terminação de futuro ou condicional, caso exista. Por exemplo, *chamar-se-ia* é transformado em *chamaria-se*.

Após o processo de análise, fazem-se as seguintes verificações:

1. Verificar a existência da forma verbal encontrada. Usar para isso o módulo de processamento dos verbos. Se não existir, a palavra não pode ser considerada um verbo com enclíticos. Se existir, obter o tempo, pessoa e número.
2. Verificar se, para além da forma verbal, existe um máximo de três partículas. Neste caso, a primeira tem necessariamente que ser *se*, a segunda uma de {*me, te, lhe, nos, vos, lhes*} e a terceira uma de {*o, a, os, as*}.
3. No caso de uma das partículas ser *se*, verificar se o verbo *se* encontra na terceira pessoa (do singular ou do plural), e a partícula *se* encontra directamente ligada ao verbo.
4. Verificar se os pronomes {*o, a, os, as*}, caso apareçam, são os últimos da lista de partículas.
5. Verificar se a terminação da forma verbal foi transportada do final da palavra para junto do radical, no caso de o tempo do verbo ser o condicional ou o futuro do indicativo. Se este processo não ocorreu, então a palavra não pode ser considerada válida.
6. No caso de haver objecto directo e objecto indirecto sob a forma de pronomes, verificar se estes inicialmente se encontravam contraídos.

### 34 Palavras compostas

As palavras compostas podem ser formadas por aglutinação ou justaposição. No primeiro caso, as componentes envolvidas no processo de composição perdem o seu acento, originando uma palavra de acento único. No segundo caso, as componentes mantêm a sua individualidade, sem perderem o seu acento (Figueiredo e Ferreira, 1974).

O Palavroso, na sua actual versão, não faz qualquer tratamento especial em relação às palavras compostas por aglutinação, processando-as como qualquer palavra primitiva. Só são sujeitas a um processamento específico as palavras compostas por justaposição, bem como as palavras derivadas

por prefixação com justaposição.

A restrição deste tratamento específico apenas aos fenómenos que envolvem justaposição deve-se a duas razões fundamentais: por um lado, a identificação de cada uma das componentes é imediata. Por outro, são estas as palavras que levantam problemas de flexão, seguindo padrões nem sempre bem definidos (cf. Barreiro et al. 1993).

Em termos de análise morfológica, poder-se-ia pensar que bastaria verificar se cada componente da palavra composta é uma forma válida do português. Esta abordagem simplista traria dois problemas:

1. Poderia aceitar composições inexistentes, como *mesa-pirilampos* ou *flor-fala*.
2. Poderia aceitar palavras compostas mal flexionadas, como *quintas-feira* ou *caminho-dos-ferro*.

O primeiro problema pode ser resolvido aceitando apenas composições que existam no dicionário. A dificuldade aumenta no caso das palavras mal flexionadas, pois não é possível analisar apenas cada uma das componentes individualmente. O problema só pode ser resolvido olhando a palavra como um todo.

Assim, foram criadas regras de flexão para palavras compostas e um dicionário destas palavras, que se encontram descritas e fundamentadas em Barreiro et al. (1993). Essas regras são do género:

- ∇ Se a palavra composta for formada por dois nomes, então ambos os nomes flexionam no plural: *couve-flor* / *couves-flores*.
- ∇ Se a palavra composta for formada por um verbo e um nome, só o segundo flexiona no plural: *guarda-sol* / *guarda-sóis*.

Estas regras exigem que cada componente seja analisada individualmente, considerando, depois, a palavra no seu conjunto. Embora a ideia genérica seja simples, esta abordagem traz problemas computacionais. Por exemplo, a palavra *guardas* tanto pode ser um verbo como um nome. Assim, *guardas-sóis* pode ser considerado um caso de palavra composta formada por dois nomes, donde ambas as componentes devem flexionar para formar o plural. Isto é, *guardas-sóis* poderia ser aceite como válida.

Este problema foi solucionado colocando no dicionário de palavras compostas a classe gramatical de cada componente. Por exemplo,

Palavra	Número	Género	Classe	Classe-Classe
guarda-sol	S	M	nome	verbo-nome

guerra-fria	S	F	nome	nome-adj
livre-pensador	S	M	nome	adj-nome
mesa-de-cabeceira	S	F	nome	nome-prep-nome

---

Quando uma palavra composta é submetida à análise, cada componente é analisada individualmente, construindo-se todas as palavras compostas possíveis a partir das análises efectuadas. Para cada palavra construída, cria-se um registo das flexões detectadas. Depois, cada uma destas palavras compostas é verificada em dicionário. Se existir, verifica-se se foi obtida por um processo de flexão válido, comparando o seu registo de flexões com o registo permitido, ditado pelas regras de flexão e a classe gramatical das componentes que se encontra em dicionário.

Para além deste processo geral, existem alguns processos específicos. Nomeadamente, para reconhecer palavras compostas com alguns prefixos específicos, como *anti*, *super*, *vice*, *ex*, etc.

Para as palavras fechadas compostas e os verbos compostos não é necessário dicionários com tanta informação como o acima descrito. O dicionário dos verbos compostos tem o mesmo formato que no caso simples, enquanto as palavras fechadas compostas são representadas como no quadro 2.20.

Palavra	Categoria	Probabilidade
ai-jesus	ij	1.0
assim-assim	adv	1.0

---

Quadro 2.20 - Extracto da lista de palavras fechadas compostas.

## 35 Utilizações do Palavroso

O Palavroso tem feito parte integrante de várias aplicações de processamento de linguagem natural.

Nomeadamente, em ferramentas de processamento automático de corpos de texto e em gramáticas computacionais. A utilização do Palavroso na correcção ortográfica será pela primeira vez descrita nesta dissertação (cf. capítulos 3 a 6).

### 36 Em ferramentas de processamento de corpos de texto

Existem quatro ferramentas de processamento de corpos de texto baseadas no Palavroso:

<b>morfolog</b>	Programa para fazer análise morfológica interactivamente e anotar texto automaticamente.
<b>encontra</b>	Aplicação que permite encontrar em corpos de texto sequências de palavras com determinadas características morfológicas, especificadas pelo utilizador.
<b>estatic</b>	Ferramenta que faz diversos tipos de contagens de características morfológicas sobre textos.
<b>orto</b>	Ferramenta de auxílio ao preenchimento do dicionário do Palavroso. Dado um texto cria uma lista de palavras, ou formas de palavras, que não se encontram no seu dicionário, com todas as classificações possíveis segundo as regras.

Estas ferramentas, e sua aplicação, vêm descritas em Medeiros (1992).

### 37 Em gramáticas computacionais

O Palavroso tem sido utilizado como módulo de análise morfológica em várias gramáticas. A primeira (no tempo) aplicação do Palavroso, e que motivou o seu desenvolvimento, foi a elaboração de uma gramática sem dicionário, para ser incluída num sistema de síntese de fala a partir de texto, como descrito em Santos et al. (1992).

Também Fonseca (1993) utilizou o Palavroso como módulo morfológico numa gramática que serviu de base a uma interface em linguagem natural para um sistema tutor.

Finalmente, o Palavroso foi integrado no ambiente de desenvolvimento de aplicações de linguagem natural, G, desenvolvido pelo grupo de linguagem natural da Microsoft Co.

## 38 Estudos efectuados com o Palavroso

O Palavroso esteve na origem de diversos estudos efectuados pelo Grupo de Linguagem Natural do INESC, umas vezes como ferramenta, outras vezes como objecto do próprio estudo. Por ordem cronológica, destacamos:

*Português quantitativo* (Medeiros et al., 1993). Estudo quantitativo sobre o português, baseado em corpos de texto. Apresenta a medida do grau de preenchimento lexical do português e o grau de ambiguidade entre algumas categorias morfológicas.

*O corpus e a classificação sintáctica dos verbos* (Nascimento et al., 1993). Neste estudo procura-se demonstrar que o uso real dos verbos nem sempre segue as regras descritas em gramáticas. Aqui foram usadas ferramentas de processamento de corpos de texto baseados no Palavroso.

*Dicionários de língua corrente: Algumas considerações* (Reis, 1993). Este artigo discute algumas falhas dos dicionários da língua portuguesa. Nomeadamente, a omissão de vocábulos e a falta de actualização de novos significados em formas já existentes. Este trabalho foi motivado pelo preenchimento lexical do dicionário do Palavroso.

*Critérios e opções linguísticas no desenvolvimento do Palavroso, um sistema computacional de descrição morfológica do Português* (Barreiro et al., 1993). Tendo como base a criação de regras e o preenchimento lexical do dicionário do Palavroso, neste estudo são feitas algumas reflexões sobre vários problemas da descrição da língua portuguesa sob o ponto de vista morfológico.

*Anotação contextual do corpus INESC 1990* (Marques, 1994-a). Relatório que descreve o trabalho manual de desambiguação da anotação automática feita pelo morfolog, ou seja, um anotador baseado no Palavroso.

*Homografia: Relações morfológicas e semânticas* (Marques, 1994-b). Estudo sobre relações morfológicas e semânticas que se estabelecem entre palavras morfológicamente ambíguas. Neste estudo é usado texto anotado pelo Palavroso.

*Português computacional* (Santos, 1994). São aqui discutidos alguns problemas do português, tendo como pano de fundo as opções linguísticas que é necessário tomar no desenvolvimento de aplicações como o Palavroso. São também apresentados alguns estudos quantitativos baseados no Palavroso, bem como sugestões de utilização de um analisador morfológico no ensino do Português.

*European and Brazilian Portuguese: Quantitative report of lexical, syntactical and orthographic differences* (Wittmann e Pereira, 1994). Relatório que descreve aspectos quantitativos da diferença entre as duas versões do português. Usaram ferramentas de processamento de corpos de texto baseadas no Palavroso.

*Uso de informação quantitativa num analisador morfológico de Português* (Medeiros, 1994). Neste trabalho são discutidos alguns aspectos da inclusão de informação quantitativa num analisador morfológico como o Palavroso.

### **39 Alguns dados sobre a implementação do Palavroso**

O Palavroso foi completamente escrito em linguagem C (ANSI). Esta característica torna-o um sistema quase completamente portátil. Neste momento, existem versões a funcionar em ambientes Unix, MS-DOS e MS Windows. A sua compilação foi testada usando o compilador GNU gcc, o que torna o Palavroso uma aplicação disponível para a vasta gama de ambientes onde se possa instalar este compilador.

Na sua versão completa, tira partido de 2300 regras (quadro 2.22) e os seus dicionários contêm cerca de 54000 entradas (quadro 2.21). No total, estima-se a sua cobertura entre 1 300 000 e 1 500 000 formas do português (cf. Barreiro et al., 1993).

<b>Componente do Dicionário</b>	<b>Número de entradas</b>
Advérbios	3641
Nomes e adjectivos	36954

Palavras fechadas	721
-------------------	-----

Verbos	13065
--------	-------

---

---

<b>Total</b>	<b>54381</b>
--------------	--------------

---

Quadro 2.21 - Número de entradas lexicais dos dicionários do Palavroso.

<b>Componente</b>	<b>Tipo de regras</b>	<b>Total Parcial</b>	<b>Total</b>
Verbos	Verbos regulares	160	
	Verbos irregulares	248	
	Regras de Transformação	407	
	Regras de restrição	789	
	<b>Total</b>	<b>1604</b>	<b>1604</b>
=====			
Nomes	Género	161	
	Número	130	
	Grau	126	
	<b>Total</b>	<b>471</b>	<b>471</b>
=====			
Enclíticos		37	37
Palavras Compostas		128	128

Interface com o exterior	60	60
Total		2300

Quadro 2.22 - Número de regras do Palavroso

## 40 Resumindo

Neste capítulo descrevemos em pormenor o analisador morfológico Palavroso. Ao longo desta descrição, foi possível evidenciar alguns aspectos que demonstram a possibilidade de utilizá-lo eficazmente como parte integrante de sistemas de processamento de linguagem natural.

Destes aspectos, relembramos o facto de basear o seu processamento fundamentalmente em regras; a possibilidade de dar sempre alguma resposta, mesmo que não disponha de toda a informação no dicionário; a sua modularidade, que permite a integração fácil com outros sistemas; a portabilidade entre várias plataformas computacionais e o tratamento de alguns problemas da morfologia portuguesa reconhecidamente difíceis, como os verbos com enclíticos e certo tipo de palavras compostas.

## 41 Correção ortográfica

A correção ortográfica consiste na tarefa de detecção e correção de erros ortográficos. Podemos fazê-la à mão ou usar ferramentas que nos auxiliem nessa tarefa. A ferramenta usada na correção automática da ortografia denomina-se corrector ortográfico.

A correção ortográfica insere-se no problema mais geral de correção de textos escritos, que inclui, também, a correção da construção de frases e a correção do estilo.

Temos, então, três níveis de manipulação de um documento. O primeiro, ortográfico, limita-se a controlar a correção das palavras isoladas, independentemente do seu uso. O nível gramatical, imediatamente acima, preocupa-se com problemas como os de concordância, pontuação e uso de preposições.

Por fim, o nível estilístico. Neste nível, trabalha-se com frases gramaticalmente bem construídas, mas que por uma questão de estilo deveriam ser corrigidas. Estas correções têm como objectivo tornar o texto mais legível, tornando as construções gramaticais mais simples, ou reduzindo o tamanho das frases. A correção de estilo pode envolver também a verificação da conformidade do texto com linhas de orientação previamente estabelecidas, como por exemplo, as elaboradas para a

escrita de documentos técnicos.

A divisão entre estes níveis nem sempre é clara, principalmente quando se passa da correcção manual para a correcção automática de documentos. Por exemplo, existem certos erros que são tradicionalmente apontados como ortográficos, mas que não podem ser detectados automaticamente sem a ajuda da análise sintáctica da frase em que aparecem, como *prezo* de *prezar* vs. *preso* de *recluso* (Filho, 1994).

A figura 3.1 apresenta a proporção de erros encontrados em textos escritos em inglês<sup>1</sup>, de acordo com os dados apresentados por Nijholt (1992).

μ §

Figura. 3.1 - Proporção dos erros encontrados em textos escritos em inglês.

Embora os erros detectáveis ao nível da palavra sejam os menos frequentes, a grande maioria das ferramentas construídas para auxiliar a correcção de textos escritos limita-se a este tipo de erros. A razão disso tem a ver com a complexidade de um verificador/corrector gramatical. Este, para além da correcção ortográfica, tem que fazer a análise sintáctica das frases (Richardson & Braden-Harder, 1988; Voss, 1992).

Nesta dissertação tratamos apenas o problema da correcção ortográfica ao nível da palavra, usando um corrector automático ou semi-automático.

## 42 O problema

A correcção ortográfica envolve dois processos distintos:

- ∇ Verificação ortográfica, em que simplesmente se verifica se cada uma das palavras de um texto pertence à língua ou não.
- ∇ Sugestão de alternativas às palavras assinaladas como incorrectas.

O primeiro processo depende bastante da representação que o sistema tem da língua. Pode recorrer a métodos estatísticos e à análise morfológica, ou pode simplesmente verificar se a palavra existe num dicionário que se deseja suficientemente representativo da língua.

O segundo processo consiste em criar uma lista de palavras parecidas com a palavra errada. A

---

<sup>1</sup> Como não temos conhecimento de estudos idênticos para a língua portuguesa, aceitamos que as proporções entre tipos de erros possam não ser independentes da língua. Apresentamos o caso do Inglês apenas a título ilustrativo.

função desta lista é fornecer possíveis palavras candidatas à substituição da palavra errada.

As primeiras ferramentas de auxílio à correcção ortográfica consistiam em simples verificadores ortográficos: os *spelling checkers*. Desde então estas ferramentas evoluíram bastante, em particular passou a haver ajuda no processo de correcção propriamente dito. Apareceram então os *spelling correctors*. No entanto, mesmo hoje em dia, existe o hábito de falar de *spelling checkers*, quando na realidade se tratam mais de *spelling correctors*. Em português, usa-se o termo "correcção ortográfica" genericamente, abrangendo os dois processos: verificação e correcção.

### 43 Modos de correcção

Há várias formas de proceder à correcção ortográfica de um texto. Pode ser um processo essencialmente interactivo ou não interactivo. Neste último caso, pode ser completamente automático, ou requerer no final a intervenção do utilizador.

Num processo interactivo, ou semi-interactivo, é o utilizador que escolhe as substituições adequadas. No primeiro caso, conforme vão sendo detectadas as palavras com erro, é fornecida ao utilizador uma lista de sugestões de entre as quais pode escolher a substituição certa. No segundo caso, o corrector começa por fazer a verificação ortográfica de todo o texto, assinalando todas as ocorrências de palavras que não reconheça. O utilizador intervém só no final, resolvendo os problemas encontrados.

Não é pacífico decidir se uma abordagem é melhor que a outra. A interactividade tem a vantagem de o utilizador acompanhar o processo de correcção, podendo ir acrescentando palavras ao dicionário de utilizador, ou especificando palavras que o corrector deve ignorar, ou ainda interromper o processo no ponto que achar aconselhável. Tem a desvantagem de, eventualmente, o utilizador ter que corrigir o mesmo tipo de erro mais do que uma vez.

O processo semi-interactivo tem a principal vantagem de o utilizador partir para a correcção propriamente dita tendo já conhecimento de todos os erros que ocorreram. Desta forma pode corrigir os erros que quiser, pela ordem que quiser. Pode também evitar a correcção do mesmo tipo de erro mais do que uma vez.

Se o processamento for completamente automático, é o próprio corrector que determina qual a melhor substituição a fazer quando encontra uma palavra com erro. Um método de correcção deste tipo está sempre sujeito a erros, que serão tanto mais quanto mais abrangente for o domínio de aplicação. Quanto mais reduzido for o vocabulário, menor é a probabilidade de haver palavras parecidas entre si, diminuindo também a probabilidade de haver ambiguidades quanto à palavra a substituir. Poderíamos, por exemplo, pensar em usar um corrector deste tipo em ambientes

computacionais onde seja necessário escrever comandos. É o caso de um sistema operativo, de um sistema de gestão de bases de dados, ou ainda de compiladores de linguagens de programação. No caso dos sistemas operativos ou das bases de dados, o dicionário do corrector seria constituído por todos os comandos do sistema mais os identificadores de variáveis ou campos que existam. No caso dos compiladores, o dicionário poderia ser constituído pelas palavras-chave da linguagem mais as variáveis que tenham sido declaradas no programa.

## 44 Cobertura lexical

A questão da restrição da utilização do corrector a um domínio específico conduz-nos ao problema da cobertura lexical que o corrector deve ter.

Se, por um lado, é importante que o corrector ortográfico reconheça o maior número possível de palavras, também se deve ter o cuidado da sua cobertura lexical não ser exagerada.

Com efeito, se o corrector conhecer muitas palavras evita-se a frequente solicitação da intervenção do utilizador para corrigir erros que na verdade o não são. Por outro lado, se a cobertura lexical for demasiado extensa, corre-se o risco de ter no dicionário palavras muito raras, arcaicas ou obsoletas. Principalmente em documentos técnicos, as ocorrências deste tipo de palavras são, provavelmente, erros que nunca serão detectados porque a palavra existe no dicionário (Peterson, 1980).

Tomemos como exemplo a palavra *consciencia*. Vamos supor que existe o verbo *conscienciar* e que o dicionário do corrector ortográfico contém as formas deste verbo. Então, sempre que o utilizador se esquecer de num texto colocar o acento em *consciência*, o corrector não será capaz de detectar este erro, pois reconhecerá a palavra como sendo uma forma do verbo *conscienciar*. Seria preferível que o corrector não reconhecesse a palavra *consciencia*, e assinalasse erro quando a encontrasse, do que deixar passar em branco todos os outros erros por falta de acento.

Este problema leva-nos a um outro, intrínseco aos correctores ortográficos: fazem a correcção ortográfica palavra a palavra, fora de contexto. Assim, nunca poderão detectar utilizações incorrectas de palavras correctamente escritas. Por exemplo, se o utilizador se enganar e escrever *numero* em vez de *número*, ou *poças* em vez de *possas*, o corrector nunca será capaz de detectar estes erros.

Um problema parecido com o da cobertura lexical é o do tamanho da lista de sugestões. Se por um lado interessa que a lista seja suficientemente grande para incluir a sugestão correcta, por outro lado deve ser suficientemente pequena para que o utilizador não tenha que procurar a palavra correcta numa lista de palavras demasiado extensa. O ideal seria que a lista de sugestões contivesse uma só palavra: a correcta. Não sendo tal possível, deve-se tentar que a palavra correcta esteja entre as primeiras.

Resumindo, a lista de sugestões deve ser tão pequena quanto possível, desde que a palavra correcta esteja presente. Um bom desempenho nesse sentido está dependente de um bom palpite de qual terá sido o erro cometido.

## 45 Medidas de distância

Criar uma lista de sugestões de possíveis correcções da palavra com erro consiste, essencialmente, em encontrar no dicionário do corrector ortográfico palavras parecidas à palavra com erro. A dificuldade está, geralmente, em decidir o que é uma palavra parecida, e, de entre as parecidas, as que constituirão a correcção mais plausível.

Existem várias medidas de distância entre duas cadeias de caracteres (palavras, no estudo presente). Vamos aqui rever três dessas medidas (Oflazer, 1994; Burillo, 1994).

## 46 Distância n-grâmica

Um n-grama é uma subsequência de caracteres de uma palavra, com comprimento  $n$ . Frequentemente é conveniente ter em conta a fronteira de palavra na definição de n-grama. Assim, adoptando o símbolo '\_' para representar um espaço, a decomposição da palavra *sequência* em trigramas será: {*\_se, seq, equ, quê, uên, ênc, nci, cia, ia\_*}.

A distância n-grâmica entre duas palavras é dada pelo número de n-gramas que não são comuns às duas palavras.

Por exemplo, se denotarmos por  $\mu$  § a distância n-grâmica entre as palavras  $X$  e  $Y$ , então  $D^2(\textit{correcto}, \textit{correto}) = 3$ , pois tem três digramas que não são comuns às duas palavras: {*et, ct, ec*}.

Se tomarmos o  $n$  igual a 3, então  $D^3(\textit{correcto}, \textit{correto}) = 5$ , em que os trigramas não comuns são: {*rec, ect, cto, ret, eto*}.

## 47 Distância de edição

A distância de edição relaciona a semelhança entre as duas palavras pelo número mínimo de operações simples (inserção de um carácter, eliminação, substituição e transposição de dois caracteres contíguos) que é necessário efectuar para converter uma na outra.

Por exemplo, a distância de edição entre *itens* e *itesm* é 2. Neste caso é necessário fazer uma substituição e uma transposição, não necessariamente por esta ordem. Temos então  $Ed(itens, itesm) = 2$ .

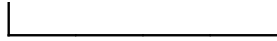
## 48 Matriz de semelhança

Uma das medidas de distância proposta por Burillo (1994) baseia-se em criar uma matriz de zeros e uns, descrevendo os caracteres que duas palavras têm em comum. Duas palavras estarão tanto mais próximas uma da outra quanto mais próximo estiverem os uns da diagonal principal. Se as palavras forem iguais, então a matriz resultante é a matriz identidade.

Depois de construída a matriz de semelhança, existe uma função que calcula um índice de semelhança entre as duas cadeias de caracteres.

Tomemos como exemplo o cálculo de semelhança entre *mesma* e *meza*. A matriz de semelhança é:

	m	e	z	a
m	1	0	0	0
e	0	1	0	0
s	0	0	0	0
m	1	0	0	0
a	0	0	0	1



Para calcular o índice de semelhança é necessário atribuir pesos aos valores da matriz por forma a valorizar os valores que se encontram próximos da diagonal principal (matriz 1). Para isso usa-se a expressão:

$$\mu_{ij}$$

em que:

$m^{ij}$  e  $p^{ij}$  são os elementos da matriz, (linha  $i$ , coluna  $j$ ), antes e depois do cálculo da expressão, respectivamente.

$n$  é o número de linhas e  $m$  é o número de colunas da matriz .

Depois é necessário anular alguns elementos, por forma a ficar apenas um elemento diferente de zero em cada linha e em cada coluna (matriz 2). Começando pela primeira, escolhe-se o maior elemento de uma linha e anulam-se todos os outros elementos da linha e coluna que intersectam nesse elemento.

	m	e	z	a	
m	1	0	0	0	
e	0	0.92	0	0	
s	0	0	0	0	
m	0.25	0	0	0	
a	0	0	0	1	
	Matriz 1				

	m	e	z	a	
m	1	0	0	0	
e	0	0.92	0	0	
s	0	0	0	0	
m	0	0	0	0	
a	0	0	0	1	
	Matriz 2				

O cálculo da distância propriamente dita entre as duas palavras é dado pela expressão:

$$\mu_{ij}$$

em que  $\mu$  §e  $S$  é a soma acumulada dos termos da matriz, obtida do seguinte modo: Sempre que se encontra uma sequência de  $n$  termos consecutivos (diferentes de zero) em diagonal, o primeiro termo dessa sequência é multiplicado por  $n$ , o segundo por  $n-1$  e assim sucessivamente.

No caso presente,  $\mu$  §

## 49 Classificação do tipo de erros

Usualmente, os erros são classificados quanto à sua origem em três classes (Damerau, 1964; Peterson, 1980; Berkel & Smedt, 1988):

1. Linguísticos, ortográficos ou de competência.
2. Tipográficos.
3. De armazenamento ou transmissão.

Os primeiros são de origem cognitiva, motivados por má formação linguística ou falta de informação do autor. A maior parte das vezes, as palavras com erro são homófonas<sup>1</sup> das palavras correctas, ou pelo menos, foneticamente muito semelhantes.

Os erros tipográficos têm uma origem de ordem motora. Acontecem quando a pessoa que dactilografou o texto cometeu um erro ao carregar nas teclas, produzindo uma sequência errada. Geralmente a palavra resultante não é homófona, nem fonologicamente parecida com a palavra correcta.

Por erros de transmissão entendemos os erros que se devem a deficiências nos canais de transmissão de dados, ou armazenamento dos mesmos.

Ao longo deste texto identificaremos estes tipos de erros (1, 2, 3) por **Linguísticos**, **Tipográficos** e de **Transmissão**. Usaremos o termo "erro ortográfico" genericamente, para nos referirmos a qualquer um dos anteriores.

Quanto ao tipo de erro, estes podem ser simples ou complexos. Os erros simples são aqueles que resultam de uma das quatro operações do quadro 3.2. Por outras palavras, se usarmos o conceito de distância de edição, as palavras com um erro simples são aquelas que se encontram à distância

---

<sup>1</sup> Gramaticalmente, palavras homófonas possuem o mesmo som mas têm grafia e **sentido** diferentes. Usamos neste texto a palavra "homófonas" num sentido mais restrito, querendo apenas dizer que se trata de palavras com a mesma realização fonética.

unitária da palavra correcta. Nas palavras com erros complexos, a distância de edição à palavra correcta é superior a um.

<b>Operação</b>	<b>Descrição</b>	<b>Exemplo</b>
Inserção	Aparece uma letra a mais na palavra.	<i>candiadaturas</i>
Substituição	Uma das letras aparece substituída por outra.	<i>subdtituição</i>
Omissão	Falta uma das letras à palavra.	<i>escever</i>
Transposição	Duas letras contíguas na palavra são trocadas entre si.	<i>algorimto</i>

Quadro 3.2 - Os quatro erros simples.

Damerau (1964) constatou que 80% dos erros são simples. Somente 20% dos erros são de ordem mais complexa, ou seja, o erro é consequência de mais do que uma das operações descritas no quadro 3.2. Por exemplo, *itesm (itens)* contem uma transposição (entre *s* e *m*) e uma substituição (de *n* por *m*).

Corrigir um erro provocado por uma das quatro operações assinaladas, implica aplicar a sua operação inversa. Por exemplo, para corrigir um erro de omissão, é necessário efectuar uma inserção. No entanto, se o erro for provocado por uma transposição, a operação inversa também é uma transposição.

A cada uma destas operações inversas, chamamos operação simples. Um erro simples pode ser corrigido pela aplicação de uma única operação simples. Para corrigir um erro complexo é exigida a aplicação de mais do que uma operação simples.

## **50 Erros linguísticos**

Assinalámos acima, como característica fundamental dos erros linguísticos, o facto de a maior parte

das vezes estes conduzirem a palavras homófonas com as palavras correctas, ou pelo menos fonologicamente muito parecidas. Também se verifica que as palavras com erros linguísticos a maior parte das vezes respeitam as regras da língua.

Estes dois factores levam-nos a concluir que uma das principais razões para as pessoas cometerem erros de origem linguística tem que ver com a diferença que existe entre a forma como a palavra é pronunciada e a forma como é escrita (Berkel & Smedt, 1988; Peterson, 1980). Quanto maior for o desvio, maior será a probabilidade de se cometer um erro.

A língua portuguesa contém vários fonemas que podem ter mais do que uma representação ortográfica. Por exemplo, o fonema /S/ pode ser obtido por *x* (*xarope*) ou por *ch* (*chave*). O fonema /s/ pode ser representado por *s* (*saco*), *ss* (*osso*), *c* (*cera*), *ç* (*poço*) ou *x* (*próximo*) (Cunha e Cintra, 1987).

Esta variedade de representações origina o aparecimento de palavras com erro homófonas das palavras correctas. Por exemplo, as palavras *chícara* e *xícara* são homófonas.

Situações de palavras com erro que são fonologicamente muito semelhantes com a palavra correcta também ocorrem com certa frequência em português. Por exemplo, *encómudo* (*incómodo*), *exspectáculo* (*espectáculo*) e *perconceito* (*preconceito*).

Este tipo de erro poderá, além disso, estar associado a uma dicção deficiente, ou à pronúncia específica de uma determinada região.

Em português, sempre que o erro consiste na omissão um acento gráfico, a palavra continua a obedecer às regras da língua: apenas se acentua de forma diferente. Não poucas vezes, também acontece essa omissão ainda originar uma palavra que existe na língua. Nestes casos, um corrector ortográfico pouco poderá ajudar. Um exemplo disso são as palavras *contínua* e *continua*.

Por vezes, o conhecimento etimológico da palavra ajuda-nos a decidir qual a forma correcta. Mas outras vezes nem isso nos ajuda. Por exemplo, escreve-se *rabugento* e *rabugice*, pois são palavras derivadas de *rabugem* (do latim, *rubigi&ne*). No entanto, escreve-se *rabuje*, pois trata-se de uma forma do verbo *rabujar*, mas que também deriva de *rabugem*.

Os erros de origem linguística são, de certa forma, previsíveis, sendo os mais persistentes. Também são os que deixam pior impressão no leitor, pois deixam transparecer ignorância (Berkel & Smedt, 1988; Nijholt, 1992).

## 51 Erros tipográficos

Por serem de origem motora, são aleatórios e imprevisíveis. Contudo, têm uma característica

interessante: frequentemente traduzem-se em sequências impossíveis da língua, como na palavra *impossível*.

A origem destes erros estará mais ligada às faculdades psico-motoras de quem escreve, ou ao teclado particular que se esteja a usar, do que às características da língua em uso.

Embora haja um elevado grau de imprevisibilidade nos erros tipográficos, se nos limitarmos a erros tipográficos provenientes de um único tipo de teclado, então acreditamos que se possa propor determinadas correcções com maiores probabilidades de estarem correctas, se forem baseadas nas distâncias entre as teclas.

Embora a proporção de erros simples apontada por Damerau (1964) seja baseada em contagens gerais, em que não separa os erros tipográficos dos erros linguísticos, quer-nos parecer que os tipos de erros simples por ele apontados são mais uma característica dos erros tipográficos, do que dos erros de origem linguística<sup>1</sup>. Tal não é difícil de aceitar, se acreditarmos que, como defende Pollock & Zamora (1984), os erros de origem linguística estão apenas entre 10% e 15% do total de erros ortográficos encontrados num texto.

Contrariamente aos erros de origem linguística, a maior parte das vezes as palavras com erros tipográficos não são homófonas nem foneticamente parecidas com a palavra correcta. Por exemplo, *dificukdades* vs. *dificuldades*.

---

<sup>1</sup> O próprio autor deixa transparecer essa ideia, quando após a descrição dos quatro tipos de erro acrescenta: "These are the errors one would expect as a result of misreading, hitting a key twice, or letting the eye move faster than the hand.". Outros autores chegam mesmo a considerar que os valores apresentados por Damerau se reportavam apenas a erros tipográficos (Berkel & Smedt, 1988; Agirre et al., 1992).

## 52 Erros de transmissão

Por erros de transmissão, entendemos os erros que se devem a deficiências nos canais de transmissão de informação. De entre estes, damos especial ênfase ao reconhecimento óptico de caracteres (OCR) e aos sistemas de envio e recepção de correio electrónico.

Os erros devidos ao reconhecimento óptico de caracteres não são muito diferentes dos tipográficos. Por exemplo, Peterson (1980) só reconhece para este tipo de processamento o erro de substituição. Embora dificilmente possamos admitir outro tipo de erro, em determinadas tipos de caracteres podemos ter erros mais complexos. Por exemplo, a sequência **rr** pode ser confundida com um **n**, ou **cl** com **d**.

Se tal acontecesse, estaríamos em presença de erros complexos. Recorrendo às operações simples, teríamos que fazer uma substituição e uma inserção para corrigir uma confusão entre **rr** e **n**.

De qualquer forma, podemos tirar partido do domínio dos erros e construir uma matriz de probabilidade de confusão. Por exemplo, a probabilidade de um **i** ser confundido com um **l** é maior que ser confundido com um **m**.

Os erros específicos que surgem no envio e recepção de correio electrónico são devidos à codificação dos caracteres acentuados.

Os caracteres acentuados são representados em ASCII usando uma tabela de oito bits. No entanto, os canais de comunicação e o dispositivo receptor das mensagens de correio electrónico funcionam muitas vezes com códigos de apenas sete bits (vd. Apêndice A).

Também pode acontecer o emissor e o receptor usarem códigos ASCII diferentes para representar a mesma letra com acento. De qualquer maneira, o resultado é quase sempre a substituição dos caracteres acentuados por outros caracteres quaisquer. E trata-se sempre de um erro de substituição. Por exemplo, *comunicação* pode ser recebido como *comunicagco*.

Como os erros de transmissão podem ser tratados como erros tipográficos, no tratamento que fizermos relativo aos erros tipográficos fica implícito o tratamento dos erros de transmissão.

## 53 Análise de erros ortográficos

Só poderemos resolver bem o problema da correcção ortográfica se conhecermos bem os erros

ortográficos que se pretende corrigir.

Nesta secção caracterizaremos os erros ortográficos da língua portuguesa que podem ser objecto de correcção considerando a palavra isolada, fora de contexto.

Usaremos, para este efeito, um corpo de erros ortográficos que compilámos, tendo em vista esta caracterização assim como o teste e a avaliação de correctores ortográficos. No apêndice A é possível encontrar uma descrição deste corpo de erros, quanto à sua organização, conteúdo e proveniência dos erros.

Faremos uma análise em separado para os erros de origem linguística e de origem tipográfica, uma vez que possuem características distintas, sendo, também, de origem diferente a motivação para as suas análises. Usaremos, pois, uma lista de erros tipográficos (CorpoTipo)<sup>1</sup> e uma lista de erros linguísticos (CorpoLing).

## 54 Erros linguísticos

Com a análise de erros de origem linguística pretendemos captar algumas características que nos possam sugerir alguns métodos de correcção dos erros que se baseiem mais no conhecimento linguístico, do que em processamentos meramente estatísticos.

Quando se usa uma motivação linguística na detecção e correcção de erros, procura-se detectar na palavra com erro alguns padrões específicos e típicos, para os quais se conhece, à priori, a solução. Esta abordagem traz várias vantagens:

- ∇ Produz menos sugestões, pois o conjunto de padrões típicos que uma palavra pode conter é bastante limitado.
- ∇ As sugestões produzidas encontram-se mais próximas da palavra correcta, devido à especificidade dos padrões que trata.
- ∇ Pode corrigir erros que sob o ponto de vista estatístico seriam considerados complexos, pois os padrões de erros que se procura envolvem, em geral, mais do que um grafema.
- ∇ Não necessita recorrer tantas vezes ao dicionário, pois o número de sugestões geradas não é muito extenso. Como consequência, pode-se esperar uma diminuição do tempo

---

<sup>1</sup> O corpo de erros encontra-se dividido em várias partes, em função da origem dos erros. Cada uma destas partes encontra-se rotulada com um acrónimo, que usaremos ao longo deste texto para as identificar. Por exemplo, CorpoTipo identifica a parte do corpo de erros que contém erros de origem tipográfica.

de resposta do corrector.

Numa análise prévia dos erros de origem linguística, constatámos que os erros se dividiam, quanto à sua origem, em dois grandes grupos:

- ∇ A palavra com erro é fonologicamente semelhante à palavra correctamente escrita, como por exemplo, *amenisar* vs. *amenizar*
- ∇ O erro é causado por ignorância a nível morfológico, reflectindo-se em fenómenos incorrectos de derivação e flexão. Por exemplo, *fazer-se-á* vs. *far-se-á*.

Por vezes, é possível encontrar palavras que reúnem as duas características anteriores. Nestes casos, consideramos que o erro é dado por desconhecimento das regras morfológicas, funcionando a semelhança fonológica apenas como elemento catalisador. É o caso *darnos* vs. *dar-nos*, ou *semicerrado* vs. *semicerrado*.

## 55 Semelhanças fonológicas

Tal como sugerido por outros autores (Berkel & Smedt, 1988), verificamos que, efectivamente, a semelhança fonológica entre a palavra com erro e a palavra correctamente escrita está na origem da grande maioria dos erros ortográficos de origem linguística.

Nas semelhanças fonológicas distinguimos os casos de homofonia de outras situações em que as palavras não são homófonas, mas são fonologicamente muito semelhantes, ao ponto de na linguagem falada se poderem confundir.

Os casos mais frequentes de confusão envolvendo homofonia dão-se nas seguintes situações:

1. Omissão ou inclusão de consoantes mudas. Por exemplo, *abtrato* vs. *abstracto*; *productivo* vs. *produtivo*.
2. Troca de s por z ou vice-versa, quando o s se encontra entre vogais, tendo realização fonética /z/. Por exemplo, *analiza* vs. *analisa*; *certesa* vs. *certeza*.
3. Confusão entre ss e ç, ou ss e c. É sempre possível substituir em qualquer palavra uma ocorrência de ss (se o seu valor fonético for /s/) por ç, sem alterar o seu valor fonético. No caso de ss estar antes de e ou i, a substituição pode ser feita com c. Por exemplo, *nessecidade* vs. *necessidade*; *barcassa* vs. *barcaça*.
4. Confusão entre ns e nç; confusão entre rs e rç. Este tipo de confusão pode ser considerado um caso particular do anterior. Exemplos: *compreenção* vs. *compreensão*; *cansão* vs.

*canção; torsão vs. torção; extorção vs. extorsão.*

5. Substituição da terminação *ez* por *ês*; substituição da terminação *iz* por *is*. Por exemplo, *gravidês vs. gravidez; actris vs. actriz; dis vs. diz.*
6. Colocação de acento gráfico quando não é necessário. Por exemplo, *campainha vs. campainha; cêra vs. cera; sòmente vs. somente.*
7. Substituição de *o* por *u*, e vice-versa, para o fonema /u/. Por exemplo, *culisão vs. colisão; borborinho vs. borburinho.*
8. Troca entre *j* e *g*. Exemplos: *alforje vs. alforge; arrange vs. arranje.*
9. Confusão entre *ch* e *x*, como em *chaile vs. xaile* e *xávena vs. chávena.*
10. Confusão envolvendo *x* e *s*, para o fonema /z/. Por exemplo, *esímio vs. exímio;*

Enquanto os erros anteriores são causados por a língua permitir que grafias diferentes tenham a mesma realização fonética, nos casos de erros que correspondem a palavras fonologicamente parecidas, a sua origem estará maioritariamente relacionada com deficiências da transmissão oral. Pode ser porque quem escreve ouviu mal, porque quem disse a palavra a pronunciou incorrectamente, ou pode acontecer a pronúncia de uma determinada região ser propensa à má compreensão de certas palavras.

Dos erros que tínhamos disponíveis, destacamos as seguintes confusões:

1. Entre prefixos ou grupos iniciais de grafemas. Em Estrela e Pinto-Correia (1988) aparecem referências apenas às confusões *des/dis*, *es/ex* e *per/pre*. A estas acrescentamos *en/in* e *por/pro*:

Confusão	Exemplos	
	Errado	Correcto
<i>des vs. dis</i>	<i>despêndio</i> <i>disfrutar</i>	<i>dispêndio</i> <i>desfrutar</i>
<i>en vs. in</i>	<i>enchado</i> <i>incómodo</i>	<i>inchado</i> <i>encómodo</i>

<i>es vs. ex</i>	<i>esclamou</i> <i>expresso</i>	<i>exclamou</i> <i>espesso</i>
<i>per vs. pre</i>	<i>perferível</i> <i>predidos</i>	<i>preferível</i> <i>perdidos</i>
<i>por vs. pro</i>	<i>portecção</i> <i>promenor</i>	<i>protecção</i> <i>pormenor</i>

2. Confusão entre duas letras. A mais frequente é a troca entre um *i* e um *e*, *chatiar* ou *creação*. Outras substituições frequentes são o *e* pelo *a*, como em *acelarar*, *farramenta* e *sociadade*, e a troca do *v* pelo *b*: *saliba*; *badagaio*.
3. Colocação de um *e* gráfico proveniente de um *e* oral originado por uma dissimilação, mas que segundo a ortografia oficial deve ser um *i*. Por exemplo, *femenino*, *defenição* (Estrela e Pinto-Ferreira, 1987).
4. Inserção da letra *i* em determinadas circunstâncias. Por exemplo, *azuleijo vs azulejo*; *decaiem vs. decaem*.
5. Omissão de um *e* ou *i* entre duas consoantes que facilmente se podem tornam num grupo consonântico. Por exemplo, *bactriana vs. bacteriana*; *bacharlato vs. bacharelato*; *utilizadores vs. utilizadores*. O inverso (inserção de uma vogal entre grupo consonântico) também ocorre. Por exemplo, *obececar vs. obcecar*.
6. Permutação de letras ou sílabas. Geralmente a permutação de letras envolve as letras *r* ou *s*, como em *cicratizar (cicatrizar)*, *largato (lagarto)*, *tartante (tratante)* e *compurscar (conspurcar)*. Um exemplo de troca de sílabas é *selinidade vs. senilidade*.

No corpo de erros linguísticos que tínhamos disponível (CorpoLing), constatámos que os erros devidos a semelhanças fonológicas ascendiam a 76.5% dos casos.

## 56 Incorreção morfológica

Nos erros originados por desconhecimento das regras morfológicas incluímos erros de flexão e erros motivados por deficiências em fenómenos de derivação e composição.

Na descrição dos erros originados por processos de derivação incluiremos todas as classes gramati-

cais. Nos erros originados por flexão incorrecta faremos a distinção entre o caso verbal e o caso nominal.

### **Derivação e composição**

Há várias causas para a ocorrência de erros ortográficos em palavras obtidas por derivação ou por composição. Geralmente, são erros relacionados com derivações que implicam a modificação do radical, como aglutinações ou justaposições mal efectuadas, e incorrecta flexão das palavras compostas.

Uma das fontes de erros na derivação é o processo de sufixação. Os erros ocorrem porque não é usado o sufixo correcto ou porque a ligação do sufixo com a palavra não é bem feita.

Alguns exemplos de sufixação incorrecta são:

<b>Incorrecto</b>	<b>Correcto</b>
<i>causadíaco</i>	<i>causídico</i>
<i>complementariedade</i>	<i>complementaridade</i>
<i>homogenizar</i>	<i>homogeneizar</i>
<i>sacrilegadamente</i>	<i>sacrilegamente</i>
<i>toráxico</i>	<i>torácico</i>

Dentro do processo de derivação, incluímos certos erros que, a nosso ver, ocorrem por influência de estrangeirismos. Por exemplo, a palavra *control*, que em português se escreve *controlo*. Outro exemplo será *kilómetro*, ou *kilograma*.

A derivação por prefixação é uma das maiores fontes de erros, geralmente associados ao uso do hífen para separar ou não o prefixo da palavra. A separação ou não do prefixo está relacionada com o prefixo em si, com a forma como termina o prefixo e como começa a palavra (Bergström e Reis,

1987; Costa e Melo, 1989). Assim, temos *semicircular* e *semi-homem*, *subgrupo* e *sub-reptício*.

Alguns erros deste tipo são:

<b>Incorrecto</b>	<b>Correcto</b>
<i>infraestrutura</i>	<i>infra-estrutura</i>
<i>autosuspenso</i>	<i>auto-suspenso</i>
<i>semi-eixo</i>	<i>semieixo</i>
<i>contra-prova</i>	<i>contraprova</i>
<i>tri-campeão</i>	<i>tricampeão</i>

Quando o prefixo é aglutinado à palavra, pode exigir algumas transformações para ficar de acordo com as regras da língua, ou não sofrer transformações a nível fonológico. Quando este pormenor é esquecido, acontecem os erros ortográficos. Evidenciamos aqui alguns destes erros:

<b>Incorrecto</b>	<b>Correcto</b>
<i>bisexto</i>	<i>bissexto</i>
<i>bisexual</i>	<i>bissexual</i>
<i>deshonra</i>	<i>desonra</i>
<i>desharmonia</i>	<i>desarmonia</i>

Em relação às palavras compostas, surgem alguns problemas como os de derivação por prefixação. Ou seja, por vezes compõem-se por aglutinação quando deveria ser por justaposição e vice-versa. Também acontece formarem-se palavras compostas que na realidade o não são.

Alguns exemplos de palavras compostas mal formadas são:

<b>Incorrecto</b>	<b>Correcto</b>
<i>pique-nique</i>	<i>piquenique</i>
<i>giradiscos</i>	<i>gira-discos</i>
<i>madre-silva</i>	<i>madressilva</i>
<i>via-satélite</i>	<i>via satélite</i>

Finalmente, temos o problema da flexão das palavras compostas. Aparecem erros como *médica-cirúrgica* (*médico-cirúrgica*), *médicos-cirúrgicos* (*médico-cirúrgicos*) e *poéticos-literários* (*poético-literários*).

### **Flexão nominal**

Em 1581 ocorrências de palavras com erro encontramos apenas 6 cujo erro era devido à má flexão, quer quanto ao género quer quanto ao número:

<b>Incorrecto</b>	<b>Correcto</b>
<i>cônjuga</i>	<i>cônjuge</i>
<i>despachanta</i>	<i>despachante</i>
<i>indivídua</i>	<i>indivíduo</i>

*imperadora*

*imperatriz*

*ilhoses*

*ilhós*

*vagãos*

*vagões*

---

As três primeiras palavras com erro correspondem à tentativa de formar o feminino de palavras que o não tem, ou que é igual ao masculino. O erro seguinte consiste na formação do feminino de forma regular, quando existe uma forma irregular.

Nos casos de flexão quanto ao número há uma tentativa de formação do plural de uma palavra que tem a mesma forma tanto no singular como no plural. O outro erro consiste na aplicação incorrecta da forma de pluralização das palavras terminadas em *ão*. Embora se possa formar assim o plural de algumas palavras terminadas em *ão*, este não é o caso.

### **Flexão verbal**

No que diz respeito à flexão verbal, os erros que aparecem estão fundamentalmente relacionados com clíticos e verbos irregulares. Vamos começar por estes últimos.

### **Formação incorrecta de formas regulares ou irregulares**

Nestes casos quem cometeu o erro tem uma noção aproximada do que pretende, mas não a escreve correctamente, ou generaliza a partir de outros verbos.

Por exemplo,

*creem, crêm (crêem)*

*dêm (dêem)*

*contêem (contêm)*

*falásteis (falastes)*

*consinteis (consintais)*

*queiremos (queiramos)*

*incluisse (incluísse)*

*constroi (constrói)*

*construido (construído)*

**Utilização de uma forma regular quando existe a irregular**

Quem escreve não tem conhecimento de que o verbo é irregular, usando uma forma construída como se o verbo fosse regular.

Por exemplo,

*deteu (deteve)*

*entretiam (entretinham)*

*intervido (intervindo)*

*odiem (odeiem)*

Os erros mais usuais que envolvem verbos com clíticos são:

**Colocação incorrecta das componentes do verbo cliticado**

Acontece com os verbos no condicional ou no futuro, em que os clíticos são colocados entre o verbo e a sua terminação.

Por exemplo,

*falariam-nos (falar-nos-iam)*

*poderia-me (poder-me-ia)*

**Confusão entre terminações de verbos e clíticos**

Existem terminações de verbos que são homófonas com clíticos, donde por vezes se escreva um verbo com clítico quando, na realidade, se pretendia escrever uma forma de um verbo. Por exemplo,

*faría-mos (faríamos); entráva-mos (entrávamos);*

*fizes-te (fizeste); consegui-se (conseguisse);*

*falas-te (falaste)*

Em certos casos, a forma verbal não existe (*faría, entráva, fizes*). Noutros, a forma verbal existe, mas o conjunto não tem sentido (*consegui-se*); noutros, ainda, existe a forma verbal e o conjunto tem sentido, só sendo possível detectar o erro através de análise sintáctica (*falas-te*).

Outra situação que pode ocorrer é o clítico aparecer como

terminação do verbo, constituindo uma forma sem sentido:

*darnos (dar-nos); mostreios (mostrei-os);  
referirnos-emos (referir-nos-emos)*

**Ligação errada entre a  
terminação do verbo e os  
clíticos**

Por exemplo, quando o complemento directo vem imediatamente a seguir ao verbo, por vezes é necessário modificar a terminação do verbo para receber harmoniosamente o clítico. Algumas vezes estas transformações dão origem a erro:

*fiz-lo (fi-lo); quis-lo (qui-lo); tem-o (tem-no);  
distribui-lo (distribuí-lo); distrai-lo (distraí-lo).*

**Utilização de formas  
regulares em vez das  
irregulares**

Esta situação ocorre mesmo sem a existência de clíticos, embora seja mais frequente quando eles estão presentes. Repare-se que é mais evidente o erro quando dizemos

*Eu trazerei o gato.*

do que quando dizemos

*Eu trazê-lo-ei.*

Temos os exemplos:

*entretreu-se (entreteve-se); manti-me (mantive-me);  
trazê-lo-ei (trá-lo-ei); trazer-nos-ia (trar-nos-ia);  
fazê-lo-ias (fá-lo-ias); dizer-te-ia (dir-te-ia);*

É de referir ainda que 50% as formas verbais com erro de flexão eram formas de verbos irregulares e que metade destes erros envolvia problemas de acentuação.

Os verbos que apresentavam mais erros deste tipo eram verbos terminados em *air (cair)* e *uir (construir)*.

## 57 Erros de acentuação

Os erros de acentuação são bastante vulgares no português. De facto, verifica-se que cerca de 20% dos erros de origem linguística envolve problemas de acentuação, daí prestarmos especial atenção a este tipo de erro.

Adicionalmente, é possível tirar partido de contagens relacionadas com este tipo de erros para fazer uma abordagem estatística da correcção de erros de acentuação. Para este efeito usámos o corpo de erros de origem linguística (CorpoLing).

Começamos por relacionar os tipos de erros de acentuação. Podem ser a colocação desnecessária do acento, a falta de acento ou a trocas de acentos. O quadro 3.3 mostra os valores encontrados para os diferentes tipos de erros de acentuação. A percentagem absoluta refere-se à percentagem do tipo de erro relativamente ao total de erros linguísticos analisados. A percentagem relativa refere-se à percentagem do tipo de erro relativamente ao total de erros de acentuação.

Surpreendem-nos os valores encontrados para as ocorrências de acento a mais. Esperávamos que fosse maior a percentagem de erros por omissão de acento.

<b>Tipo de erro</b>	<b>Número de ocorrências</b>	<b>Percentagem absoluta</b>	<b>Percentagem relativa</b>
Acento a mais	175	11.1	49.9
Falta de acento	158	10.0	45.0
Trocas de acentos	18	1.1	5.1
<b>Total</b>	<b>351</b>	<b>22.2</b>	<b>100.0</b>

Quadro 3.3 - Distribuição dos tipos de erro de acentuação.

Do ponto de vista da correcção ortográfica usando métodos estatísticos, interessa-nos caracterizar melhor a omissão de acentos do que as ocorrências de acentos a mais. Para corrigir um erro de acento a mais, basta encontrar a sua posição na palavra e removê-lo. No caso de omissão de acento, não só temos que saber onde o inserir, como temos que saber que acento colocar.

As trocas de acentos podem ser consideradas como a composição de duas operações: retirar um acento e inserir outro. Assim, as 18 trocas de acentos que vêm assinaladas no quadro 5.1 também contam para o total de omissões de acentos, que assim perfaz 176 ocorrências.

O quadro 3.4 dá uma panorâmica da forma como as omissões de acentos acontecem. Os valores vêm expressos em percentagens relativamente ao total de erros por omissão de acentos. As casas sombreadas correspondem a situações impossíveis no português.

<b>Acento</b>	<b>a</b>	<b>e</b>	<b>i</b>	<b>o</b>	<b>u</b>	<b>Total</b>
<b>Agudo</b>	9.1	16.5	40.9	14.8	7.4	88.6
<b>Circunflexo</b>	0.6	7.4		1.7		9.7
<b>Til</b>	0.6			0.6		1.1
<b>Grave</b>	0.6					0.6
<b>Total</b>	10.8	23.9	40.9	17.0	7.4	100.0

Quadro 3.4 - Distribuição das omissões de acentos em função das vogais e os acentos.

É na letra *i* que na maioria das vezes se omite o acento, que só pode ser o agudo. Em relação ao acento circunflexo, a maior parte das omissões acontece na letra *e*. Em termos de correcção ortográfica, e para o corpo de erros que usámos, se experimentássemos colocar o acento agudo em cada uma das vogais das palavras e o acento circunflexo somente nos *e*, conseguiríamos corrigir 96% dos erros por falta de acento.

## 58 Criação de regras

Tomando como base os erros de origem linguística até agora analisados, é possível criar heurísticas que ajudem a corrigir erros ortográficos. Essas heurísticas podem ser baseadas em regras de

reescrita, como em Nilsson (1971).

Todas as palavras que se enquadrem nos erros motivados por semelhanças fonéticas podem ser corrigidos usando regras de reescrita, desde que o erro não se encontre disperso pela palavra (como em *largato* ou *selinidade*).

Quanto aos erros originados pelo desconhecimento de regras morfológicas, a maior parte das vezes só poderão ser corrigidos recorrendo a algoritmos especializados para detecção e correcção desses erros. Por exemplo, para corrigir a palavra *fazer-se-ão* é necessário saber que se trata de uma forma do futuro do indicativo do verbo *fazer*, obter a sua forma irregular, substituindo de seguida a forma regular por esta.

Para decidir a sintaxe das regras de reescrita, tomámos o corpo de erros de origem linguística e analisámos todas as ocorrências de palavras com erro, comparando-as com a palavra correcta. Cada par foi rotulado com uma pequena regra de reescrita que permitisse identificar o tipo de erro dado. Por exemplo, a palavra *abstrato* tem correcção *abstracto*. Este par foi rotulado com a etiqueta "t-ct".

Depois de todas as palavras rotuladas, reunimos sistematicamente todos os pares de palavras que tivessem o mesmo rótulo e tentámos construir regras que fossem o mais abrangentes possível.

Por exemplo, com o rótulo anterior encontrámos 20 ocorrências. Tomando desses pares apenas a palavra com erro e alinhando-as pela posição do erro obtemos:

abstrato  
arquiteto  
atriz  
bissetriz  
carateres  
coletivizac<sup>^</sup>a<sub>~</sub>o  
dejetos  
letivo  
olfativo  
perspetiva  
refletir  
reto  
setor  
setores  
sinta'tico  
ta'tica  
ta'tico  
trajeto  
trajeto'ria  
veredito

Utilizámos aqui o formato usado pelo Palavroso para a representação de caracteres acentuados (vd. secção 2.2.1), já que isso tem influência na análise dos erros.

Destas vinte palavras, em quinze casos o *t* encontra-se ladeado de vogais. Em três casos é precedido

de acento agudo e em dois é seguido de *r*. Estes três casos poderiam ser expressos da seguinte forma:

$$*\{vogal1\}t\{vogal2\}* \quad \Rightarrow \quad *\{vogal1\}ct\{vogal2\}*$$
$$*\{vogal1\}'t\{vogal2\}* \quad \Rightarrow \quad *\{vogal1\}'ct\{vogal2\}*$$
$$*\{vogal1\}tr\{vogal2\}* \quad \Rightarrow \quad *\{vogal1\}ctr\{vogal2\}*$$

Em que o asterisco substitui qualquer sequência de zero ou mais caracteres e {vogal1} e {vogal2} substituem uma vogal.

Poderíamos tentar reunir duas regras numa:

$$*t\{vogal2\}* \quad \Rightarrow \quad *ct\{vogal2\}*$$

Desta forma já não precisávamos da segunda regra, com o acento agudo.

Fazer as regras muito genéricas pode ser desejável. Se por um lado cobrem mais situações de erro, por outro, são aplicadas mais vezes sem qualquer sucesso.

A cada regra assim criada fica associado um determinado número de palavras com erro às quais a regra se aplica. Decidimos não criar qualquer regra se esse número fosse inferior a quatro. A razão da escolha deste valor está relacionada com a cobertura das regras: cerca de 85% das palavras do corpo de erros de origem linguística são abrangidas pelas regras que se aplicam a quatro ou mais palavras.

Ao criar regras que são aplicadas menos de quatro vezes, corre-se o risco de estar a sobrecarregar demasiado o processamento, sem que isso implique uma melhoria significativa no desempenho do corrector.

As palavras que não são abrangidas por estas regras heurísticas ou têm um tratamento especial, como os verbos com clíticos, ou têm de ser processadas usando as técnicas dos erros tipográficos, como se de tal se tratassem.

## 59 Erros tipográficos

O corpo de teste dos erros tipográficos é composto por uma lista de 312 palavras. Como as regras heurísticas não cobrem todos os erros linguísticos, alguns erros linguísticos vão ser tratados como tipográficos. Estas palavras ascendem a 223 ocorrências.

Temos então duas fontes de erros para analisar. Como têm origens diferentes, vamos estudá-los em separado, mas segundo os mesmos critérios, já que vão estar sujeitos ao mesmo tipo de processamento. Começaremos pelos erros tipográficos, analisando-os pormenorizadamente. Depois, faremos uma breve análise dos erros de origem linguística, tratando-os como se fossem de origem tipográfica.

Para além dos quatro tipos de erros conhecidos, omissão, inserção, transposição e substituição, criámos mais dois grupos: erros complexos e erros por omissão de espaço separador entre palavras. Os erros complexos são aqueles em que é necessário efectuar mais do que uma das quatro operações simples para corrigir o erro. Por exemplo, *efificio*, ou *desenvimento*. No primeiro caso é necessário substituir o primeiro *f* e acrescentar um acento agudo no segundo *i*. No segundo caso é necessário inserir três caracteres.

Um exemplo de erro de espaço é: *setiveres*, que deveria ser substituído por *se tiveres*.

## 60 Distribuição dos tipos de erro

O quadro 3.5 mostra a forma como os erros tipográficos se distribuem quanto ao tipo de erro.

Os valores encontrados vêm confirmar os que aparecem na literatura (Damerau, 1964), nomeadamente que 80% dos erros são causados por uma das quatro operações, sendo os outros 20% de origem mais complexa. Nestes 20% incluímos os erros originados por uma falta de espaço.

Origem dos erros	Nº de ocorrências	Percentagem
Omissão	88	28.2
Transposição	68	21.8
Inserção	54	17.3

Substituição	40	12.8
Complexo	36	11.6
Falta de espaço	26	8.3
<b>Total</b>	<b>312</b>	<b>100.0</b>

Quadro 3.5 - Distribuição dos erros tipográficos quanto ao tipo.

Na essência, o erro de falta de espaço é o mesmo de omissão. Exige, contudo, um tipo de manipulação diferente. Repare-se que nos outros casos temos uma palavra errada, fazemos uma determinada operação e obtemos outra palavra, que pode estar correcta, ou não. Neste caso de omissão, quando inserimos um espaço, estamos a produzir duas palavras, que podem estar ou não correctas.

Assim sendo, somente 11.6% dos erros tipográficos exigem mais do que uma operação simples para os corrigir. Isto é, apenas 11.6% das palavras com erros tipográficos se encontram a uma distância de edição superior a 1 da palavra correcta.

Ordenando os tipos de erro por ordem decrescente da sua ocorrência, obtemos uma ordem idêntica à encontrada por Pollock e Zamora (1984), descrita no quadro 3.6.

<b>Origem</b>	<b>Ordenação</b>
Nosso corpo	Omissão > Transposição > Inserção > Substituição > Complexo > Espaço
Pollock e Zamora	Omissão = Transposição > Inserção > Substituição

Quadro 3.6 - Ordem da frequência de ocorrência dos erros, segundo duas fontes.

É curioso comparar o tipo de erro com o comprimento médio das palavras onde esse erro ocorre. O quadro 3.7 contém os valores do tamanho médio das palavras, para cada tipo de erro.

Os dados apresentados reportam-se às palavras com erro e não às suas correcções. Quer isto dizer que pode realmente haver diferenças no tamanho das palavras com erro. As transposições e as substituições conduzem a palavras com o comprimento igual ao da

palavra correcta. As inserções conduzem a palavras maiores. As omissões conduzem a palavras menores.

Consultando o quadro 3.7 constatamos que as substituições, transposições e inserções ocorrem em palavras sensivelmente com o mesmo tamanho médio, entre 7.7 e 7.9 caracteres. Os casos interessantes são os restantes três: as omissões, os erros complexos e as faltas de espaço.

<b>Origem dos erros</b>	<b>Média</b>
Omissão	8.9
Substituição	7.9
Inserção	8.9
Transposição	7.7
Falta de espaço	8.4
Complexo	10.3

Quadro 3.7 - Distribuição dos tamanhos das palavras com erro em função do tipo de erro.

As omissões têm uma média que se evidencia por si. Se tivermos em conta o facto de estarmos a trabalhar com omissões, verificamos que o comprimento médio das palavras que originam este tipo de erro sobe para 9.9 caracteres.

O facto das palavras com erro complexo terem em média 10.3 caracteres não é de todo surpreendente. São as palavras compridas que exigem maior atenção ao serem dactilografadas, sendo por isso mais sujeitas a erros por distração.

Surpreendente é, sem dúvida, o tamanho médio das palavras em que falta um espaço. Se falta um espaço, temos, na realidade, duas palavras justapostas. Logo, a nossa intuição dir-nos-ia que a palavra resultante deveria ter um comprimento maior. Contrariamente, as palavras resultantes deste processo têm um comprimento idêntico ao das outras palavras com erro. Daqui concluímos que este tipo de erro, geralmente, envolve palavras pequenas, quer sejam duas palavras pequenas que se aglutinam, quer seja uma palavra pequena que aparece ligada a uma maior.

## 61 Posição do erro

Por vezes, os autores de correctores ortográficos comentam que os erros tipográficos são pouco frequentes no início das palavras (Agirre et al., 1992). Embora isso possa ser verdade, o conceito de início de palavra não está bem definido.

Procurámos caracterizar as ocorrências de erros quanto a este aspecto. Não o fizemos só para as ocorrências no início, mas tentámos saber qual era a distribuição global. Excluímos desta análise os erros de falta de espaço, pois à partida sabemos que têm uma distribuição própria, com predominância para os extremos.

Inicialmente definimos três formas de divisão das palavras:

1. Divisão de motivação computacional. No Correcto, o processamento dos erros de origem tipográfica é diferente conforme este se dê na primeira, na última, ou em qualquer outra posição da palavra (cf. secção 5.3.7). Portanto, considerámos como início da palavra o primeiro carácter, o fim o último carácter e o meio todo o resto.
2. Dividir equitativamente a palavra por início, meio e fim: o início corresponderia ao primeiro terço da palavra, o meio ao segundo terço e o fim ao terceiro terço da palavra.
3. Tomar como início os três primeiros caracteres da palavra. Da parte restante, os três últimos caracteres seriam o final da palavra e tudo o que restasse seria o meio.

O quadro seguinte reporta os valores totais encontrados para cada um dos métodos:

<b>Método</b>	<b>Início</b>	<b>Meio</b>	<b>Fim</b>	<b>Total</b>
<b>1</b>	11	246	30	287
<b>2</b>	70	152	65	287
<b>3</b>	89	136	62	287

O método **1** é muito específico quanto aos extremos. Deixa uma zona demasiado grande para o meio da palavra, caindo nesta zona a maioria dos casos.

Os métodos **2** e **3** são relativamente parecidos. Os resultados de um não acrescentam nada em relação ao outro. Como este último é computacionalmente mais simples, adoptaremos este para apresentar os resultados.

Assim, a distribuição obtida para as zonas onde ocorre o erro é a que vem descrita na figura 3.8.

μ §

Figura 3.8 - Distribuição dos erros tipográficos relativamente à sua posição na palavra.

O comprimento médio de todas as palavras analisadas é 8.6 caracteres. Como estabelecemos (método **3**) que o início é constituído pelos primeiros três caracteres, constituindo os três últimos o fim da palavra, restam, em média, 2.6 caracteres para o meio da palavra. Assim, é significativo o valor encontrado para o número de vezes que o erro ocorre no meio da palavra: 47%.

Se agora tirarmos partido dos resultados obtidos pelo método **1** e assumirmos que as palavras com erro têm comprimento médio de 9 caracteres (arredondamento do valor verdadeiro, de 8.6), então a distribuição dos erros atendendo à sua posição na palavra será, em média, a da figura 3.9.

Figura 3.9 - Distribuição fina dos erros em função da posição na palavra.

## 62 Distância no teclado

Existe uma ideia generalizada de que os erros de inserção e substituição de caracteres ocorrem principalmente com letras bastante próximas no teclado. Procuramos aqui verificar esta teoria.

Para isso definimos quatro distâncias:

Distância Zero - Quando se trata de um erro de repetição da mesma tecla, como em *possível*.

Distância Um - Quando o erro envolve as teclas que se encontram contíguas à letra correcta, na mesma fila, ou quando o erro envolve a mesma tecla diferindo por uma situação de minúscula/maiúscula.

Distância Dois - Quando se tratam de teclas que se encontram na vizinhança imediata da tecla correcta, mas na linha acima ou abaixo.

Distância Infinita - Qualquer outro caso.

Figura 3.10 - Esquema de parte de um teclado.

No desenho da figura 3.10, as teclas [F] e [H] encontram-se à distância Um da tecla [G]. As teclas [T], [Y], [V] e [B] encontram-se à distância Dois e as teclas [R] e [C] encontram-se a distância Infinita de [G].

Nem todos os teclados são iguais. Embora possa haver partes comuns à maioria dos teclados, existem diferenças significativas nas teclas com sinais de pontuação, acentos e outros símbolos que não sejam letras do alfabeto nem números.

Para manter este estudo o mais genérico possível, analisámos apenas erros que envolviam letras do alfabeto, aplicando-se, portanto, a qualquer teclado cuja disposição das teclas seja QWERTY. O quadro 3.11 mostra os valores encontrados para uma contagem efectuada num conjunto de 88 pala-

vras com erros de inserção e substituição, satisfazendo a condição anterior.

<b>Tipo de erro</b>	<b>Distância 0</b>	<b>Distância 1</b>	<b>Distância 2</b>	<b>Distância <math>\infty</math></b>
Inserção	17	12	4	21
Substituição	0	18	1	15
Total	17	30	5	36

Quadro 3.11 - Distâncias encontradas entre as teclas que originam erros de inserção e substituição.

No total, 41% dos erros têm origem em teclas distantes, sendo 59% provocados por teclas próximas. No caso das inserções esta diferença é ligeiramente acentuada: encontramos 61% de casos de proximidade e 39% dos casos são de longa distância.

Na generalidade podemos aceitar que a maioria dos erros de inserção ou substituição são causados por bater numa tecla próxima da tecla correcta.

## **63 Erros de origem linguística considerados como tipográficos**

Uma parte dos erros de origem linguística não é abrangido pelas regras heurísticas, havendo a necessidade de ser corrigida por métodos orientados para a correcção de erros tipográficos. Por esta razão, tomámos uma lista de palavras nestas condições e procurámos analisá-las como se de erros tipográficos se tratassem.

O quadro 3.12 contém resultados relativos à distribuição dos erros segundo o seu tipo. Repare-se no contraste de valores entre os erros linguísticos tratados como tipográficos e os verdadeiros erros tipográficos (quadro 3.5).

Por um lado verifica-se que os erros complexos aumentam consideravelmente. Por outro lado, constata-se que este valor é relativamente pequeno: só 30% dos erros de origem linguística não podem ser corrigidos por uma operação de edição simples.

<b>Origem dos erros</b>	<b>Nº de ocorrências</b>	<b>Porcentagem</b>
Omissão	42	18.8
Substituição	69	30.9
Inserção	37	16.6
Transposição	9	4.0
Complexo	66	29.6
<b>Total</b>	<b>223</b>	<b>100.0</b>

Quadro 3.12 - Distribuição dos erros linguísticos tratados como tipográficos.

A grande diferença entre os dois tipos de erros reside no número de ocorrências dos vários tipos de erro e respectiva ordem. No quadro 3.13 estão relacionadas as duas origens dos erros quanto ao tipo de erro e sua ordem.

<b>Origem</b>	<b>Ordenação</b>
Erros tipográficos	Omissão > Transposição > Inserção > Substituição > Complexo > Espaço
Erros linguísticos	Substituição ≈ Complexo > Omissão > Inserção >> Transposição

Quadro 3.13 - Relação entre erros de origem linguística e erros de origem tipográfica, quanto ao tipo de erro.

Em relação à posição do erro na palavra, não há grandes diferenças. A tendência é haver mais erros

no meio das palavras do que nos extremos, tal como acontecia com os erros genuinamente tipográficos. Isto leva-nos a pensar que esta será uma característica dos erros ortográficos em geral e não uma característica específica dos erros de origem tipográfica.

<b>Zona</b>	<b>Início</b>	<b>Meio</b>	<b>Fim</b>
<b>Percentagem</b>	32	45	23

Quadro 3.14 - Distribuição dos erros de origem linguística em relação à zona da palavra

Também a distância das letras erradas no teclado (quadro 3.15) é bastante diferente entre os erros de origem linguística e os erros de origem tipográfica. Aqui, predominam as grandes distâncias: 66% dos erros são deste tipo. Este valor é bastante importante, pois tem tendência completamente contrária à encontrada para os erros de origem tipográfica. Isto permite-nos aceitar que realmente existe uma relação entre os erros tipográficos (de inserção e substituição) e a distância no teclado entre as letras envolvidas no erro.

<b>Tipo de erro</b>	<b>Distância 0</b>	<b>Distância 1</b>	<b>Distância 2</b>	<b>Distância ∞</b>
Inserção	5	13	3	14
Substituição	0	7	8	56
<b>Total</b>	<b>5</b>	<b>20</b>	<b>11</b>	<b>70</b>

Quadro 3.15 - Número de ocorrências para cada tipo de distância entre teclas envolvidas em erros de substituição e inserção.

Resumindo, existem grandes diferenças entre erros de origem tipográfica e origem linguística, quando analisamos estes últimos pela perspectiva dos erros tipográficos. Se os resultados aqui obtidos forem usados para melhorar o desempenho de um corrector no tratamento de erros tipográficos e se parte dos erros de origem linguística forem tratados como erros tipográficos, coloca-se a seguinte questão: o processamento estatístico envolvido na geração de sugestões de palavras com erros tipográficos não deveria ser ponderado pelos valores encontrados para os erros de origem linguística?

Segundo Pollock e Zamora (1984), e no caso da língua inglesa, apenas 15% dos erros ortográficos são de origem linguística. Observando que as regras heurísticas cobrem pelo menos 85% destes erros, apenas 2.25% dos erros ortográficos são de origem linguística e processados como erros tipográficos. No cômputo geral, as regras de processamento de erros de origem tipográfica são aplicadas a palavras com erros deste tipo 97.4% das vezes, contra 2.6% das vezes a palavras com erro de origem linguística.

Concluindo, não nos parece que estes valores mereçam que se faça tal ponderação.

## 64 Processamento de trigramas

Uma vez que o Correcto tira partido de processamento trigrâmico, achamos adequado fazer algum tipo de caracterização, a nível de trigramas, das palavras com erro.

Contando com as 26 letras do alfabeto, mais onze vogais acentuadas (*á, à, â, ã, é, ê, í, ó, ô, õ, ú*), o ç, o espaço e o hífen, temos um conjunto de 40 grafemas que se podem combinar para formar trigramas. Com este conjunto podemos então formar, no máximo,  $\mu$  § trigramas. Mas destes, nem todos são possíveis. Uns devido à forma como extraímos a informação e outros porque não fazem parte da língua. Por exemplo, nunca poderão aparecer trigramas com um espaço no meio ou dois espaços seguidos, pois extraímos os trigramas de palavras isoladas. Também se constata que não fazem parte do português trigramas como *sch, lsd* e *wrt*.

Para o processamento dos trigramas necessitamos saber, por um lado, quais são os trigramas possíveis na língua e, por outro lado, qual a frequência com que estes trigramas aparecem.

Para o primeiro objectivo podemos usar qualquer conjunto de palavras do português; quanto mais palavras diferentes, maior será a cobertura dos trigramas existentes. Usámos para este efeito um corpo de teste do qual extraímos cerca de 18 000 formas distintas do português, mais todos os dicionários do Palavroso, perfazendo 63 414 formas distintas.

Para o segundo objectivo interessa contar os trigramas tal como aparecem em texto. Usámos para esse efeito um corpo de teste de cerca de 155 000 ocorrências de formas.

Obtivemos assim um conjunto de 7 277 trigramas possíveis para o português, com a respectiva frequência de ocorrência. Este valor representa cerca de 11% do número de trigramas teoricamente possíveis, para o conjunto de grafemas com que estamos a trabalhar. Destes trigramas destacamos no quadro 3.16 os mais frequentes.

---

**Os trigramas mais frequentes**

---

---

---

_de	_qu	_do	con
os_	ent	_es	ar_
de_	da_	nte	es_
as_	ue_	est	_pr
_co	em_	_um	_ca
_a_	ra_	com	ia_
ão_	te_	_pa	_ma
_o_	_se	_da	ma_
do_	_e_	_po	_te
que	to_	se_	ção

---

Quadro 3.16 - Listas dos trigramas mais frequentes na língua portuguesa. O mais frequente é o do canto superior esquerdo, decrescendo a sua frequência ao longo das colunas.

Como seria de esperar, a frequência dos trigramas está bastante relacionada com a frequência das palavras, com especial ênfase para as palavras gramaticais. Por exemplo, entre os primeiros lugares

aparecem os trigramas que constituem a preposição *de*, bem como as partículas *a*, *o*, *e*.

O número de trigramas conseguido é, de algum modo, representativo. Mesmo recorrendo a corpos de texto maiores, dificilmente se conseguirá alcançar um número significativamente superior. Seria, contudo, possível melhorar os valores obtidos para as frequências dos trigramas. Note-se que cerca de 30% dos trigramas (2299 ocorrências) só ocorreram uma vez.

### Trigramas de frequência zero

Usando estes trigramas, contamos as vezes em que uma palavra com erro continha um trigrama de frequência zero. Para estes casos, contamos o número de vezes que esse trigrama assinalava a posição onde se encontrava o erro. Os resultados vêm expressos no quadro 3.17.

	<b>Tipográficos</b>	<b>Linguísticos</b>
Número de palavras com erro	265	1580
Porcentagem de palavras com algum trigrama zero	36%	17%
Porcentagem de vezes que o trigrama de frequência zero indicava a posição do erro	100%	98%

Quadro 3.17 - Porcentagens das vezes que uma palavra com erro contém um trigrama de frequência zero e das vezes que esse trigrama indica a posição do erro.

Se assumirmos que a frequência de ocorrência de um trigrama ser zero significa que este não é válido na língua em questão, nos erros de origem linguística não aparecem tantos trigramas inexistentes. Tal já era esperado.

Contudo, não deixaram de ser inesperados os valores encontrados para a porcentagem de vezes que um trigrama de frequência zero indica a posição exacta onde se encontra o erro. Esperaríamos que um maior número de ocorrências de trigramas de frequência zero fosse devido a uma deficiente cobertura dos trigramas.

## 65 Abordagens utilizadas

A correcção ortográfica apresenta duas questões de fundo:

1. O dicionário e a forma como se faz a verificação ortográfica.
2. A detecção do erro cometido, sugerindo depois a palavra correcta.

Embora se possa construir um dicionário usando uma simples lista de palavras e resumir a verificação ortográfica a uma pesquisa binária nessa lista, essa não é a melhor maneira de o fazer. Pelo menos para o português, devido ao elevado número de formas que se obtém das palavras por um processo de flexão. Para evitar a colocação de todas as formas no dicionário, é necessário dotar o acesso ao dicionário com um processo de análise morfológica, com toda a complexidade daí resultante.

A detecção do erro cometido, e conseqüente sugestão da palavra correcta, nem sempre é um caso simples. Pode haver várias palavras candidatas à correcção, podendo todas elas ser correcções bastante plausíveis.

Suponhamos que a palavra errada era *cnto*. As palavras *canto*, *conto*, *cento*, *cinto*, *unto*, *coto*, *cato*, *cito* são oito possíveis correcções. As quatro últimas parecem pouco prováveis, já que nas substituições que são feitas em nenhum caso a letra que se substitui se encontra no teclado próxima da letra *n*. Das quatro palavras que ficam, dificilmente se conseguirá encontrar um critério baseado em palavras isoladas que nos permita escolher uma palavra mais provável.

É essencialmente com este aspecto da correcção ortográfica que nos preocuparemos daqui em diante: a construção de uma lista de sugestões e a escolha da correcção mais provável entre elas.

Ao longo dos anos têm aparecido várias propostas no sentido de resolver este problema. Geralmente os métodos propostos são orientados para um determinado tipo de erro – linguístico ou tipográfico – descurando o outro. Como os erros tipográficos tendem a aparecer em maior número que os erros de origem linguística, também os métodos de correcção tendem a ser orientados para os erros tipográficos.

Por outro lado, os erros linguísticos deixam pior impressão no leitor do que os erros tipográficos. Por esta razão alguns autores apostam mais nos métodos de correcção orientados para os erros linguísticos.

Um bom corrector ortográfico terá capacidade para lidar tão bem com um tipo de erros, como com o outro.

No seguimento desta secção iremos descrever algumas abordagens conhecidas. Não pretendemos ser exaustivos. A ideia é dar uma panorâmica geral para situar o nosso trabalho. O critério de

escolha das abordagens aqui apresentadas regulou-se principalmente pelo pioneirismo e pela inovação.

## 66 Métodos de correcção básicos

Se 80% dos erros encontrados em textos são erros simples, ao fazer um corrector ortográfico que trate só este tipo de erros, à partida já se esperam resultados satisfatórios. Por outras palavras, com relativamente pouco trabalho é possível construir um corrector ortográfico que dê bons resultados. Esses resultados serão tanto mais impressionantes quanto menos os utilizadores estiverem habituados a usar correctores ortográficos.

Existem dois trabalhos clássicos nesta área, que se baseiam única e exclusivamente na detecção de erros tipográficos simples. Vêm descritos em Damerau (1964) e Peterson (1980) e são, ainda hoje, ponto de referência para muitos autores.

Ambos os correctores se baseiam em encontrar palavras cuja distância de edição à palavra com erro não seja superior a uma unidade. A diferença reside na forma como o fazem.

## 67 Corrector de Damerau

O programa de Damerau baseava-se essencialmente num método de pesquisa em dicionário. Para facilitar a pesquisa, o dicionário era inicialmente preparado, fazendo associar a cada palavra o seu comprimento e um código que descrevia o conjunto de letras que era usado na palavra, independentemente de serem repetidas ou não. Assim, cada entrada no dicionário era composta por três campos: a palavra, o seu comprimento e o código das letras.

O código de letras consistia numa sequência de dígitos binários, que eram estabelecidos de forma idêntica à que de seguida se ilustra<sup>1</sup>.

Por exemplo, vamos estabelecer um código para a palavra *efectivamente*. Então a sequência binária vem:

1	0	1	0	1	1	0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z

Se o dígito binário é 1, então a letra homóloga está presente na palavra. Se for zero, a letra não se

<sup>1</sup> Alguns pormenores da descrição original foram propositadamente omitidos ou modificados para tornar o exemplo mais simples e claro.

encontra na palavra.

Representando esta sequência binária como um inteiro na base decimal, obtemos o valor 45232208, que representa, no nosso exemplo, o código de *efectivamente*.

As palavras a serem testadas no dicionário também são sujeitas a este processamento. Mas nem todas: as palavras com 3 caracteres, ou menos, nunca são verificadas. As palavras com 4 e 5 caracteres são verificadas primeiro numa lista de palavras mais frequentes. Só se não forem lá encontradas é que são sujeitas ao processamento indicado, para se poder efectuar a pesquisa no dicionário.

A pesquisa não é feita sempre do mesmo modo. Difere de quando se está a verificar se a palavra existe no dicionário, ou quando se está à procura de palavras parecidas.

A verificação no dicionário é feita da seguinte forma:

O tamanho da palavra é comparado com o tamanho da entrada no dicionário.

Se são iguais, e se os seus códigos também o forem, então compara as duas palavras letra por letra, até encontrar um par que difira, ou concluir que as palavras são iguais.

Se os tamanhos, os códigos ou as palavras não são iguais, passa-se à entrada seguinte no dicionário, até se encontrar a palavra, ou atingir-se o final do dicionário.

Se alguma palavra não foi encontrada no dicionário, então inicia-se o processo de pesquisa de palavras parecidas, assumindo um dos quatro erros tipográficos simples.

Se a diferença entre o tamanho da palavra e o tamanho da entrada no dicionário for superior a 1, ou se for superior a dois o número de bits diferentes entre código da entrada do dicionário e o código da palavra, então passa-se à palavra seguinte. Senão fazem-se os testes descritos no quadro 3.18.

Se em algum dos casos descritos no quadro 3.18 as palavras não coincidirem depois de efectuada a operação indicada, a pesquisa continua.

O autor não é explícito acerca do número de sugestões geradas. No entanto, fica-se com a ideia que o processo pára quando encontra a primeira palavra no dicionário, tomando automaticamente essa palavra para substituir a que se considera conter erro.

A razão para escolher a primeira palavra que encontra prende-se com o facto de o corrector ter sido usado num domínio muito específico (*coordinate indexing and retrieval system*) e de usar um dicionário bastante reduzido: o número máximo de palavras que o dicionário podia ter era 5000. Com este número de entradas, dificilmente se encontrarão no dicionário palavras parecidas, pelo menos que difiram entre si apenas por uma letra. Assim, existe uma forte possibilidade de a primeira palavra que se encontrar no dicionário e verificar que alguma das condições anteriores realmente seja a palavra correcta.



	<b>Descrição</b>	<b>Exemplo</b>
<b>Situação</b>	As palavras têm o mesmo tamanho e diferem apenas numa letra.	<b>Dic:</b> EFECTIVAMENTE <b>Pal:</b> EFRCTIVAMENTE
<b>Operação</b>	---	
<b>Conclusão</b>	Considera-se que houve um erro de substituição, tomando a palavra no dicionário como correcção.	
<b>Situação</b>	As palavras têm o mesmo tamanho e diferem em duas letras contíguas.	<b>Dic:</b> EF <u>E</u> CTIVAMENTE <b>Pal:</b> EFC <u>E</u> TIVAMENTE
<b>Operação</b>	As letras da palavra são transpostas e as palavras comparadas novamente.	<b>Pal:</b> EF <u>E</u> CTIVAMENTE
<b>Conclusão</b>	Se as palavras forem iguais, ocorreu um erro de transposição, sendo a palavra no dicionário tomada como correcção.	<b>Dic:</b> EF <u>E</u> CTIVAMENTE <b>Pal:</b> EF <u>E</u> CTIVAMENTE
<b>Situação</b>	A palavra com erro é mais comprida um carácter do que a palavra do dicionário.	<b>Dic:</b> EF <u>E</u> CTIVAMENTE <b>Pal:</b> EF <u>E</u> FECTIVAMENTE
<b>Operação</b>	A primeira letra que difira com a entrada no dicionário é eliminada.	<b>Pal:</b> EFECTIVAMENTE
<b>Conclusão</b>	Se as duas palavras coincidirem, houve um erro de inserção. A palavra do dicionário é a correcção.	<b>Dic:</b> EFECTIVAMENTE <b>Pal:</b> EFECTIVAMENTE
<b>Situação</b>	A palavra do dicionário é mais comprida um carácter do que a palavra.	<b>Dic:</b> EF <u>E</u> CTIVAMENTE <b>Pal:</b> EF <u>E</u> TIVAMENTE
<b>Operação</b>	A primeira letra em que diferem é eliminada da entrada no dicionário.	<b>Dic:</b> EFETIVAMENTE
<b>Conclusão</b>	Se as palavras coincidirem, houve um erro de omissão e toma-se a palavra do dicionário	<b>Dic:</b> EFETIVAMENTE <b>Pal:</b> EFETIVAMENTE

	como correcção.	
--	-----------------	--

Quadro 3.18 - Conjunto de operações e comparações efectuadas entre uma palavra do texto e uma entrada do dicionário.

Prós	Contras
<ul style="list-style-type: none"> <li>∇ Simplicidade do sistema</li> <li>∇ Adequado para um domínio específico</li> </ul>	<ul style="list-style-type: none"> <li>∇ Só processa erros tipográficos simples.</li> <li>∇ Todas as palavras do dicionário têm que lá estar de forma explícita.</li> <li>∇ Se for alargado a domínios gerais, pode originar um processamento demasiado pesado.</li> </ul>

## 68 Corrector de Peterson

Peterson (1980) faz um apanhado de vários métodos de verificação ortográfica, sugere um método para fazer correcção e aponta algumas direcções a seguir no sentido de o melhorar.

No âmbito da correcção, sugere um método estatístico, baseado em digramas e trigramas, que ordena as palavras por ordem decrescente de um factor de peculiaridade, calculado a partir da frequência com que os digramas e trigramas aparecem no texto. Quanto mais alto esse nível, maior a probabilidade de a palavra conter algum erro. Cada uma das palavras deveria ser depois analisada manualmente.

Num nível mais avançado de sofisticação aponta os métodos baseados em dicionário. No actual estado da arte, já não se põe outra hipótese: é ponto assente que deve ser baseado num dicionário.

Em termos de correcção, o método que sugere é baseado nos quatro erros típicos. No entanto, não é tão dependente da estrutura do dicionário como o de Damerau.

Genericamente, o método baseia-se em, partindo da palavra com erro, construir todas as palavras que, através de um dos quatro erros típicos, poderiam ter originado a palavra em questão. Depois, cada item desta lista de palavras é verificado no dicionário. Dá-se ao utilizador, como lista de sugestões, apenas as palavras geradas que existirem no dicionário.

Tomando como exemplo a palavra *rais*, o quadro 3.19 mostra todas as palavras que o algoritmo origina e, dessas, as que existem no dicionário.

Nas posições onde aparece um asterisco, a palavra pode tomar qualquer letra. Verifica-se que há 241 palavras que, devido a um dos quatro erros tipográficos típicos simples, poderiam ter originado a palavra com erro *rais*. Dessas, apenas 24 são palavras pertencentes à língua.

<b>Processo</b>	<b>Palavras que a possam originar</b>	<b>Palavras existentes em dicionário</b>
<b>Transposição</b>	aris, rias, rasi	rias
<b>Inserção</b>	ais, ris, ras, rai	ais, ris
<b>Substituição</b>	*ais, r*is, ra*s, rai*	cais, dais, mais, pais, saís, tais, vais, reis, raio, raie, raia, raiz
<b>Omissão</b>	*rais, r*ais, ra*is, rai*s, rais*	trais, irais, orais, arais, reais, riais, raies, raios, raias

Quadro 3.19 - Palavras geradas a partir da palavra *rais* e, dessas, as palavras que existem no dicionário.

Independentemente do número de palavras que fazem ou não parte da língua, seriam necessárias 241 consultas ao dicionário. O próprio autor reconhece que se trata de um processo computacionalmente pesado, mas por outro lado simples.

Para minimizar este problema, o autor apresenta algumas sugestões, tais como usar frequências de digramas e trigramas, entrar em consideração com a forma como as teclas se dispõem num teclado, ou ainda usar uma representação fonética das palavras.

<b>Prós</b>	<b>Contras</b>
<ul style="list-style-type: none"> <li>∇ Conceptualmente simples</li> <li>∇ Existe separação entre o algoritmo de geração de sugestões e a sua verificação em dicionário.</li> </ul>	<ul style="list-style-type: none"> <li>∇ Só processa erros tipográficos simples.</li> <li>∇ Computacionalmente pesado, principalmente para palavras compridas.</li> </ul>

## 69 Chaves de semelhança e abreviaturas

Pollock & Zamora (1984) apresentam um método para corrigir erros ortográficos, que se baseia em criar chaves (as chaves de semelhança) que propositadamente distorcem a palavra, mas retêm dela o fundamental.

Embora interessante, esta ideia não é de todo original. Já Blair, citado por Damerou (1964), havia sugerido uma ideia semelhante: cada palavra do dicionário era abreviada por quatro caracteres. Se coincidissem com alguma abreviatura existente, era abreviada por cinco caracteres e assim sucessivamente.

A palavra incorrectamente escrita também é abreviada, resumindo-se a procura de palavras semelhantes à comparação de abreviaturas.

No caso de Pollock e Zamora, as chaves de semelhança são construídas usando processos diferentes dos de Blair. Também é usada uma medida de distância entre as chaves do dicionário e as chaves construídas a partir da palavra com erro.

São usados dois tipos de chaves de semelhança: a chave-esqueleto e a chave de omissão.

A primeira constrói-se colocando a primeira letra da palavra, seguida das restantes consoantes. Por fim, aparecem as vogais restantes, pela ordem que aparecem, sem repetições. Por exemplo, a chave-esqueleto da palavra *interdisciplinar* será intrdscpleia.

Para a chave de omissão, entra-se com a frequência com que as consoantes são omitidas. Assim, segundo os autores, na língua inglesa as consoantes são omitidas pela seguinte ordem, da mais frequente para a menos frequente: RSTNLCHDPGMFBYVWZXXQKJ. A chave de omissão obtém-se colocando as consoantes da palavra por ordem crescente de probabilidade de omissão e acrescentando as vogais por ordem de aparecimento. A chave de omissão de *interdisciplinar* será: pdclntsrlea.

A forma como se faz a procura de sugestões é bastante idêntica à apresentada por Damerou, só que aplicada às chaves de semelhança e não às palavras propriamente ditas.

<b>Prós</b>	<b>Contras</b>
∇ Explora bem as propriedades estatísticas dos erros tipográficos, tendo um bom desempenho nos mais frequentes.	∇ O dicionário não pode ser de utilização genérica. ∇ Depende bastante da correcção dos caracteres iniciais. ∇ É relativamente pobre na correcção de erros linguísticos.

## 70 Trigramas

Berkel & Smedt (1988) descrevem um método de correcção ortográfica que se baseia na utilização de trigramas como índices de palavras no dicionário.

O processo de geração de sugestões baseia-se fundamentalmente numa lista de trigramas, que associa a cada trigrama uma lista de ponteiros para as palavras do dicionário que o contêm. Por exemplo, o trigrama *ato* tem ponteiros para palavras como *gato*, *pato*, *chato*, *sapato*, *atolado*, *anatomia*, ou *batota*.

Quando uma palavra é considerada incorrecta, começa-se por a decompor em trigramas. Depois, estes trigramas são usados para obter do dicionário as palavras que os contêm.

Vamos ilustrar todo o processo com um exemplo concreto. Suponhamos que dispúnhamos do dicionário seguinte:

1	arais	5	raia	9	reis
2	cais	6	raias	10	rias
3	dais	7	raio	11	ris
4	orais	8	raios	12	trais

A tabela de trigramas deste dicionário será composta por:

Trig.	Ponteiros
_ar	1
_ca	2

aia	5, 6
aio	7, 8
ais	1, 2, 3, 4, 12

io_	7
ios	8
os_	8

_da	3
is_	1, 2, 3, 4, 9, 11, 12
_or	4
_ra	5, 6, 7, 8
_re	9
_ri	10, 11
_tr	12

ara	1
as_	6, 10
cai	2
dai	3
eis	9
ia_	5
ias	6, 10

ora	4
rai	1, 4, 5, 6, 7, 8, 12
rei	9
ria	10
ris	11
tra	12

Uma vez mais, vamos exemplificar usando a palavra *rais*.

Os trigramas desta palavra são *\_ra*, *rai*, *ais*, *is\_*

Consultando a tabela de trigramas, verificamos que os trigramas da palavra com erro fornecem índices para:

Trigramas	Palavras
_ra	raia, raias, raio, raios
rai	arais, orais, raia, raias, raio, raios, trais
ais	arais, cais, dais, orais, trais
is_	arais, cais, dais, orais, reis, ris, trais

Concluindo, a lista de sugestões seria composta por todas as palavras do dicionário, excepto *rias*. Por outro lado, todas as palavras do dicionário que terminassem em *is*, ou comesçassem por *ra* seriam potenciais sugestões. Até mesmo as palavras *radiotelegrafista* ou *parêntesis*.

Para evitar aparecer na lista de sugestões palavras pouco prováveis de virem a constituir uma correcção, é necessário haver um processo que elimine determinado tipo de palavras. Por exemplo, não considerar palavras com um tamanho bastante diferente da palavra com erro, ou tomar como correcções mais prováveis as palavras do dicionário cuja distância trigrâmica à palavra com erro não seja superior a um determinado valor previamente estabelecido. A frequência com que os trigramas aparecem na língua também pode ajudar na escolha das sugestões mais adequadas.

Prós	Contras
<ul style="list-style-type: none"> <li>∇ Simplicidade do algoritmo.</li> <li>∇ Dá sugestões quando o erro é complexo com a mesma facilidade do que quando é simples.</li> <li>∇ Não depende do tipo de erro.</li> </ul>	<ul style="list-style-type: none"> <li>∇ Exige a manutenção de um elevado número de ponteiros.</li> <li>∇ O algoritmo de sugestões depende da estrutura do dicionário.</li> <li>∇ Pode não cobrir todos os erros, mesmo os mais simples, como no exemplo anterior.</li> <li>∇ Em palavras demasiado curtas pode não haver um único trigrama que seja</li> </ul>

válido. Por exemplo, em "qye" todos os trigramas contêm o "y", donde não se encontraria qualquer sugestão para esta palavra.

## 71 Fonemas e trifones

A ideia de se usar um código fonético em correctores ortográficos baseia-se no facto de as palavras contendo erros linguísticos serem a maior parte das vezes homófonas com a palavra correcta, ou pelo menos fonologicamente muito semelhantes. Assim, ao dar as sugestões só são apresentadas palavras fonologicamente parecidas com a palavra com erro.

Esta ideia aparece em Peterson (1980), mas é em Berkel & Smedt (1988) que encontramos descrições de correctores baseados em métodos fonéticos.

Estes autores apresentam dois correctores dentro desta linha. O mais simples baseia-se num dicionário em que cada entrada é composta por uma palavra com a respectiva representação fonética.

Dada uma palavra com erro, primeiro converte-se a palavra em código fonético. Depois, a representação fonética encontrada para a palavra é usada para aceder o dicionário. O objectivo é encontrar palavras do dicionário que tenham a mesma representação fonética (figura 3.20).

μ §

Figura 3.20 - Encontrar uma sugestão para a palavra *meza*.

Para melhorar o algoritmo, as palavras são convertidas em mais do que um código fonológico, para poder ultrapassar alguns problemas de pronúncia. Por outro lado, algumas distinções finas a nível fonológico são propositadamente omitidas, para assim abranger um maior número de variações.

Os próprios autores reconhecem alguns problemas nesta abordagem. Nomeadamente, não é adequada para lidar com erros tipográficos e só permite corrigir palavras que sejam perfeitamente homófonas da palavra correcta.

A correcção ortográfica baseada em trifones é um método que ultrapassa os problemas anteriores, apesar de continuar a seguir a ideia da utilização de representações fonéticas.

A ideia genérica desta nova abordagem consiste em aplicar técnicas de comparação de cadeias de caracteres a códigos fonológicos e não às palavras propriamente ditas. O método baseia-se em usar a técnica dos trigramas, acima descrita (secção 3.5.3.), aplicada ao código fonológico. Assim, em vez de usar unidades de três caracteres, usa unidades de três fonemas: os trifones.

A principal diferença entre esta abordagem e a dos trigramas reside na natureza dos índices do

dicionário. Isto é, continua a haver um dicionário de palavras, mas em vez de termos uma lista de trigramas com ponteiros para as palavras do dicionário que os contêm, usa-se trifones.

Para se obter uma lista de sugestões para uma palavra incorrectamente escrita, esta é convertida em código fonológico e daí extraídos todos os seus trifones. Estes são usados para obter do dicionário todas as palavras que têm esse trifone.

Tomando o exemplo da figura 3.20, temos os seguintes dicionário e lista de trifones:

Trifones	
_m'e	1, 2, 3
m'eS	1
'eS_	1
m'ez	2
'ezÇ	2
zÇ_	2, 4
m'eS	3
'eSm	3
SmÇ	3
mÇ_	3

Dicionário	
1	<i>mês</i>
2	<i>mesa</i>
3	<i>mesma</i>
4	<i>asa</i>

'az	4
'azɛ	4

A palavra *meza* tem a representação fonética /m'ezɛ/. Os seus trifones apontam para as palavras:

Trifone	Palavras
_m'e	<i>mês, mesa, mesma</i>
m'ez	<i>mesa</i>
'ezɛ	<i>mesa</i>
zɛ_	<i>mesa, asa</i>

Desta forma, não são só as palavras homófonas com a palavra com erro que são seleccionadas para fazer parte da lista de sugestões.

Prós	Contras
<ul style="list-style-type: none"> <li>∇ Dá cobertura a erros tipográficos e ortográficos.</li> <li>∇ Dá sugestões quando o erro é complexo com a mesma facilidade do que quando é simples.</li> </ul>	<ul style="list-style-type: none"> <li>∇ Exige a manutenção de um elevado número de ponteiros.</li> <li>∇ O algoritmo de sugestões depende da estrutura do dicionário.</li> <li>∇ Em palavras pequenas, determinados erros podem provocar alterações de tal maneira profundas na pronúncia</li> </ul>

da palavra, que os trifones obtidos não tenham nada de comum com a palavra correcta.

## 72 Autómatos finitos

Os autómatos finitos são um meio comumente utilizado para representar o léxico de uma língua e fazer processamento morfológico (Koskenniemi, 1983; Agirre et al., 1992; Oflazer, 1994; Pentheroudakis & Higinbotham, 1991). Nestas circunstâncias, a geração de uma lista de sugestões pode tirar partido desta representação específica do léxico, tal como sugerido por Oflazer (1994), que apresenta um método de geração de sugestões baseado nestas ideias.

O processo é dividido em duas partes. A primeira consiste em encontrar um conjunto de radicais que possam originar o radical da palavra com erro. A segunda consiste em percorrer o autómato finito que representa as regras de flexão, construindo palavras que não se afastem muito da palavra com erro.

A semelhança entre as palavras é calculada pela medida de distância de edição. Para esta medida é estabelecido um valor máximo que as palavras candidatas à correcção não podem ultrapassar.

As regras de flexão têm dois níveis: um nível superficial e um nível lexical. Ao percorrer o autómato finito, vai-se aplicando as várias regras de flexão. Simultaneamente, a palavra vai sendo construída (ao nível superficial). De cada vez que uma regra de flexão é aplicada, a parte superficial construída até ao momento é comparada com a palavra com erro e calculada a distância de edição. Se a distância for inferior a um limite pré-estabelecido, a construção da palavra continua pelo mesmo caminho, senão desiste-se do caminho corrente e volta-se a tentar outro, até encontrar um nó terminal.

<b>Prós</b>	<b>Contras</b>
∇ Tira bom partido da representação lexical e morfológica para encontrar as sugestões.	∇ O algoritmo de sugestões está intrinsecamente ligado à representação específica do léxico, não podendo usar um dicionário de
∇ Não é orientado para nenhum tipo es-	

pecífico de erros, podendo oferecer resultados com o mesmo nível, seja qual for a origem do erro.

utilização genérica.

∇ É computacionalmente muito pesado.

## 73 O método ideal

O método ideal para fazer correcção ortográfica possivelmente não existe. Quanto muito, existirão métodos que serão mais aconselháveis em determinadas circunstâncias do que outros.

Dos métodos que descrevemos, a maior parte foi criada com uma motivação específica, ou com certas condições que à partida eram impostas.

Uma das condições que por vezes é imposta é a língua a que o corrector se destina. Por exemplo, as línguas aglutinativas tendem a ter um conjunto extremamente elevado de formas que se obtêm por flexão. O dicionário de um corrector ortográfico para uma língua deste tipo não pode conter explicitamente todas estas formas, sendo necessário recorrer à análise morfológica. Esta limitação tem como principal consequência a impossibilidade de se usar em correcção ortográfica algoritmos que necessitem ter as palavras no dicionário de forma explícita, como no método dos trigramas, ou dos trifones.

Embora não possamos escolher o método ideal, podemos estabelecer características que os correctores ortográficos devem possuir. Este assunto será retomado no capítulo 5, dedicado aos métodos de avaliação de correctores ortográficos. Aqui, salientamos apenas dois aspectos:

1. Os correctores devem processar erros tipográficos e linguísticos igualmente bem. Se por um lado os erros tipográficos são os mais frequentes, por outro lado os erros linguísticos são mais nefastos. Transmitem uma ideia de ignorância, ou falta de formação, e, por isso, deixam pior impressão no leitor. Por outro lado são persistentes. Se alguém comete um erro linguístico, o mais certo é voltar a fazê-lo, a não ser que seja corrigido.
2. Deve haver uma separação bem definida entre a correcção ortográfica e os dicionários usados na correcção ortográfica. Isto é, os algoritmos usados na correcção ortográfica, tanto na verificação, como na geração de sugestões, não devem ser dependentes da estrutura interna do dicionário usado.

Existem várias razões para esta separação:

- a. Havendo esta separação, é possível usar diferentes dicionários, de diferentes origens, com o mesmo corrector. Pode-se mesmo fazer correcção ortográfica de documentos multilingue (Microsoft, 1991).
- b. A independência entre os processos de correcção e acesso ao dicionário, permitem usar o dicionário para outras tarefas. Por exemplo, Prószyński (1994) apresenta várias ferramentas de ajuda à elaboração de documentos, todas elas baseadas num analisador/gerador morfológico e correspondente léxico.
- c. A estrutura interna do dicionário pode ser alterada, sem que isso constitua um problema para a correcção ortográfica. Em particular, é possível usar um analisador morfológico que desempenhe o papel de dicionário, como em Agirre et al. (1992).

Estas ideias baseiam-se no pressuposto de que o corrector ortográfico que se constrói é para uso geral, não se limitando a um domínio específico. Se este for o caso, então poderá ser mais vantajoso usar um algoritmo de correcção que tire o máximo partido das condições de funcionamento, sem olhar às características acima apontadas.

Contudo, parece-nos mais simples adaptar um bom corrector de utilização genérica a um domínio específico, do que tornar mais abrangente um corrector que seja eficiente à custa do domínio limitado em que é usado.

## 74 Correcto, um corrector que usa o Palavroso

O Correcto é um corrector ortográfico que na sua generalidade obedece às directrizes apontadas no final do quarto capítulo. Nomeadamente:

- ∇ Tem uma larga cobertura quanto a tipo de erros que processa, não sendo orientado para um tipo de erro específico. Na sua forma original processa erros linguísticos e erros tipográficos. Os erros de transmissão são tratados como se de erros tipográficos se tratassem. No caso de o corrector ser usado num ambiente em que predominem os erros de transmissão, é possível acrescentar regras específicas que tratem deste tipo de erros.
- ∇ O processo de geração de sugestões é completamente independente do dicionário usado. Começa por criar uma lista de sugestões recorrendo a métodos heurísticos e estatísticos. Estes métodos geralmente criam bastantes palavras que não existem na língua. Por esta razão, cada uma destas palavras é verificada pelo dicionário. A lista final é constituída única e exclusivamente pelas sugestões que existirem no dicionário.

Para além destas características, é de evidenciar a utilização do analisador morfológico Palavroso com função de dicionário. Em algumas línguas, a utilização de um analisador morfológico como dicionário de algum corrector ortográfico é uma imposição. É o caso das línguas aglutinativas, como o Turco (Offlazer, 1994), Húngaro (Prószéky, 1994), ou o Basco (Agirre et al., 1992). No

caso do Português e de outras línguas fusivas, e para o caso específico da correcção ortográfica, esta abordagem pode não ser uma imposição, mas é de certeza uma grande vantagem, como aliás é reconhecido por outros autores (Andrade et al., 1993).

De entre essas vantagens apontamos:

- ∇ Maior e melhor cobertura da língua. Quando o dicionário de um corrector consiste numa simples lista de formas, nem sempre são colocadas todas as flexões das palavras, esquecendo-se, por vezes, algumas palavras importantes (Andrade et al., 1993). Se, por outro lado, for usado um processo de análise morfológica, quando se introduzem novas palavras no dicionário, a maior parte das vezes as flexões dessa palavra ficam automaticamente cobertas. Por outro lado, como não é necessário colocar todas as formas, o espaço que seria ocupado por estas formas pode ser usado para colocar palavras diferentes.

No caso do Palavroso, o seu dicionário tem cerca de 54 000 (cinquenta e quatro mil) entradas, que, juntamente com as regras morfológicas, permitem uma cobertura, no mínimo, de cerca de 1 300 000 (um milhão e trezentas mil) formas do português (cf. Barreiro et al., 1993).

- ∇ Certos fenómenos da língua portuguesa só podem ser tratados convenientemente por recurso à análise morfológica. É o caso das conjugações verbais pronominal e pronominal reflexa. Neste caso, se não se usar a análise morfológica, ou se coloca no dicionário as formas deste tipo de conjugação, ou então reconhece-se apenas os elementos individuais que compõem estas formas. No primeiro caso, dificilmente se poderá fazer uma cobertura suficientemente abrangente. No segundo caso, permite-se que o corrector aceite como correctas certas formas que na verdade estão erradas, já que não se verifica a palavra como um todo.

Um problema idêntico é o das palavras compostas.

- ∇ A existência do analisador morfológico facilita a evolução do corrector ortográfico no sentido de detectar e corrigir outros tipos de erro, como erros de concordância. Tendo esta ferramenta disponível, também é mais fácil construir outras ferramentas de auxílio à elaboração de documentos.

## **75 Descrição global do Correcto**

De acordo com o que acima foi dito, a utilização do Palavroso no corrector ortográfico é feita numa base de cliente-servidor. O corrector solicita um serviço de verificação e o dicionário/analisador morfológico dá a resposta adequada.

O processo de geração de palavras alternativas é completamente independente do processo que

verifica se uma palavra se encontra em dicionário. É possível identificar, tanto a nível de código como a nível de processamento, o que faz parte da verificação e o que faz parte do algoritmo de sugestões. Na realidade, o Correcto poderia usar qualquer outro dicionário, desde que a interface entre os dois fosse normalizada.

A figura 4.1 esquematiza a interacção entre o corrector e o dicionário.

μ §

Figura 4.1 - Interacção entre o corrector e o dicionário.

O módulo de verificação ortográfica recebe uma palavra, prepara-a para ser processada e invoca o dicionário para verificar se a palavra existe ou não. Como o dicionário é de utilização genérica, esse pedido de verificação tem de ser efectuado através de uma interface construída para o efeito. Se a palavra não existe, então o módulo de sugestões entra em acção e gera as palavras alternativas. A maior parte destas palavras não existe na língua, por isso, antes de fornecer palavras ao exterior, o sistema usa o dicionário para verificar se cada uma das palavras existe ou não na língua (no dicionário).

## 76 Interface com outras aplicações

Do esquema da figura 4.1 ressalta a existência de duas componentes fundamentais no módulo de correcção ortográfica. Uma é responsável pela verificação ortográfica, a outra pela geração de uma lista de sugestões de correcções possíveis às palavras assinaladas como erradas. Estas duas componentes identificam-se com duas rotinas que fazem parte da interface do Correcto com outras aplicações. Para além destas, existem outras que são usadas na inicialização e finalização do processo de correcção ortográfica. O quadro 4.2 descreve resumidamente estas rotinas.

Rotina	Descrição	Observações
<b>Inicialização</b>	Procede à inicialização de todos os recursos necessários à correcção ortográfica: dicionários, regras e outro tipo de informação que seja necessário estar acessível	

	em memória.	
<b>Verificação</b>	Verifica se uma palavra se encontra no dicionário, ou não.	Devolve um valor inteiro que indica se a palavra existe ou não. No caso de não existir, indica se se trata de um verbo com clíticos, palavra composta, ou se é um erro genérico.
<b>Sugestões</b>	Procura palavras parecidas no dicionário, que sejam uma possível correcção à palavra com erro.	É o módulo mais importante no algoritmo do corrector. Esta rotina está para o corrector como a rotina anterior está para a análise morfológica.
<b>Libertação de sugestões</b>	Liberta a memória usada pela lista de sugestões. Esta é reservada na rotina que cria as sugestões.	
<b>Libertação de recursos</b>	Liberta a memória de todos os recursos reservados para a correcção ortográfica.	

Quadro 4.2 - Conjunto de elementos funcionais que permitem a utilização do Correcto.

A forma como se usam estas rotinas depende do tipo de correcção que se pretende fazer. Se a correcção for interactiva, é necessário fazer geração de sugestões sempre que é detectada uma palavra com erro; se a correcção for semi-interactiva, esse processamento pode ser feito só no final; se se tratar de um caso de correcção automática, pode ser feita de uma maneira ou outra.

Apesar destas variantes, existem elementos comuns às várias abordagens, que é necessário ter presente. Assim,

- ∇ É sempre necessário proceder à inicialização do corrector quando se começa uma sessão de correcção ortográfica. Esta é sempre a primeira tarefa a ser feita numa sessão de correcção. Não se pode invocar qualquer um dos processos restantes sem ter passado pela inicialização.
- ∇ Deve-se fazer sempre a libertação dos recursos usados pelo Correcto durante uma sessão de correcção. Esta é a última operação a ser efectuada durante uma sessão. Não se pode invocar

mais nenhum dos outros processos após esta tarefa, exceptuando uma nova inicialização do sistema.

- ∇ É necessário libertar os recursos usados pela rotina de sugestões, de cada vez que é gerada uma lista de sugestões. Por cada palavra com erro, é gerada uma lista de sugestões que necessita de memória para a conter. Após a manipulação das sugestões obtidas, a lista deixa de ter interesse e pode-se libertar os recursos consumidos. Caso não sejam libertados, corre-se o risco de esgotar a memória disponível, ou no mínimo, piorar o desempenho do corrector.

## 77 Sessão de correcção

Vamos descrever genericamente como se processa uma sessão de correcção ortográfica.

O primeiro passo consiste em inicializar o corrector ortográfico. Depois, cada uma das palavras do documento é analisada individualmente. Isto é, a rotina de verificação ortográfica é invocada por cada palavra do documento. Se, para alguma palavra, a rotina de verificação devolve uma mensagem indicando que a palavra não é conhecida, invoca-se a rotina de sugestões. Esta, caso consiga, fornece uma lista de palavras que são potenciais correcções da palavra que foi identificada como errada.

A lista de sugestões é, depois de manipulada pela aplicação que solicita o serviço de correcção ortográfica, eliminada. Os recursos computacionais usados por esta lista são libertados para serem usados na elaboração de nova lista de sugestões, ou para outras necessidades do sistema.

O processo de verificação de novas palavras é retomado até a sessão de correcção ser terminada. Aqui, solicita-se a libertação de todos os recursos usados pelo corrector ortográfico.

Todo este processo encontra-se descrito no fluxograma da figura 4.3.

μ §

Figura 4.3 - Fluxograma descrevendo uma sessão de correcção ortográfica.

### Legenda

**Inicialização** - Inicializa o corrector ortográfico.

**Termina?** - Verifica se a sessão termina ou não.

**Finalização** - Liberta os recursos usados pelo corrector.

**Obter Palavra** - Obtém

nova palavra para analisar.

**Existe?** - Verifica se a palavra existe ou não no dicionário.

**Sugestões** - Gera uma lista de possíveis correções à palavra com erro.

**Libertar Sugestões** - Liberta os recursos usados pela lista de sugestões.

Para terminar esta descrição genérica do Correcto, importa fazer as seguintes observações:

- ∇ No processo de inicialização, para além de reservar recursos para uso do corrector e de inicializar variáveis próprias do Correcto, é necessário efectuar a inicialização do analisador morfológico Palavroso, invocando as rotinas adequadas.
- ∇ Tendo em vista a utilização do Correcto no ambiente MS Windows, é possível haver várias instâncias do corrector e do analisador morfológico a funcionar ao mesmo tempo. Por uma questão de economia de espaço, os dicionários do Palavroso e regras, tanto do corrector como do analisador morfológico, são carregados em memória apenas uma vez: na primeira invocação do processo de inicialização. Nas vezes seguintes, só se criam os recursos necessários a uma nova instância. Por exemplo, as variáveis onde se guardam as opções de cada instância.
- ∇ De forma idêntica ao processo de inicialização, os recursos usados pelos dicionários e regras só são libertados uma vez. Se houver mais do que uma instância do corrector ou do analisador morfológico a funcionar ao mesmo tempo, só a última a instância a requerer a libertação de recursos os liberta completamente. Nas outras situações são libertados apenas os recursos usados exclusivamente pela instância que solicita a libertação de recursos.

Na parte restante deste capítulo prestaremos especial atenção aos processos de verificação ortográfica e geração de sugestões, com especial ênfase para este último.

## 78 Verificação

O processo de verificação é uma tarefa da responsabilidade quase exclusiva do analisador morfológico. Apenas é necessário fazer o pré e o pós-processamento do Palavroso.

No pré-processamento são feitas três operações fundamentais:

1. Converter as letras maiúsculas em minúsculas. Não se faz nada no caso de haver mistura destes casos no meio da palavra. O quadro seguinte resume o processo:

<b>Palavra Inicial</b>	<b>Resultado</b>
xxxxxx	xxxxxx
Xxxxxx	xxxxxx
XXXXXX	xxxxxx
xxXxxX	xxXxxX

2. Conversão de caracteres acentuados. Estes são convertidos em sequências de dois caracteres, com o acento a seguir à letra. Por exemplo, o *ã* é convertido em *a~*. O *ç* é decomposto em *c^* (vd. secção 2.2.1.)
3. Estabelecer as opções do Palavroso, com vista ao funcionamento mais adequado e correcto. Por exemplo, de acordo com a configuração das letras maiúsculas, são estabelecidos parâmetros de funcionamento que as caracterizam. Desta forma, no final, é possível restabelecer as características da palavra inicial nas sugestões que são dadas.

Em relação ao pós-processamento, apenas é necessário converter a resposta do Palavroso para um valor inteiro que a caracterize.

O Palavroso, tendo em vista o problema da correcção ortográfica, mas enquadrando-se na filosofia de ser uma ferramenta de utilização genérica, foi modificado por forma a poder passar para o exterior informação relativa ao seu processamento interno, nomeadamente indicar os pontos onde a

análise morfológica eventualmente possa ter falhado, como é o caso dos verbos com clíticos e palavras compostas.

Assim, a operação de pós-processamento limita-se a verificar se o Palavroso reconheceu a palavra ou não. No caso da verificação ter falhado por algum problema detectado no processamento dos verbos com clíticos, ou no processamento das palavras compostas, codifica uma mensagem exprimindo estas situações.

## 79 Sugestões

Diferentes tipos de erros exigem diferentes técnicas de correcção. Assim, no módulo responsável pela criação da lista de sugestões, podemos identificar dois tipos de processamento essencialmente diferentes. Encontramos um conjunto de processos orientados para a correcção de erros de origem linguística e um conjunto de processos orientados para a correcção de erros de origem tipográfica.

O processamento de erros de origem linguística tende a tirar maior partido de heurísticas. Nomeadamente, grande parte deste processamento é baseado em regras de reescrita, como sugerido na secção 3.4.1.4.

No que respeita aos erros de origem tipográfica, o processamento principal consiste em gerar todas as palavras possíveis que se encontrem à distância de edição 1 da palavra com erro. Este processo recorre ao auxílio de trigramas para limitar o número de sugestões geradas.

Embora os erros tipográficos sejam os mais abundantes, aplicam-se em primeiro lugar os métodos heurísticos para verificar se a palavra contém um erro linguístico típico. Só depois, e se nenhuma das heurísticas originar palavras válidas, é feito o processamento relativo aos erros tipográficos. Esta organização deve-se a questões de eficiência. Nomeadamente:

- ∇ A aplicação dos métodos heurísticos consiste num processamento menos pesado computacionalmente do que a detecção de erros tipográficos, pois não gera tantas alternativas que é necessário verificar no dicionário.
- ∇ Os métodos heurísticos têm uma aplicação muito específica. Se uma heurística for aplicada e originar uma sugestão válida, provavelmente será essa a palavra correcta. Assim sendo, quando algum método heurístico origina sugestões válidas, já não se procura encontrar outros erros, nomeadamente os tipográficos, economizando assim um considerável esforço computacional.

Os métodos heurísticos recorrem a vários tipos de soluções para detectarem e corrigirem os erros ortográficos:

- ∇ Correção segundo as regras da morfologia. Este método aplica-se no caso de ter sido detectado um erro nas regras morfológicas. É usado para corrigir erros de verbos com clíticos e flexão das palavras compostas.
- ∇ Correção de erros de acentuação. Devido à grande quantidade de erros ortográficos que envolvem problemas de acentuação, existe um módulo que trata exclusivamente da acentuação das palavras, procurando encontrar um acento que se aplique à palavra com erro.
- ∇ Correção de erros por aplicação de regras de reescrita. Estas regras permitem corrigir um largo espectro de palavras com erros de origem linguística, principalmente as motivadas por semelhanças fonéticas entre a palavra com erro e a palavra correctamente escrita.
- ∇ Correção de erros que envolvem a utilização indevida de letras maiúsculas ou minúsculas.

De forma idêntica, os métodos orientados para a correção de erros tipográficos usam duas abordagens. Uma, de índole mais genérica, procura sugerir palavras que se encontrem à distância de edição um da palavra correctamente escrita. A outra procura corrigir os casos específicos de omissões do espaço separador entre palavras.

Os erros causados por aplicação incorrecta das regras morfológicas são corrigidos através de algoritmos específicos. Os outros são processados usando métodos genéricos.

A distinção entre uns e outros está relacionada com a capacidade que o analisador morfológico tem de detectar o erro cometido. Se o é capaz de fazer, então podemos ter algoritmos de correção para tipos de erro específicos. O processo de geração de sugestões é escolhido em conformidade com a resposta dada pelo processo de verificação ortográfica.

Para já, existem três possibilidades, podendo eventualmente este número ser alargado para outros tipos de correções. Os casos que agora são tratados são os verbos com clíticos, as palavras compostas e o caso geral, que engloba o tipo de geração de sugestões mais genérico.

## **80 Verbos com enclíticos e palavras compostas**

Existe um módulo especializado na criação de sugestões relacionadas com erros que envolvam verbos com clíticos ou palavras compostas.

Neste tipo de erros, o Palavroso dá um diagnóstico de qual terá sido o erro cometido. A criação da

lista de sugestões baseia-se em fazer alterações na palavra motivadas pelo diagnóstico dado. Por exemplo, se a palavra com erro fosse *cantarão-lhe*, o diagnóstico do Palavroso seria que a terminação do futuro não se encontrava separada do verbo. A correcção consistiria em encontrar a terminação do futuro e movê-la para o final da palavra, resultando a sugestão em *cantar-lhe-ão*.

Todas as sugestões estão intimamente ligadas com o diagnóstico dado pelo analisador morfológico. Algumas correcções são directas. Por exemplo, se o diagnóstico do Palavroso indicar ordem incorrecta dos clíticos, o processo de correcção consiste apenas em modificar a sua ordem. Os erros que são tratados desta forma são:

<b>Verbo sem clíticos</b>	Trata erros como <i>cantar-emos</i> . Ao fazer a análise morfológica, o Palavroso junta a terminação característica do futuro ou condicional ao infinitivo. Depois de fazer essa operação, nestes erros, deixam de haver clíticos. O Palavroso reporta o erro e o Correcto cria uma sugestão limitando-se a juntar as duas componentes.
<b>Ordem dos clíticos inválida</b>	Corrige erros da forma <i>deu-me-se</i> . Para isso, analisa cada um dos enclíticos, determina se são reflexivos, se são objecto directo ou se são objecto indirecto e coloca-os na ordem correcta.
<b>Objecto directo e indirectos não contraídos</b>	Trata erros da forma <i>vendeu-me-os</i> . A sugestão criada obtém-se contraindo as duas partículas.
<b>Terminação do futuro ou condicional não separada do infinitivo</b>	Processa erros da forma <i>cantarão-lhe</i> . Cria uma sugestão colocando a terminação após o clítico.
<b>A terminação do verbo ou o enclítico não estão modificados de acordo um com o outro.</b>	Corrige erros como <i>fiz-lo</i> , ou <i>tem-o</i> . Cria uma sugestão baseando-se na terminação do verbo e no clítico que se segue.

**Divisão indevida do radical**

Trata situações como *estáva-mos*, *cantá-se* e *consegui-se*. Na maioria dos casos limita-se a criar uma sugestão retirando o hífen.

**Utilização de uma forma de verbo regular, quando devia ser irregular**

Corrige erros como *dizer-te-ia*. As sugestões criadas baseiam-se fundamentalmente na resposta do Palavroso que, quando detecta um erro deste tipo, devolve a forma correcta, sempre que possível.

Em relação às palavras compostas, só são tratados dois problemas:

### **Composição incorrecta**

A palavra foi composta por justaposição, quando a composição correcta é por aglutinação, como em *contra-prova*. A correcção baseia-se apenas em retirar o hífen.

### **Flexão incorrecta**

A forma como as várias componentes estão flexionadas é incorrecta, como por exemplo, *médicos-cirúrgicos*, em que apenas última componente deveria flexionar.

Para criar as sugestões, o corrector usa o Palavroso para analisar morfologicamente cada uma das componentes individualmente. Usando estes resultados com as componentes flexionadas, constroem-se as várias sugestões. Tomando o exemplo anterior, a análise morfológica daria as palavras *médico* e *cirúrgico*, permitindo construir as palavras *médico-cirúrgicos* e *médico-cirúrgico*.

## **81 Caso geral**

O caso geral, que usa métodos de resolução genéricos, aplica-se na maioria dos casos de palavras com erro. Um erro que tenha sido assinalado como resultante da má aplicação das regras morfológicas só é processado pelos métodos específicos. Os restantes são processados usando os métodos genéricos.

Para se obter as sugestões, ou palavras parecidas, de uma palavra com erro, são efectuados vários processos seguindo uma ordem sequencial. Para cada um destes processos, só se passa ao processamento seguinte se os anteriores tiverem falhado em dar sugestões. No esquema da figura 4.4 vêm enumerados os vários processos, pela ordem por que são executados.

### **1. Processamento de acentos**

#### 1.1. Colocar/retirar acentos

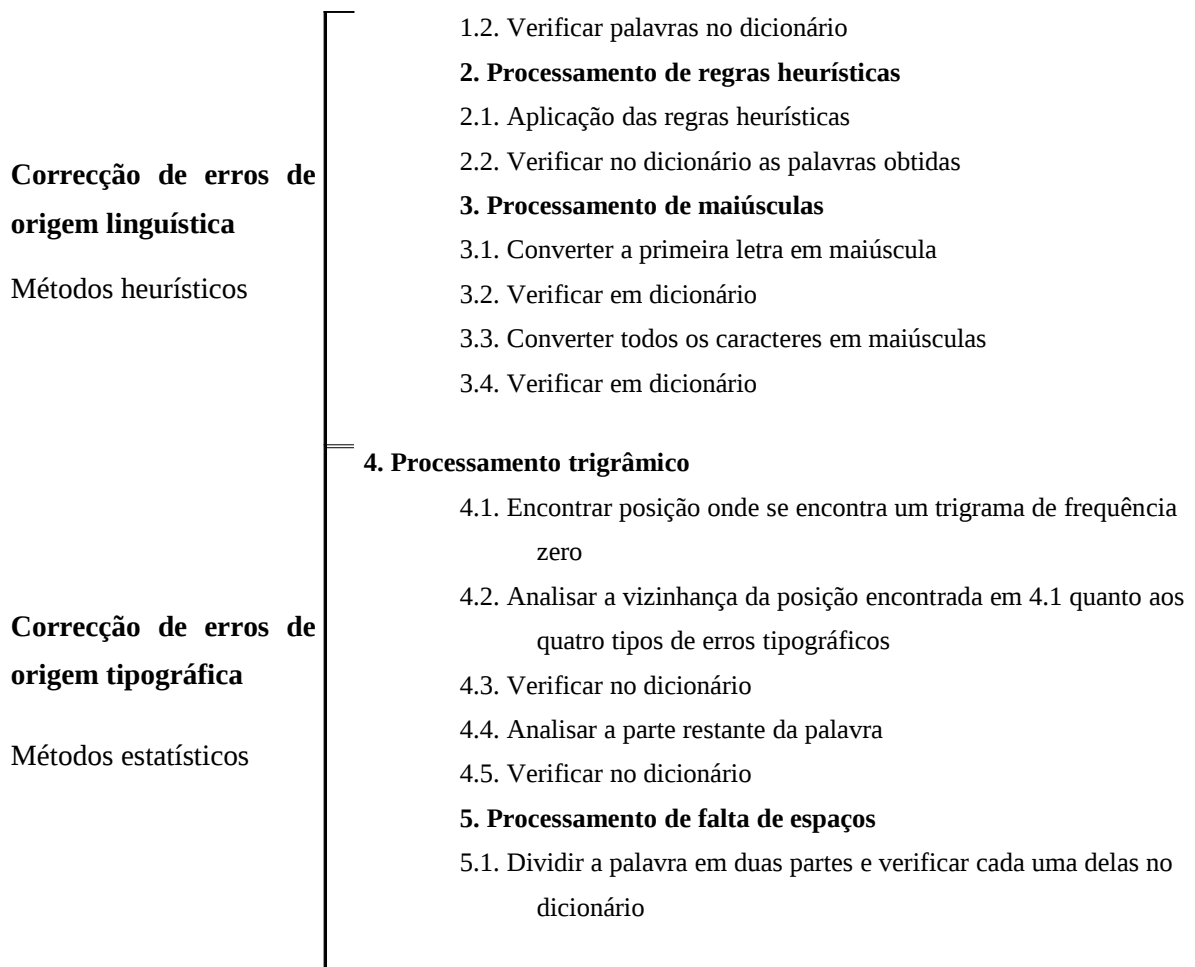


Figura 4.4 - Processos usados na geração de sugestões do caso geral.

De acordo com o que foi dito, após cada processo de verificação no dicionário decide-se se o processamento pára ou continua. Pára se tiver encontrado alguma sugestão válida.

## 82 Erros complexos

Se, depois de fazer todas as tentativas descritas na figura 4.4 para encontrar correcções para a palavra com erro, não tiver sido encontrada qualquer sugestão válida, considera-se que a palavra contém um erro complexo.

O Correcto procura resolver os casos de erros complexos submetendo as sugestões obtidas a nova análise. Desta forma pode detectar erros tipográficos que originem palavras que se encontrem à distância de edição dois da palavra correcta. Pode também detectar erros que envolvam mais do que uma regra heurística.

O problema desta abordagem reside no elevado número de sugestões que são geradas, se aplicarmos este segundo passo a todas as sugestões obtidas no primeiro. O Correcto gera uma média de 60 sugestões por palavra com erro. Se nenhuma destas for válida, procede-se com o passo seguinte que se for aplicado a cada uma das sugestões anteriores, serão geradas cerca de  $\mu$  § sugestões, que é necessário verificar no dicionário.

Este valor médio é demasiado elevado, acarretando um esforço computacional muito pesado. Para aliviar o processamento, só seis sugestões são usadas na segunda iteração: as duas de grau de confiança mais elevado para o processamento dos acentos, as duas de grau de confiança mais elevado para o processamento das regras heurísticas e as duas de grau de confiança mais elevado para o processamento dos trigramas.

Se nenhuma destas sugestões produzir resultados, o processamento pára e nenhuma sugestão é fornecida.

### 83 Processamento relativo aos acentos

Os erros de acentos são três: falta de acento (*comiamos*), acento a mais (*genéricamente*) e acentos trocados, quer em tipo (*estómago*), quer em posição (*âmbigua*).

Para resolver estes três tipos de erros, usa-se duas operações fundamentais:

1. Retirar os acentos das palavras
2. Experimentar colocar os vários tipos de acentos na palavra resultante<sup>1</sup>

Com o primeiro passo detectamos os erros de acento a mais. Nos casos de falta de acento, este passo não influi em nada. No caso de trocas de acentos, preparamos as palavras para o passo seguinte: a inserção de acentos. Uma vez que o til é usado a maior parte das vezes nos ditongos *ãõ* e *õe*, e que raramente acontece as pessoas não colocarem o til nestes casos, optou-se por não criar sugestões com este acento. De forma idêntica, exceptuando as palavras *prà* e *prò*, o acento grave só é usado em contracções da preposição *a* com outros vocábulos, aparecendo sempre no início das palavras. Optámos por processar estes casos de acentuação usando as regras heurísticas.

---

<sup>1</sup> Na versão actual, o Correcto experimenta todos os acentos possíveis em todas as vogais, o que é pouco eficiente. Este será um dos aspectos a melhorar num futuro próximo (vd. Conclusão).

Imaginemos que a palavra com erro era *numéro*. Então obtemos as sugestões seguintes:

<b><i>numero</i></b>	Por aplicação do passo 1
<b><i>número</i></b>	Passo 2, aplicado a numero
<i>numéro</i>	idem
<i>numeró</i>	
<i>numêro</i>	
<i>numerô</i>	

Se a palavra fosse *comiamos*, as sugestões resultantes seriam:

*cóiamos*  
***comíamos***  
*comiámos*  
*comiamós*  
*cômiamos*  
*comiâmos*  
*comiamôs*

A lista gerada é depois ordenada por ordem decrescente de probabilidade da sugestão estar correcta (grau de confiança na aplicação de determinada operação), para depois ser verificada em dicionário.

Para ordenar as sugestões usamos o seguinte critério:

1. Os erros cometidos por a palavra ter um acento a mais são mais prováveis (cf. secção 3.4.1.3). Considera-se, portanto, uma correcção mais provável aquela que for obtida por remoção de um acento de uma palavra com erro. Depois vêm os erros por falta de acento e, por fim, os erros por troca de acento.
2. Dentro dos erros cometidos por falta de acento, assim como nos erros de troca de acento, as sugestões são ordenadas pela probabilidade de ocorrência de omissão de um acento numa determinada letra, de acordo com o quadro 3.3.
3. Para ordenar automaticamente as sugestões por estes critérios, usam-se os valores do quadro 4.5.<sup>1</sup> No caso de troca de acentos usamos o valor da probabilidade para o caso de inserção de acento ponderado pela probabilidade de ser uma troca de acentos. Por exemplo, uma troca que envolvesse a inserção de um acento agudo num *a* teria grau de confiança

$$0.046 * 0.05 = 0.0023.$$

Ao fazer a verificação no dicionário das sugestões geradas por este método aceitam-se quanto muito duas sugestões: uma que tenha sido obtida por remoção de um acento e outra por inserção. Isto prende-se com o facto de que, embora seja frequente aparecerem palavras que diferem pela existência ou não de um acento, é muito pouco frequente aparecerem palavras que diferem apenas pela troca de acentos. Por exemplo, temos o nome *evidência* e temos *evidencia*, forma do verbo *evidenciar*. No entanto não existe qualquer uma de *évidencia*, *êvidencia*, *evídencia*, *evidéncia*, *evidenciá*, *evidenciá*, *evidenciâ*, *evidenciã*.

Correcção	Grau de confiança
Remover acento	0.500

---

<sup>1</sup> Os valores deste quadro são baseados no quadro 3.2 e uma versão modificada do quadro 3.3, em que usamos a percentagem relativa ao total de erros por acento, em vez de ser relativa apenas aos erros por omissão de acento.

Acentuar í	0.205
Acentuar é	0.083
Acentuar ó	0.074
Acentuar á	0.046
Acentuar ú	0.037
Acentuar ê	0.037
Acentuar ô	0.009
Acentuar â	0.003
Troca de acentos	0.05 * (Acentuar x)

Quadro 4.5 - Graus de confiança usados na ordenação das sugestões.

Em termos práticos, isto resume-se a terminar a verificação no dicionário assim que seja encontrada alguma sugestão com grau de confiança inferior a 0.5. Repare-se que este é o valor máximo que um erro de acentuação toma, correspondente à remoção de um acento.

Tomando como base os exemplos anteriores, obteríamos:

Sugestão	Grau de confiança
<i>numero</i>	0.500

Sugestão	Grau de confiança
<i>comíamos</i>	0.205 ←

<i>numéro</i>	0.00415		<i>cómiamos</i>	0.074
<i>numeró</i>	0.0037		<i>comiamós</i>	0.074
<i>número</i>	0.00185	←	<i>comiámos</i>	0.046
<i>numêro</i>	0.00185		<i>cômiamos</i>	0.009
<i>numerô</i>	0.00045		<i>comiamôs</i>	0.009
<hr/>			<i>comiâmos</i>	0.003
			<hr/>	

No primeiro caso seriam verificadas quatro sugestões em dicionário, enquanto que no segundo caso basta verificar uma única sugestão, excluindo logo as restantes seis.

## 84 Regras heurísticas

O processamento com as regras heurísticas visa essencialmente a correcção de erros de origem linguística. É baseado em regras de reescrita (Nilsson, 1971), que são construídas procurando detectar e corrigir erros típicos do português. Por exemplo, detectando palavras em que falte o *c* antes do *t*, como em *deteta*.

Dado as regras serem de reescrita, podemos usá-las para corrigir não só erros ortográficos, mas outro tipo de erros que sejam muito característicos. Por exemplo, erros de transmissão. No caso do problema da representação dos caracteres em 7 bits, pode-se encontrar certos padrões nos erros. Por exemplo, em geral *ç* aparece como *g*, enquanto *ã* aparece como *c*. Em particular, a partícula *çã* aparece *gco*. Poderíamos então ter um regra que sempre que uma palavra com erro tivesse a sequência *gco*, aparecesse uma sugestão em que aquela sequência fosse substituída por *çã*.

Com base no estudo efectuado na secção 3.4.1.4, estabelecemos uma sintaxe limitada e rígida para

expressir as regras de reescrita.

Cada regra é composta por quatro partes:

<b>Padrão</b>	Expressão regular que detecta o erro.
<b>Correcção</b>	Expressão regular que reescreve o erro.
<b>Grau de confiança</b>	Valor real que exprime a confiança que se tem no resultado da aplicação da regra.
<b>Assinatura</b>	Rótulo que identifica as regras que são aplicadas para obter determinada sugestão.

A aplicação de uma regra faz-se da seguinte forma: Se uma palavra com erro tem uma sequência de caracteres que obedece ao **Padrão**, então acrescenta à lista de sugestões uma alternativa substituindo na palavra original o **Padrão** pela **Correcção**. A sugestão terá um grau de confiança igual ao grau de confiança da regra.

O **Padrão** e a **Correcção** são descritos através de pequenas expressões regulares com uma sintaxe bastante limitada.

As expressões regulares criadas para este efeito são compostas por cinco partes. Uma das partes (a Sequência) identifica a parte da palavra que contém o erro, se for a Sequência do **Padrão**, ou a correcção do erro, no caso de se tratar da Sequência da **Correcção**. As outras partes representam o contexto em que a sequência deve aparecer. O quadro 4.6 descreve as componentes das expressões regulares usadas.

Componente	Descrição	Valores que pode tomar
Início	Início da palavra. É composto por tudo o que se encontra entre a primeira letra e a pré-condição.	<ul style="list-style-type: none"><li>• Vazio</li><li>• '*'</li></ul>
Pré-condição	Carácter que pode anteceder a sequência.	<ul style="list-style-type: none"><li>• Vazio</li></ul>

		<ul style="list-style-type: none"> <li>• Identificador de variável</li> </ul>
Sequência	Sequência de caracteres. Errados, ou sua correção.	<ul style="list-style-type: none"> <li>• Qualquer sequência de caracteres.</li> </ul>
Pós-condição	Carácter que pode seguir-se à sequência.	<ul style="list-style-type: none"> <li>• Vazio</li> <li>• Identificador de variável</li> </ul>
Resto	Tudo o que se segue à pós-condição.	<ul style="list-style-type: none"> <li>• Vazio</li> <li>• '*'</li> </ul>

Quadro 4.6 - Componentes de uma expressão regular usada no **Padrão** e na **Correcção**. O asterisco (\*) representa qualquer conjunto de zero ou mais caracteres. A existência de "Vazio" implica o extremo da palavra na posição onde este ocorrer.

As regras são declaradas num ficheiro que é dividido em duas secções. Na primeira secção declaram-se as variáveis que vão ser usadas, enquanto na segunda secção aparecem as regras propriamente ditas.

Uma variável define-se associando a um identificador um conjunto de caracteres que a variável pode representar. Como as variáveis só podem representar um carácter de cada vez, o conteúdo da variável é definido através de uma sequência de caracteres, que se encontra separada do identificador por um ou mais espaços.

Para usarmos uma variável numa expressão, envolvemos o seu identificador com chavetas. Por exemplo, {vogal} é uma variável que poderá representar uma das letras {a, e, i, o, u}.

A figura 4.7 contém um exemplo de ficheiro com regras heurísticas. Nessa figura, os números à esquerda servem para identificarmos as linhas durante esta exposição. Nos verdadeiros ficheiros contendo regras, estas não se encontram numeradas.

```

1:  %variaveis%

2:  vogal          aeiou

3:  cons           bcd fghjklmnpqrstvwxyz

4:

5:  %regras%
```

6:	*ct{vogal}*	*t{vogal}*	0.57	%ct
7:	por{cons}*	pro{cons}*	0.50	%por
8:	*ge*	*je*	0.54	%ge

Figura 4.7 - Extracto de um ficheiro de regras heurísticas.

Nas linhas 1 e 5 encontramos os dois identificadores especiais que dividem o ficheiro em secções.

A primeira regra (linha 6) indica que se uma palavra contém a sequência *ct* seguida de uma vogal, então deve ser criada uma sugestão em que *ct* é substituído por *t*. Por exemplo, a palavra *productivo* seria reescrita como *produtivo*.

A segunda regra diz que se uma palavra começa por *por* e é seguida de uma consoante, então esta sequência deve ser substituída por *pro*. Por exemplo, a palavra *porlongue* seria transformada em *prolongue*.

A última regra permite trocar as ocorrências de *ge* por *je*, independentemente do contexto. Permitem, por exemplo, corrigir *geito* em *jeito*.

Estes pequenos exemplos servem para ilustrar não só o funcionamento das regras na sua globalidade, mas também vários aspectos da sintaxe destas regras. Nomeadamente:

1. Não é necessário que todas as componentes das expressões regulares apareçam numa regra.
2. Quando o Início e a Pré-condição são ambos vazios, então a Sequência só se aplica a prefixos de palavras (regra da linha 7). De forma idêntica em relação ao final da palavra.
3. Existe uma relação biunívoca entre as componentes do **Padrão** e as componentes da **Correcção**. Isto é, o Início de palavra do **Padrão** corresponde ao Início da **Correcção**, a Pré-condição do **Padrão** corresponde à Pré-condição da **Correcção**, e assim sucessivamente.

Por outras palavras, não é possível trocar a ordem pela qual estes elementos aparecem na palavra. O máximo que é possível fazer é eliminar uma das componentes.

Em geral, as componentes da **Correcção** são as mesmas que o **Padrão**, excepto a sequência de caracteres que se está a corrigir.

O grau de confiança das regras obteve-se recorrendo a contagens em corpos de teste. O seu valor calcula-se dividindo o número total de vezes que uma regra foi aplicada, pelo número de vezes que a sua aplicação produziu uma palavra existente no dicionário. Assim, uma regra que seja aplicada muitas vezes, mas a palavra resultante raramente exista no dicionário, tem um grau de confiança baixo.

Este parâmetro permite ordenar as sugestões obtidas por ordem decrescente de probabilidade de serem a ser uma correcção, privilegiando as palavras que foram obtidas através da aplicação de regras com grau de confiança mais elevado.

A inclusão da assinatura nas regras foi motivada pela existência de um elevado número de regras que se anulavam mutuamente. Ou, de outra forma, constatou-se que a maioria das regras admitiam uma regra inversa. Por exemplo, existem as regras

*ch{vogal}*	*x{vogal}*	0.47	%chx
*x{vogal}*	*ch{vogal}*	0.20	#chx

em que uma faz exactamente o inverso da outra.

Quando se trabalha com os erros complexos, pode acontecer uma regra inverter o que a outra já havia feito, produzindo a palavra inicial, que não interessa para a lista de sugestões. Por exemplo, se aparecesse a palavra *chicara* (*xícara*), aplicando a primeira regra obtínhamos *xicara*, que ainda não era a palavra correcta. Quando se fizesse a segunda iteração, a segunda regra estaria em condições de ser aplicada, produzindo *chicara*, novamente.

Optando pelo processo das assinaturas, aquelas duas regras podem ser assinaladas como sendo uma a inversa da outra e ao efectuar a segunda iteração no processamento dos erros complexos nunca aplicar uma regra que seja a inversa de outra que já tenha sido aplicada. Para identificar duas regras inversas usa-se o mesmo identificador, com a diferença de um ser precedido de "#" enquanto que o outro deve ser precedido de "%".

A esta discussão resta acrescentar duas observações:

1. Tenta-se sempre aplicar todas as regras a uma palavra. Daí, a ordem por que elas aparecem listadas não ser importante.
2. A mesma regra pode ser aplicada várias vezes à mesma palavra, produzindo sugestões diferentes. Por exemplo, na palavra *arquitecto* a regra #ct (a inversa de %ct, acima descrita) é aplicada duas vezes, nos dois *t* da palavra, dando origem às sugestões *arquicteto* e *arquitecto*.

## 85 Maiúsculas

O processamento relativo a maiúsculas é extremamente simples: basta alterar um parâmetro de operação do Palavroso.

Quando se faz a normalização das palavras estas são convertidas em minúsculas (cf. secção 2.2.1 e secção 4.2). Durante este processo, para não se perder informação, é estabelecido num parâmetro de operação qual o tipo de conversão que foi feito. Quando se experimenta se uma sugestão pode ser toda em maiúsculas, ou se a primeira letra pode ser maiúscula, basta modificar o respectivo parâmetro de operação, "enganando" o Palavroso. Este irá dar uma resposta como se tal fosse verdade.

Se desta forma conhecer a palavra que dantes era considerada errada, então considera-se que a sugestão de alterar o modo dos caracteres é válida, sendo uma correcção bastante provável.

O Palavroso é invocado duas vezes, uma para cada tipo de modo de caracteres.

Por exemplo, se a palavra com erro fosse *inesc*, então o Correcto experimentava primeiro só colocando a primeira letra maiúscula: *Inesc*. Na realidade esta transformação não é feita, dizendo-se apenas ao Palavroso que a palavra original teria a primeira letra maiúscula. Como esta tentativa iria falhar, a segunda tentativa seria *INESC*, que o Palavroso reconheceria.

## 86 Processamento trigrâmico

Usamos aqui a expressão "processamento trigrâmico" para identificar o processamento relativo aos erros tipográficos, que envolve a utilização de trigramas.

O fundamental desta parte é tentar encontrar palavras que, se forem sujeitas a um dos quatro erros simples, originem a palavra com erro. Isso equivale a aplicar à palavra com erro a operação inversa daquela que terá originado o erro. Por exemplo, se acharmos que o erro cometido foi a inserção de um carácter, então devemos procurar palavras que se obtenham a partir da palavra com erro através da eliminação de um carácter.

A operação inversa de uma inserção é uma eliminação e a operação inversa de uma omissão é uma inserção. No caso das transposições e substituições, estas operações são inversas delas mesmas. Por exemplo, a transposição de uma transposição dá a sequência inicial.

Em termos práticos, partindo da palavra com erro, constroem-se todas as palavras que se encontrem à distância de edição um.

Das palavras assim produzidas nem todas pertencem à língua. Daí a necessidade de verificar se as

palavras obtidas existem ou não no dicionário.

O problema desta abordagem reside no elevado número de alternativas que tem que ser verificada em dicionário.

As operações de edição são efectuadas sobre um conjunto possível de 40 grafemas: as 26 letras do alfabeto, as vogais acentuadas, o ç, o hífen e o espaço. Dada uma palavra de tamanho  $n$ , o número de alternativas produzidas é:

<b>Tipo de operação</b>	<b>Número de alternativas</b>
Eliminação	$n$
Inserção	$(n+1)*40$
Substituição	$n*40$
Transposição	$n-1$
<b>Total</b>	<b><math>82n + 39</math></b>

Por exemplo, se uma palavra tiver comprimento 3, são geradas 285 alternativas que têm de ser verificadas em dicionário. Se o comprimento da palavra for 10, o número de alternativas a verificar sobe para 859. Mesmo nos casos mais optimistas (palavras de tamanho reduzido), o número de acessos ao dicionário é demasiado elevado.

Para limitar o número de alternativas que à partida são produzidas, recorreremos ao uso de trigramas. A ideia consiste em efectuar as operações de edição apenas nos casos em que essas operações não conduzem a trigramas impossíveis na língua. Por exemplo, se a palavra com erro fosse *possivel*, então nunca seria permitido inserir um s por forma a produzir a palavra *posssivel*, pois o trigrama sss não existe em português.

Desta forma, por cada operação de edição efectuada, surgem novos trigramas a compor a palavra que é necessário verificar se existem na língua. O quadro 4.8 contém todas as verificações necessárias para cada uma das operações de edição.

Este tipo de verificação faz decair em cerca de 80% o número de sugestões que é necessário verificar no dicionário. Se o número médio de sugestões por palavra com erro fosse 600, com este processamento passaríamos a ter uma média de 120 alternativas para verificar no dicionário.

Embora este número de alternativas já nos permita encarar como viável o algoritmo de correcção de erros tipográficos, ainda é possível fazer melhor. Por um lado, é possível diminuir o número de sugestões. Por outro lado, é possível ordená-las por ordem decrescente de probabilidade de estarem correctas.

<b>Operação</b>	<b>Sequência original</b>	<b>Sequência final</b>	<b>Trigramas que é necessário verificar</b>
<b>Inserção</b>	<i>axzb</i>	<i>axyzb</i>	<i>axy, xyz, yzb</i>
	<i>_zb</i>	<i>_yzb</i>	<i>_yz, yzb</i>
	<i>ax_</i>	<i>axy_</i>	<i>axy, xy_</i>
<b>Eliminação</b>	<i>axyzb</i>	<i>axzb</i>	<i>axz, xzb</i>
	<i>_yzb</i>	<i>_zb</i>	<i>_zb</i>
	<i>axy_</i>	<i>ax_</i>	<i>ax_</i>
<b>Transposição</b>	<i>abxycd</i>	<i>abyxcd</i>	<i>aby, byx, yxc, xcd</i>
	<i>_xycd</i>	<i>_yxcd</i>	<i>_yx, yxc, xcd</i>
	<i>abxy_</i>	<i>abyx_</i>	<i>aby, byx, yx_</i>
	<i>axyzb</i>	<i>axkzb</i>	<i>axk, xkz, kzb</i>

<b>Substituição</b>	<i>_y<b>z</b>b</i>	<i>_k<b>z</b>b</i>	<i>_k<b>z</b>, k<b>z</b>b</i>
	<i>a<b>xy</b>_</i>	<i>a<b>xk</b>_</i>	<i>a<b>xk</b>, x<b>k</b>_</i>

Quadro 4.8 - Conjunto de verificações que é necessário fazer aos trigramas, para cada uma das operações de edição. O carácter '\_' representa o espaço.

A diminuição do número de sugestões é conseguido à custa dos trigramas de frequência zero, como discutido na secção 3.4.2.5. O Correcto começa por tentar identificar a primeira posição da palavra que contém algum trígama de frequência zero. Se tal for encontrado, então define uma vizinhança desta posição à qual se vai restringir a procura do erro da palavra. Se não encontrar nenhum trígama de frequência zero, então a análise é efectuada em toda a palavra.

Se a posição em que o trígama de frequência zero foi encontrado for  $n$ , então a vizinhança é definida de  $n$  a  $n + 2$ .

No caso destes valores estarem a um carácter de diferença do extremo da palavra, então alarga-se o processamento a esse extremo. Por exemplo, se  $n$  for 1, então a zona de processamento é alargada desde a posição zero à posição 3.

Uma vez definida a zona em que deve ser feito o processamento, passa-se ao processamento propriamente dito, invocando as rotinas para detectarem cada um dos quatro tipos de erro tipográfico: omissão, substituição, transposição e inserção.

Após a fase de geração de sugestões, passa-se à verificação das sugestões no dicionário. Se nenhuma sugestão for validada em dicionário, então alarga-se a geração de sugestões ao resto da palavra.

Para não haver sobreposição de processamento, é necessário manter o rasto do que foi feito anteriormente. Existem quatro situações possíveis, de acordo com o quadro 4.9.

Após a a geração de novo conjunto de sugestões é necessário fazer nova verificação no dicionário.

O processamento com os quatro tipos de erros tipográficos é muito produtivo. Conduz a muitas sugestões válidas, que é necessário organizar por ordem decrescente de probabilidade de serem a sugestão correcta.

<b>Zona analisada</b>	<b>Zona a analisar</b>
-----------------------	------------------------

---



---

Toda a palavra	XXXXXXXXXX	∅
Início da palavra	XXXXxxxxx	xxxxXXXXX
Final da palavra	xxxxxXXXX	XXXXXxxxx
Meio da palavra	xxxXXxxxx	XXxXXxXXX

---

Quadro 4.9 - Situações possíveis para a divisão de processamento.

Para estabelecer estas probabilidades recorreremos aos resultados da secção 3.4.2, que nos dizem quais as tendências das ocorrências de erros. Entramos em linha de conta com os seguintes factores:

1. Frequências dos trigramas.
2. Frequência do tipo de erro. As omissões são os erros mais frequentes. Portanto, as inserções serão as correcções mais prováveis.
3. Posição do erro. Pode ser determinante. Os erros nos caracteres do meio são os mais frequentes, tornando mais plausíveis as correcções a esta zona.
4. Distância das teclas. Uma substituição é mais provável se a letra que substitui for próxima da substituída. Quando se faz uma remoção, esta tem mais peso se a letra a remover estiver próxima no teclado das suas duas letras vizinhas.

Feitas as considerações anteriores, o grau de confiança das sugestões obtidas a partir de métodos tipográficos é calculado da forma que a seguir se descreve.

Considera-se que a correcção mais provável é uma inserção no meio da palavra. Neste caso o grau de confiança é dado exclusivamente pela frequência dos trigramas.

Nos restantes casos a frequência dos trigramas é ponderada pelos outros factores, como seja o tipo de erro e a posição do erro.

Uma remoção ou substituição de letras próximas no teclado são correcções preferenciais. Por essa

razão, as sugestões obtidas por um destes métodos são multiplicadas por um factor aumentativo. A verificação da distância entre teclas é feita através de uma tabela bidimensional, simétrica e composta por zeros na diagonal principal.

## 87 Falta de espaço

Este tipo de processamento aparece em último lugar porque não é um tipo de erro muito frequente.

A ideia genérica baseia-se em inserir um espaço entre cada duas letras da palavra, verificando se a parte que fica à esquerda e a parte que fica à direita existem no dicionário. Vai-se percorrendo a palavra até encontrar um par de segmentos que verifique a condição, ou encontrar o final da palavra.

Este processo pode ser melhorado por forma a evitar alguns acessos ao dicionário. Para isso, basta verificar se a divisão da palavra em dois segmentos produz segmentos tendo trigramas válidos nos extremos. Se algum deles não for, pode-se passar imediatamente à segmentação seguinte, sem verificar a existência dos segmentos no dicionário.

Tomemos como exemplo a palavra *permaneciacom*

Segmentos	Trigrama 1	Possibilidade	Trigrama 2	Possibilidade	Verificação
p ermaneciacom	#p#		#er	Possível	
pe rmaneciacom	pe#	Possível	#rm		
per maneciacom	er#	Possível	#ma	Possível	A verificar
perm aneciacom	rm#		#an		
perma neciacom	ma#	Possível	#ne	Possível	A verificar
perman eciacom	an#		#ec	Possível	
permane ciacom	ne#	Possível	#ci	Possível	A verificar

permanec iacom	ec#		#ia	Possível	
permaneci acom	ci#	Possível	#ac	Possível	A verificar
permanecia com	ia#	Possível	#co	Possível	A verificar

Se usarmos a informação relativa à frequência dos trigramas, verificamos que, em dez pares de verificações que seria necessário fazer, basta na realidade efectuar cinco pares de verificações no dicionário.

## 88 Resumindo

Neste capítulo foram descritas algumas ideias usadas na resolução de vários problemas que colocam na correcção ortográfica. Algumas são baseadas em ideias de outros autores (já discutidas anteriormente), como a criação de sugestões que se encontrem à distância de edição 1 da palavra com erro ou a utilização de informação estatística para encontrar a melhor sugestão.

Outros métodos não foram encontrados na documentação disponível. É o caso da geração de sugestões dos verbos com clíticos e palavras compostas, que segue uma motivação linguística e só é possível utilizando um analisador morfológico.

No geral, o Correcto trata-se de um corrector que processa um largo espectro de erros ortográficos e, dada a sua modularidade de construção, a utilização de regras heurísticas e a utilização de um analisador morfológico, oferece-nos a possibilidade de expansão futura, nomeadamente no tratamento de erros mais complexos e correcções seguindo uma motivação linguística (vd. discussão do trabalho futuro, na conclusão).

## **89 Parâmetros de avaliação de correctores ortográficos**

Por vezes, é necessário comparar o desempenho de vários sistemas computacionais. Outras vezes é necessário avaliar o nível de desempenho de uma determinada aplicação ainda em fase de desenvolvimento.

É possível encontrar alguns trabalhos relacionados com a avaliação de gramáticas computacionais (Erbach, 1992; Volk, 1992) ou de sistemas de tradução automática (Santos, 1988). Por vezes, fazem-se competições entre sistemas da mesma área, o que implica o estabelecimento de parâmetros objectivos que permitam avaliar e comparar os vários sistemas. Este tipo de parâmetros também pode ajudar a orientar o desenvolvimento de novos produtos.

O tipo de testes que conhecemos mais próximos do que seria necessário para um corrector ortográfico, são os testes a analisadores morfológicos. Em particular, as 1<sup>as</sup> Morpholympics, realizadas entre 7 e 8 de Março de 1994 em Nuremberga, opuseram oito analisadores morfológicos para a língua alemã. Neste caso, as únicas medidas apresentadas nos resultados reportavam-se à

velocidade de processamento (número de palavras processadas por segundo) e à cobertura dos sistemas (percentagem de formas reconhecidas).

No que se refere à correcção ortográfica, o mais usual é encontrarmos no final dos artigos científicos alguns comentários relativos ao desempenho dos sistemas que os autores acabam de descrever. Por vezes o desempenho apresentado reporta-se unicamente à velocidade de processamento (Agirre et al., 1992). Noutros casos são fornecidos mais alguns dados, como a percentagem de vezes em que a palavra correcta aparece na primeira posição da lista de sugestões, a percentagem de vezes que a palavra correcta aparece na lista de sugestões (Berkel & Smedt, 1988; Oflazer, 1994) e o número médio de sugestões dadas por palavra com erro (Oflazer, 1994).

Uma fragilidade encontrada nestas avaliações, relatadas pelos próprios autores, é o conjunto de condições em que os testes foram efectuados. Por exemplo, os corpos de teste variavam entre 141 palavras com erro e um máximo de 463 (das quais apenas 188 eram erros únicos). Num dos casos (Berkel & Smedt, 1988), o dicionário do corrector continha apenas 254 palavras.

Embora as medidas anteriores possam dar uma ideia do desempenho dos correctores, pensamos ser possível fazer uma caracterização melhor.

Partindo de algumas ideias extraídas das avaliações encontradas, usando a experiência adquirida no desenvolvimento do Correcto e aproveitando a nossa experiência enquanto utilizadores de outros correctores, tentámos definir um método objectivo e sistemático para avaliar correctores ortográficos.

Há três aspectos fundamentais a considerar:

<b>Velocidade de processamento</b>	É um aspecto clássico a ter em conta em informática: quer-se uma ferramenta rápida.
<b>Correcção dos resultados</b>	É importante que um corrector ortográfico não aceite como certas palavras erradas, dê sempre sugestões válidas, tenha uma boa cobertura lexical, etc.
<b>Funcionalidade</b>	Existe uma série de características de que hoje em dia já não se abdica, como por exemplo, os dicionários de utilizador. Também é desejável que o utilizador tenha o máximo controlo possível sobre a forma de actuação do corrector.



## 90 Velocidade de processamento

O ideal seria activarmos a correcção ortográfica de um texto e no instante seguinte ter todo o documento corrigido. Isso não é possível, até porque, no caso de serem detectados erros, poderá ser necessária alguma supervisão por parte do utilizador.

Mas será sempre desejável que o corrector não leve demasiado tempo a concluir que um parágrafo está completamente correcto. Ou então, depois de detectada uma palavra com erro, demore muito tempo para encontrar um conjunto de sugestões. Pior ainda, quando os correctores demoram bastante tempo a tentar encontrar sugestões, para no final decidirem que não têm sugestões a dar.

Quanto mais cedo se evitarem essas situações, melhor. É na escolha dos algoritmos que se deve começar a garantir que o processamento seja feito em tempos razoáveis.

A maior parte do tempo de processamento é gasto, em princípio, durante a fase de verificação ortográfica. É neste módulo, portanto, que a velocidade de processamento adquire maior importância para o utilizador.

Nesta fase, a velocidade depende essencialmente do acesso ao dicionário, medindo-se em número de palavras processadas por unidade de tempo. A unidade de tempo comumente usada é o segundo. Usamos então a medida palavras/segundo.

Em relação às sugestões, o que interessa para o utilizador é o tempo total de resposta por cada palavra com erro (medido em milissegundos por erro, por exemplo). Como este valor pode depender do número total de sugestões dadas, podemos usar uma outra medida, complementar à anterior: o número médio de sugestões apresentadas por unidade de tempo.

Resumindo, temos três parâmetros que caracterizam a velocidade de um corrector ortográfico:

<b>Parâmetro</b>	<b>Descrição</b>	<b>Unidade de medida</b>
Velocidade de verificação	Número médio de palavras que verifica por segundo	palavras/segundo
Rapidez de resposta	Tempo médio usado para dar uma lista de sugestões	milissegundos/erro
Velocidade de sugestões	Número médio de sugestões que fornece por segundo	sugestões/segundo

--	--	--

Quadro 5.1 - Parâmetros que caracterizam a velocidade de processamento de um corrector ortográfico.

Podem existir limitações sérias ao cálculo destes valores. O ideal seria fazer medições de tempos exactas, sem intervenção humana directa, através do acesso ao código dos correctores. Na impossibilidade disso, somos obrigados a recorrer à cronometração manual, com todos os seus inconvenientes. Deste modo, mesmo usando resultados imprecisos, pelo menos podem-se obter ordens de grandeza.

Em qualquer dos casos, é necessário recorrer a corpos de teste de tamanho aceitável e suficientemente diversificado quanto à classe gramatical das palavras. Para o cálculo do primeiro parâmetro é necessário um corpo de teste composto por palavras sem qualquer erro. Assim, basta obter o tempo total que o corrector necessita para verificar todo o texto e entrar em conta com o número de palavras processadas.

Para o cálculo dos outros dois parâmetros usa-se um corpo de teste composto apenas por palavras com erro. A ideia básica consiste em medir o tempo total gasto para dar as sugestões de todas as palavras com erro, usando, depois, o número de palavras com erro e o número total de sugestões geradas para calcular a Rapidez de Resposta e a Velocidade de Sugestões, respectivamente.

A Rapidez de Resposta obtém-se dividindo o tempo total gasto a dar sugestões, pelo número de palavras com erro. A Velocidade de Sugestões obtém-se dividindo o número total de sugestões produzidas pelo tempo total gasto.

## 91 Correção dos resultados

Este é um dos aspectos fundamentais na avaliação de um corrector ortográfico.

Com os parâmetros que a seguir definimos pretende-se avaliar até que ponto as sugestões que o corrector dá são adequadas ou não.

Quando submetemos um texto à análise de um corrector ortográfico, esperamos acima de tudo duas coisas:

1. Que ele detecte todas as ocorrências de palavras com erro.
2. Que o corrector seja capaz de sugerir a palavra correcta. Assim, com um mínimo de esforço podemos substituir a palavra, eliminando o erro.

Por outro lado, é pouco desejável que o corrector aponte erros que na realidade não existem. Esta falha é sintoma de uma fraca cobertura da língua.

Em certos correctores disponíveis no mercado, aparecem com bastante frequência listas de sugestões com cerca de 30 itens (cf. capítulo 6 - "O Correcto e os outros"). Parece haver aqui a ideia de que um corrector será tanto melhor quanto maior for o número de sugestões que apresentar. Na realidade, quando o utilizador pede sugestões, o seu único objectivo é que o corrector lhe sugira a correcção mais provável, pronta a usar.

Em resumo, a qualidade das sugestões é muito mais importante do que a sua quantidade. O ideal seria que, por cada palavra com erro, se obtivesse apenas uma sugestão: exactamente a palavra que gostaríamos de ter escrito. Na maior parte dos casos tal não é possível. Impomo-nos então o objectivo genérico da palavra correcta aparecer sempre entre as primeiras da lista de sugestões, vindo de preferência em primeiro lugar.

Nas próximas secções deste capítulo procuraremos formalizar as ideias genéricas que acabamos de discutir.

## **92 Dispersão das sugestões**

Parece óbvio que se deve evitar que as listas de sugestões sejam dispersas, ou seja, contendo muitas sugestões que não têm nada a ver com a palavra em questão.

Existe uma medida empírica para o número máximo de sugestões que um corrector deve fornecer. Em geral, as sugestões são fornecidas ao utilizador numa pequena janela, normalmente de tamanho fixo (Microsoft, 1991; IBM, 1991). A lista de sugestões nunca deve ultrapassar, em tamanho, esta janela. Se ultrapassar, o utilizador pode ter que procurar a palavra correcta para além do seu campo de visão imediato, o que passa a ser incómodo. Pensamos, portanto, que uma lista de sugestões com mais de cinco palavras é demasiado grande. Outros autores adotam limites diferente. Agirre et al. (1992), por exemplo, apresenta listas com um máximo de três sugestões. Berkel & Smedt (1988) não impõem um máximo absoluto, mas dão maior importância às três primeiras sugestões na avaliação que apresentam.

Para medir a dispersão das sugestões, temos dois valores: o número médio de sugestões por palavra e o factor de dispersão.

Para estes parâmetros, só entramos com as palavras para as quais o corrector foi capaz de dar sugestões. As outras contam como insucessos.

**Número médio de sugestões por palavra**

Obtém-se dividindo o número total de sugestões pelo número de palavras com erro que obtiveram resposta:

$$\mu \xi$$

em que:

$\mu \xi$  - número de palavras com erro que obtêm resposta.

$n$  - número de sugestões para a palavra  $i$ .

**Factor de Dispersão** Definimos este valor por:

$$\mu \xi$$

em que:

$\mu \xi$  - número de palavras com erro, para as quais houve sugestões.

$n$  - número de sugestões para a palavra  $i$ .

$F^d$  toma valores no intervalo ]0, 1]. É máximo quando  $n$  toma sempre o valor 1. É tanto menor quanto maiores forem os valores de  $n$ .

O valor óptimo de  $F^d$  é 1. Quer isso dizer que a dispersão é nula e que se obteve exactamente uma sugestão por cada palavra.

## 93 Ordenação das sugestões

O ideal é que a primeira sugestão seja a correcta. Sempre que tal não acontece, infere-se que o desempenho do corrector poderia ser melhor nesses casos. Não está a ordenar correctamente a lista de sugestões.<sup>1</sup>

Usámos apenas uma medida: o Factor de Ordenação.

**Factor de Ordenação**                      O seu valor calcula-se através da expressão:

$$\mu \xi$$

em que:

$\mu \xi$  - número de vezes que o corrector fornece a sugestão correcta.

$\mu \xi$  - posição que a palavra  $i$  ocupa na lista de sugestões.

Tal como o factor de dispersão, o factor de ordenação toma valores no intervalo ]0, 1]. É máximo (valor óptimo) quando a sugestão correcta ocupa sempre a primeira posição, sendo tanto menor quanto maiores forem os valores de  $\mu \xi$ .

## 94 Insucesso

Este é um elemento a que se deve dar a máxima importância. Não é agradável trabalhar com um corrector lento, nem é agradável procurar a palavra correcta numa enorme lista de sugestões. Mas que o corrector não detecte alguns erros ortográficos é intolerável .

Podemos avaliar o insucesso de um corrector a dois níveis:

- ∇ Nível das sugestões, quando o corrector não dá sugestões, ou quando a palavra cor-

---

<sup>1</sup> Note-se que este é um parâmetro estatístico. O facto de, geralmente, uma palavra ser a correcção desejada, não quer dizer que o utilizador nunca queira outras correcções.

recta não se encontra entre a lista de sugestões.

- ∇ Nível do léxico, quando ele não reconhece alguma palavra correctamente escrita, ou quando aceita como válida uma palavra com erro.

Sugerimos quatro medidas para avaliar o insucesso de um corrector. Duas que avaliam o insucesso ao nível das sugestões (Factor de Insucesso e Factor Zero) e outras duas que avaliam o insucesso ao nível lexical (Factor de Completude e Factor de Robustez)<sup>1</sup>.

**Factor de Insucesso** Descreve a incapacidade de o corrector incluir a sugestão correcta entre as palavras da lista de sugestões. Obtém-se dividindo o número de vezes em que a palavra correcta não se encontra na lista, pelo número de vezes que o corrector deu uma lista de sugestões:

$$\mu \xi$$

em que:

$\mu \xi$  - número de vezes em que a palavra correcta não se encontra na lista de sugestões.

$\mu \xi$  - número de vezes que o corrector fornece lista de sugestões.

$\mu \xi$  toma valores no intervalo [0, 1]. É máximo quando  $\mu \xi = \mu \xi$ , ou seja, se a palavra correcta nunca se encontrar na lista de sugestões. Toma o valor zero se a lista de sugestões contiver sempre a sugestão correcta.

O valor óptimo é 0 (zero).

**Factor zero** Está relacionado com a não obtenção de qualquer sugestão (**zero** sugestões) para uma palavra com erro. Obtém-se dividindo o número de vezes em que tal acontece pelo número de palavras com erro:

$$\mu \xi$$

---

<sup>1</sup> Escolhemos estes termos por analogia com a lógica. Quer-se que o corrector conheça todas (o maior número possível) as palavras válidas (completo) e não conheça mais nenhuma (robusto).

em que:

$\mu \xi$  - número de vezes em que o corrector não dá qualquer sugestão.

$\mu \xi$  - número de palavras com erro.

A análise deste factor é idêntica ao factor de insucesso. O valor óptimo é zero, situação em que o corrector fornece sempre alguma sugestão.

**Factor de  
Compleitude**

Descreve a cobertura da língua que o corrector oferece. Quer-se que a cobertura seja completa, que reconheça todas as palavras da língua.

Conta-se o número de palavras correctamente escritas que o corrector não conhece e assinala como erro, dividindo-se este valor pelo número total de palavras processadas. O resultado obtém-se subtraindo à unidade o valor encontrado para a razão:

$\mu \xi$

em que:

$\mu \xi$  - número de palavras desconhecidas.

$\mu \xi$  - número total de palavras verificadas.

Toma valores no intervalo [0, 1]. Deseja-se que este valor seja o maior possível. Idealmente, um.

**Factor de Robustez**

Quer-se um corrector robusto. Está relacionado com o número de palavras erradas que o corrector aceita como certas.

$\mu \xi$

em que:

$\mu \xi$  - número de palavras com erro que foram aceites como certas.

$\mu \xi$  - número de palavras com erro.

Toma valores no intervalo [0, 1]. Toma o valor mínimo se todas as palavras com erro forem aceites. Toma o valor óptimo (um) se nenhuma palavra com erro for aceite.

## 95 Eficiência do motor

As medidas que apresentamos a seguir não se aplicam a todos os correctores, pois pressupõem o acesso ao código para permitir o cálculo. Destinam-se, principalmente, a uma avaliação na fase de desenvolvimento de um corrector. Além disso, só têm sentido em correctores em que haja uma separação clara entre o algoritmo que gera a lista de sugestões e a verificação dessas sugestões no dicionário, como é o caso do Correcto.

Temos duas medidas para avaliar a eficiência do motor: o número médio de sugestões que são verificadas no dicionário e o Factor de Eficiência.

**Número médio de sugestões verificadas no dicionário**      Obtém-se dividindo o número total de sugestões geradas para verificar no dicionário pelo número de palavras com erro:

$$\mu \xi \mu \xi$$

em que:

$\mu \xi$  - número total de palavras com erro.

$\mu \xi$  - número total de sugestões verificadas no dicionário.

Pretende-se que este valor seja o mais reduzido possível.

**Factor de Eficiência**      Nestas circunstâncias, definimos a eficiência do algoritmo de sugestões como a razão entre o número total de sugestões que são geradas e o número de sugestões que realmente existem em dicionário:

$\mu \xi$

em que:

$\mu \xi$  - número de sugestões válidas, que existem em dicionário.

$\mu \xi$  - número total de sugestões geradas pelo algoritmo.

A eficiência será tanto maior, quanto mais se aproximar de 1.

Significa isto que o número de sugestões válidas é idêntico ao número total de sugestões geradas.

## 96 Corpos de teste. Índice de correcção

Todos os factores relativos à correcção dos resultados de um corrector ortográfico, que acabámos de descrever, só podem ser obtidos fazendo análise de corpos de teste, que deverão ser de dois tipos:

1. Corpos de palavras com erro. São usados para o cálculo de todos os factores, excepto o factor de completude. Em certas situações, como para calcular o Factor de Ordenação, é necessário saber qual das sugestões é a palavra correcta. Para determinar isso, a lista de erros deve ser composta por palavras com erro associadas à sua correcção.
2. Corpos de teste constituídos por textos sem qualquer tipo de erro. Estes corpos são usados para calcular o Factor de Completude.

Os parâmetros que aqui descrevemos não são mutuamente exclusivos. Com efeito, existe uma certa redundância. Por exemplo, estando definido o Factor de Dispersão, de certa forma é dispensável saber qual o número médio de sugestões por palavra. No entanto, estamos mais habituados a trabalhar com médias. Assim, este valor pode ajudar-nos a formar mais rapidamente uma ideia do desempenho do sistema.

Por outro lado, nem todos os parâmetros têm a mesma importância. Por exemplo, os factores relativos ao insucesso são mais importantes do que os outros. A nosso ver, não é muito vantajoso possuir

um corrector ortográfico com um bom factor de dispersão, se a maior parte das vezes a palavra correcta não constar na lista de sugestões.

Criámos então uma medida que engloba os vários aspectos da correcção de resultados de um corrector ortográfico: o Índice de Correcção.

Do cálculo deste índice vamos excluir a informação relativa à eficiência do motor de sugestões, pois não é uma medida genérica. Excluimos também o número médio de sugestões por palavra, pois este valor não acrescenta mais informação ao factor de dispersão.

Os factores que usamos na definição do Índice de Correcção tomam sempre valores entre zero e um. Uns têm valor óptimo em 1, outros atingem o óptimo em 0. Assim, construímos uma razão de combinações lineares, colocando em numerador os factores que atingem o óptimo em 0 e em denominador os factores que atingem o óptimo em 1. Desta forma, o índice será tanto melhor, quanto menor for o valor da razão.

Temos então:

$$\mu \xi$$

Uma vez que o factor de completude é calculado com base em corpos diferentes dos restantes factores, avançamos com outra aproximação, que também pode ser útil:

$$\mu \xi$$

Agora atribuímos os valores aos escalares  $\mu \xi$  e  $\mu \xi$  de acordo com a importância relativa que os factores têm.

Assim, consideramos que o factor de dispersão é mais importante que o factor de ordenação. Se a dispersão for reduzida, mesmo que na maior parte das vezes a palavra correcta possa não estar na primeira posição, temos sempre a certeza de que se encontra entre as primeiras posições. Por esta razão, atribuímos os valores:

$$\mu \xi$$

Em relação ao insucesso, consideramos o Factor de Robustez como o mais importante. O factor que menos peso deve ter é o de completude. Entre os outros dois, o Factor Zero adquire mais importância do que o Factor de Insucesso. Atribuímos então os valores:

$$\mu \xi$$

A expressão do Índice de correcção é então dado por:

$\mu \xi$

$\mu \xi$

Note-se que, dado que estes valores são baseados em contagens sobre corpos de teste, só têm fundamento em comparações se as condições de obtenção dos resultados forem as mesmas, ou se os corpos usados, embora semelhantes, não pertencerem a um domínio específico e forem suficientemente grandes.

## 97 Funcionalidade

Embora os factores anteriores (velocidade de processamento e índice de correcção) sejam os mais importantes a considerar na avaliação de um corrector ortográfico, existem outros aspectos que podem parecer cosméticos mas que contam para a valorização de um corrector. São os aspectos relacionados com a funcionalidade do corrector ortográfico.

Já Peterson (1980) havia avançado com algumas ideias a respeito da funcionalidade de correctores ortográficos. Algumas dessas características tornaram-se quase obrigatórias. A sua omissão desvalorizará, portanto, o corrector.

Outras características, embora não obrigatórias, são desejáveis e alguns correctores possuem-nas. A sua inclusão num corrector ortográfico tenderá a valorizá-lo. Podem servir também como linha de orientação no futuro desenvolvimento de correctores ortográficos.

Consideramos obrigatórias as seguintes características:

1. Possibilidade de existência de dicionários de utilizador. Em particular, possibilidade de uso simultâneo de vários dicionários de utilizador.
2. Memorização temporária de uma palavra. Quando termina a sessão de correcção, essa palavra é esquecida.
3. Memorização de palavras para mais do que uma sessão de correcção. Isto é, permitir acrescentar palavras ao dicionário de utilizador ao longo de uma sessão.
4. Memorização temporária de uma alteração. A partir desta operação, sempre que a palavra

com erro for encontrada na sessão de correcção corrente é imediatamente alterada pela palavra correcta, sem intervenção do utilizador.

5. Possibilidade de o utilizador especificar se deseja que o corrector ortográfico ignore ou não palavras contendo números ou palavras completamente escritas em maiúsculas.

Consideramos desejável a inclusão das características que se seguem nos correctores ortográficos:

1. Existência de subdicionários de termos técnicos e possibilidade do utilizador especificar quais os subdicionários a usar.
2. Possibilidade do utilizador desfazer a última operação (ou  $n$  últimas operações).
3. Possibilidade de o utilizador obter uma lista de palavras existentes no dicionário com um determinado prefixo.
4. Inclusão de palavras existentes no dicionário do utilizador na lista de sugestões.
5. Sinalização de utilização indevida de espaços. Por exemplo, dois espaços entre palavras, existência de um espaço entre uma palavra e um sinal de pontuação, ou omissão de um espaço a seguir a um sinal de pontuação.
6. Possibilidade de correcção ortográfica multilingue de documentos.
7. Possibilidade de o utilizador escolher entre um modo de correcção interactiva, semi-interactiva ou automática.

Contrariamente aos aspectos anteriormente discutidos em relação à avaliação de correctores ortográficos, a funcionalidade dos correctores não pode ser objectivamente medida. Contudo, não deixa de ser importante verificar a inclusão das características acima descritas. Genericamente diremos que um corrector ortográfico será tanto mais funcional quanto mais das características mencionadas incluir.

## 98 Concluindo

De acordo com os critérios até agora discutidos, na avaliação de um corrector ortográfico entram

vários tipos de informação que dificilmente se conseguirá resumir num só parâmetro. Esta dificuldade é ainda acrescida pela impossibilidade de quantificar o aspecto funcional dos correctores.

Por outro lado, a avaliação e comparação entre vários correctores ortográficos depende bastante de quem avalia e o tipo de características que mais aprecia. Por exemplo, pode considerar a correcção dos resultados um aspecto fundamental, ou pode preferir um corrector que seja rápido, acima de tudo.

Resumindo, os principais aspectos a ter em consideração quando se avalia um corrector ortográfico são:

- ∇ Velocidade de verificação
- ∇ Rapidez da resposta
- ∇ Índice de Correção
- ∇ Existência das características funcionais descritas

## 99 O Correcto e os outros

Este trabalho não ficaria completo se não avaliássemos o desempenho do Correcto, e, implicitamente, do Palavroso, face a outros correctores ortográficos para português.

Dedicamos, pois, este capítulo à avaliação de alguns correctores ortográficos para português europeu presentemente comercializados em Portugal e sua comparação com o corrector que desenvolvemos. Esta avaliação assenta nas ideias desenvolvidas no capítulo anterior.

Embora naquele capítulo tenhamos referido aspectos como a funcionalidade do corrector e a sua velocidade de processamento, aqui daremos mais ênfase aos aspectos linguísticos e mais formais da avaliação.

Esta opção deve-se a vários factores:

1. O aspecto funcional depende, na maior parte das vezes, da aplicação que o corrector usa. Hoje em dia é pouco usual encontrarem-se correctores funcionando isoladamente. Em geral, o corrector encontra-se integrado noutra aplicação, normalmente um editor/processador de texto. Dois dos correctores avaliados são usados pelo mesmo processador de texto, enquanto um terceiro é usado por uma versão mais avançada do mesmo processador. Embora cada um deles possa ter outras funcionalidades, aquela que é visível é semelhante.
2. Os correctores têm plataformas de funcionamento diferentes. Um foi construído para o sis-

tema operativo MS-DOS, outro para ambientes Unix e três deles para o sistema operativo MS Windows 3.x. O Correcto, naquilo que está exclusivamente relacionado com os algoritmos de correcção ortográfica, pode funcionar nas três as plataformas.

Esta diferença pode trazer disparidades significativas nos tempos calculados. Por outro lado, enquanto nas versões de Windows e Unix podemos fazer uma contagem bastante exacta dos tempos gastos, o mesmo não acontece com o corrector construído para o MS-DOS, pois não temos acesso às rotinas responsáveis pela verificação ortográfica e geração de sugestões.

## 100 Os outros

Para além do Correcto, corrector aqui apresentado, desenvolvido pelo autor e propriedade do Grupo de Linguagem Natural do INESC, são aqui analisados cinco correctores comerciais actualmente existentes no mercado. O único critério que presidiu à escolha dos correctores foi a sua acessibilidade, ou disponibilidade.

Os correctores analisados vêm apresentados no quadro 6.1.

Identificação	Proprietário	Plataforma	Aplicações que o usam	Versão
DW5	IBM	MS-DOS	DisplayWrite 5	Versão 1.5
Bruxo	SMD Informática	Unix	Elenix Clássico	
Lince	ILTEC; Priberam	MS Windows	Aplicações da Microsoft para Windows: WinWord, Excel, PowerPoint.	Versão 1.0
HM2	Houghton Mifflin	MS Windows	Aplicações da Microsoft para Windows: WinWord, Excel, PowerPoint.	Versão 1.20
HM3	Houghton Mifflin	MS Windows	Aplicações da Microsoft para	Versão 1.30

			Windows: WinWord, Excel, PowerPoint.	
--	--	--	---	--

Quadro 6.1 - Conjunto de correctores usados para avaliar e comparar com o Correcto.

## 101 Preparação do teste

A avaliação dos sistemas é fundamentalmente baseada em testes sobre os corpos de texto descritos no apêndice A. Estes corpos encontram-se divididos em várias partes, em função das suas características. Como o tamanho destas partes difere bastante e as proporções entre elas não reflectem as proporções reais em que os vários tipos de erro ocorrem em textos, a análise vai ser segmentada em função dos corpos de teste. Contudo, no final procuraremos apresentar uma avaliação global dos sistemas.

Os testes a efectuar são:

**Erros linguísticos** – Procura-se avaliar o desempenho dos correctores no que diz respeito a erros de origem exclusivamente linguística. É usado o corpo de teste CorpoLing.

**Erros tipográficos** – Avalia-se o comportamento dos correctores no que diz respeito a erros meramente tipográficos, quer sejam simples, complexos ou de omissão de espaços. O corpo com este tipo de erros, CorpoTipo, continha algumas ocorrências de erros por transposição, em que o acento aparecia antes da vogal. Este tipo de erros é característico das mensagens de correio electrónico, não se esperando que apareça em textos normais. Para que a comparação seja justa, foi criado um outro corpo de erros tipográficos, em que estas formas não aparecem: CorpoTipo2.

**Cobertura do corrector** – Temos dois corpos de teste: CorpoTeste e CorpoDic. O primeiro foi usado para testar o Correcto e o Palavroso durante a fase de desenvolvimento, pelo que a sua utilização pode falsear os resultados do Correcto face aos seus competidores. É, pois, mais objectivo usar só o segundo.

## 102 Resultados

Seguem-se os resultados obtidos para cada um dos testes acima descritos. Em cada teste só aparecem os parâmetros que tem sentido avaliar com o corpo de teste usado. Por exemplo, se o corpo de teste é composto só por erros ortográficos, então não tem sentido avaliar a cobertura do corrector.

Para tornar o quadro mais legível, associado a cada parâmetro está indicado o intervalo em que o parâmetro em questão pode tomar valores, assim como o valor óptimo.

### Erros linguísticos

Corpo de teste: *CorpoLing*

Número de palavras: 1581

Medida	Intervalo, Óptimo	DW5	Lince	Bruxo	HM2	HM3	Correcto
Média de sugestões por palavra	$\mu$ § Opt: 1.0	4.63	10.40	10.40	1.39	1.47	1.22
Factor de Dispersão	$\mu$ § Opt: 1.0	0.28	0.26	0.26	0.90	0.88	0.95
Factor de Ordenação	$\mu$ § Opt: 1.0	0.87	0.85	0.85	0.98	0.97	0.98
Factor de Insucesso	$\mu$ § Opt: 0.0	0.18	0.16	0.16	0.24	0.19	0.14
Factor Zero	Opt: 0.0	0.00	0.10	0.10	0.10	0.05	0.16
Factor de Robustez	$\mu$ § Opt: 1.0	0.92	0.95	0.95	0.96	0.96	0.91
Índice de Correção	$\frac{c}{I'}$	0.0705	0.1199	0.1626	0.1178	0.0807	0.1166

	Opt: 0.0						
--	----------	--	--	--	--	--	--

Os três primeiros correctores apresentam um fraco desempenho a nível de dispersão e ordenação das sugestões. Mesmo assim, o DW5 consegue obter o melhor índice de correcção, principalmente devido ao valor conseguido no factor zero.

O desempenho do Correcto não é excepcional, apresentando um índice de correcção mediano. O seu pior desempenho é devido ao factor zero e ao factor de robustez.

## Erros tipográficos

Corpo de teste: *CorpoTipo2*

Número de palavras: 266

Medida	Ótimo	DW5	Lince	Bruxo	HM2	HM3	Correcto
Média de sugestões por palavra	$\mu$ § Opt: 1.0	4.65	9.11	1.97	1.46	1.68	1.93
Factor de dispersão	$\mu$ § Opt: 1.0	0.28	0.31	0.78	0.86	0.82	0.78
Factor de correcção	$\mu$ § Opt: 1.0	0.86	0.86	0.86	0.95	0.94	0.93
Factor de insucesso	$\mu$ § Opt: 0.0	0.20	0.19	0.07	0.22	0.22	0.13
Factor Zero	Opt: 0.0	0.01	0.12	0.25	0.05	0.08	0.17
Factor de robustez	$\mu$ § Opt: 1.0	0.99	0.99	0.99	0.99	0.99	1.00
Índice de Correcção	$\frac{c}{I'}$ Opt: 0.0	0.0765	0.1360	0.1390	0.0890	0.1040	0.1186

Os valores encontrados para os erros de origem tipográfica são idênticos aos de origem linguística. Embora o Correcto tenha melhorado o seu factor de robustez, continua a ter um factor zero demasiado elevado.

## Cobertura do corrector

Corpo de teste: *CorpoDic*

Número de palavras: 52482

Medida	Ótimo	DW5	Lince	Bruxo	HM2	HM3	Correcto
Factor de cobertura	$\mu$ §Opt: 1.0	0.99	0.99	0.95	0.95	0.98	0.97

No que se refere à cobertura lexical, o Correcto também não apresenta resultados entusiasmantes, situando-se num plano médio. É de salientar o desempenho do DW5 e do Lince que não reconheceram apenas um pequeno conjunto de palavras.

### 103 Avaliação global

Os valores que apresentámos acima dão uma ideia razoável da qualidade relativa dos correctores. Tratam-se, contudo, de parâmetros estatísticos, que dependem bastante dos corpos de teste usados.

No caso presente, não podemos dizer que os corpos sejam os ideais. Por exemplo, nos testes de cobertura lexical foi usado um corpo que continha muitos nomes próprios. A consequência disso está patente nos maus resultados do Correcto, que não privilegia o processamento deste tipo de palavras. Pensamos que deve ser mantido um mínimo, deixando para o utilizador final a decisão de os incluir ou não no dicionário pessoal.

Para além desta descrição formal dos correctores, podemos avançar um pouco mais e exprimir algumas impressões que só é possível conseguir através da utilização directa dos correctores.

A maior dificuldade encontrada prende-se com o tratamento das palavras compostas e dos verbos com clíticos.

Para resolver estes problemas constatámos a existência de duas abordagens com filosofias opostas. O DW5 analisa cada uma das componentes individualmente, resultando daí muitas situações de incorrecção, ou pelo menos de possível incorrecção. Por exemplo, aceita como correcto qualquer nome composto, desde que reconheça cada uma das componentes. Pode aceitar, por exemplo, *caminhos-dos-ferro*.

No caso dos verbos, para reconhecer as componentes individualmente, necessita de ter no seu dicionário formas como *los* e *á*. Também as formas verbais não fogem a esta regra. Assim, possui no seu dicionário a palavra *cantá*, para que seja possível reconhecer *cantá-los-ei*. Como resultado, reconhece como certa a palavra *cantá-se* e nunca assinala como erro a "contração" *á*.

Contudo, a pior impressão fica quando se constata que estas palavras também aparecem nas listas de sugestões.

A outra abordagem a este problema consiste em dicionarizar as palavras completas. No caso dos verbos com clíticos aparece a dificuldade em armazenar tão elevado número de flexões. Como resultado, só são colocadas no dicionário as flexões mais usuais, falhando em bastantes casos.

Esta solução tem o mérito de não reconhecer como certas formas erradas nem de dar como sugestões palavras incorrectas.

Neste aspecto, o Correcto tem vantagens sobre os outros correctores, pois não só analisa qualquer conjugação ou nome composto, como é capaz de sugerir correcções para os problemas que encontra.

O maior problema do Correcto está relacionado com o factor zero. Isto é, num número significativo de casos não é capaz de encontrar qualquer sugestão válida.

## **Conclusão**

É sabido que a componente lexical e morfológica é fundamental em qualquer sistema de processamento de linguagem natural. Qualquer grupo que inicie o seu trabalho na área do processamento automático de linguagem natural, tem de despender um apreciável esforço na componente lexical e morfológica.

O Palavroso é o resultado desse esforço no Grupo de Linguagem Natural do INESC.

Nesta dissertação descrevemos o analisador morfológico Palavroso e, mais do que isso, apresen-

támos o Correcto, um corrector ortográfico para português que usa o Palavroso.

Para além da descrição do corrector ortográfico, descrevemos e caracterizámos um conjunto de erros ortográficos de português, tendo em vista a orientação mais adequada no desenvolvimento do Correcto.

Porque era necessário a qualquer momento avaliar objectivamente o desempenho do Correcto e porque verificámos haver uma falha nesta área, sugerimos uma série de medidas de desempenho que nos permitem avaliar a eficácia de qualquer corrector.

Utilizando as medidas descritas, fizemos uma comparação entre vários correctores ortográficos actualmente comercializados e o Correcto. Concluimos que o corrector aqui apresentado tem um comportamento bastante positivo e que, uma vez optimizado e convenientemente testado, poderá competir com vantagem com os correctores ortográficos actualmente existentes para a língua portuguesa.

Apesar da avaliação das ferramentas apresentadas nesta dissertação ser genericamente positiva, isto não quer dizer que seja impossível fazer melhor, ou que tenha terminado o trabalho referente ao processamento morfológico e à correcção ortográfica do português. Por um lado, prevê-se a comercialização das ferramentas aqui apresentadas, o que exigirá um trabalho constante de manutenção e actualização. Por outro lado, durante a elaboração deste trabalho foram aparecendo muitas ideias novas para a melhoria tanto do Correcto como do Palavroso. Considerámos útil discutir estas ideias aqui, à guisa de conclusão.

O que de mais urgente há a fazer é a optimização do código tanto do corrector, como do analisador morfológico. Essa optimização passa essencialmente pelo aspecto da velocidade e pela fiabilidade. O Palavroso na sua forma actual usa dicionários de tamanhos apreciáveis e de acesso bastante moroso. É necessário proceder, portanto, à compressão dos dicionários, bem como à diminuição do tempo de acesso aos mesmos.

São também possíveis outras melhorias no código com vista ao aumento de velocidade. Algumas alterações poderão ser pontuais, mas outras poderão ser mais significativas, especialmente se as melhorias passarem por alterações como as sugeridas em Medeiros (1994).

No que diz respeito à fiabilidade, é necessário testar as ferramentas com utilizadores reais, trabalhando em ambientes reais. Pode-se detectar, assim, certos erros dos quais não nos tenhamos apercebido.

No campo específico da morfologia, dever-se-ia trabalhar a parte de derivação, que está pobremente tratada na actual versão do Palavroso.

Este é o trabalho para o futuro imediato. Visando a continuação deste trabalho num futuro mais distante, podem ser tratados os seguintes assuntos:

1. Analisar até que ponto se pode tirar maior partido da análise morfológica para a correcção, em particular na geração de sugestões. Por exemplo, se o analisador morfológico devolver a menor parte da palavra que ele não foi capaz de reconhecer, eventualmente o erro poderá ser mais facilmente detectado, seja de que origem for. Imaginemos que a palavra com erro era *pertendíamos*. O analisador morfológico reconheceria a terminação da palavra *íamos*, não sendo capaz de reconhecer nenhuma terminação maior. Consequentemente, *pertend* seria a menor parte não reconhecível, limitando-se a pesquisa do erro a este fragmento. Desta forma, reduziríamos para quase metade a zona onde se procuraria o erro.

Este tipo de processamento deve ser analisado com cuidado. Numa breve análise que fizemos ao corpo de teste de erros, verificámos que na maioria dos erros a parte reconhecível pelo analisador morfológico reduzia-se a um ou dois dos últimos caracteres da palavra. Se efectivamente assim for, é possível que este tipo de processamento não se justifique.

2. Ainda no campo da geração de sugestões, uma das fontes de erro linguístico é a utilização de formas de palavras regulares, incorrectas, em vez de formas de palavras irregulares. Por exemplo, usar *intervido* em vez de *intervindo*. Um processo de gerar sugestões poderia ser analisar a palavra com erro usando unicamente regras regulares. Se desta forma se obtivesse alguma palavra existente em dicionário, então gerar-se-ia a forma que tivesse as características morfológicas analisadas, mas agora usando todo o tipo de regras. Principalmente as irregulares. A forma assim obtida seria a correcção mais provável.

Se aplicássemos ao exemplo *intervido* apenas regras regulares, o analisador morfológico diria que se trata do particípio passado do verbo *intervir*. Usando agora também as regras irregulares, ao procurar gerar a forma com aquelas características morfológicas, obter-se-ia *intervindo*, que é a sugestão correcta.

Embora esta abordagem possa ser bastante eficiente, tem a desvantagem, por agora, de necessitar de geração morfológica.

3. Avançar um pouco mais na complexidade dos erros detectados, tirando partido das palavras que se encontrem próximas, considerando pelo menos a palavra anterior, ou a palavra seguinte. Uma vez que a análise morfológica se encontra disponível, poder-se-ia usar uma gramática simplificada e restrita que detectasse determinado tipo de erros. Por exemplo erros de concordância, como "...tu fizestes...".
4. Sempre que possível, para além das sugestões, o corrector deveria dar uma explicação do erro cometido, ou da correcção sugerida. Em determinados erros de origem linguística pode ser bastante educativo.

Por exemplo, se a palavra com erro fosse *cantarão-se*, o corrector deveria dar uma resposta do tipo:

Sugestões	Explicação
<b>cantar-se-ão</b>	Na conjugação reflexiva de um verbo, no futuro ou no condicional, a partícula "se" fica sempre entre o infinitivo e a terminação.

5. Para determinado tipo de confusões conhecidas, o corrector deveria chamar a atenção do utilizador, mesmo que a palavra exista em dicionário. Por exemplo, se encontrasse num texto a palavra *imigrante*, inquirir o utilizador a respeito do significado que ele pretende:

	Palavra	Significado
? Quer referir-se a	<b>imigrante</b>	entra no país.
? Ou a	<b>emigrante</b>	sai do país.

6. Usar um analisador sintáctico para dar só sugestões de palavras coincidentes com a classe morfológica. Num determinado contexto, ou é *prezo* (Verbo), ou *preso* (Nome).
7. Há situações em que tanto o Palavroso como o Correcto têm que experimentar a colocação de vários acentos numa palavra, até encontrar uma forma válida. Em vez de se usar um método de acentuação cego, que experimenta vários acentos em todas as vogais da palavra, poder-se-ia usar um processo mais inteligente.

Geralmente, se uma palavra não tem acento gráfico, a palavra é grave e o acento tónico é feito na penúltima sílaba. Logo, não se deveria tentar colocar acentos nesta sílaba. O mais natural é a palavra ser esdrúxula, caso necessite de acento gráfico.

8. Da mesma forma que o corrector deve dispor de dicionários de utilizador, também deve ser dada a possibilidade de o utilizador poder especificar regras heurísticas.
9. A análise morfológica deve ser alargada à informação existente nos dicionários de utilizador. Isto é, em vez de o utilizador acrescentar várias formas do mesmo verbo ao seu dicionário, deveria ser possível o utilizador especificar apenas o infinitivo impessoal do verbo, com alguma informação morfológica adicional, e o corrector (ou analisador morfológico) passaria a reconhecer imediatamente todas as formas desse verbo.

Esta facilidade exige, acima de tudo, a criação de um módulo bastante evoluído para fazer a introdução de dados no dicionário.

10. O corrector deve alertar o utilizador para a ocorrência de estrangeirismos, sugerindo termos portugueses equivalentes. De forma idêntica, alertar para a utilização de termos em estilo demasiado familiar.

Um trabalho que poderá seguir logicamente o descrito nesta dissertação é a criação de um gerador morfológico que use as mesmas regras e os mesmos dicionários que o Palavroso.

## Bibliografia

- Agirre, E.; Alegria, I.; Arregi, X.; Artola, X.; Ilaraza, A. Díaz de; Maritxalar, M.; Sarasola, K.; Urkia, M. (1992). "Xuxen: A spelling checker/Corrector for Basque Based on Two-Level Morphology". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 119-125.
- d'Andrade, Ernesto; Soares, Helena; Fraústo, Isabel (1992). "Lince, um corrector ortográfico português". *Actas do 1º Encontro de Processamento da Língua Portuguesa - EPLP'93*, Lisboa, pp. 97-100.
- Barkey, Chuck (1992). "Indexing Large Quantities of Documents Using Computational Linguistics". *Proceedings of the Second Twente Workshop on Language Technology*, Twente, pp. 59-63.
- Barreiro, Anabela; Pereira, Maria de Jesus; Santos, Diana (1993). *Critérios e opções linguísticas no desenvolvimento do Palavroso, um sistema computacional de descrição morfológica do Português*, Relatório INESC RT/54-93, Lisboa.
- Bergström, Magnus; Reis, Neves (1987). *Prontuário Ortográfico e Guia da Língua Portuguesa, 18ª Edição*, Editorial Notícias, Lisboa.
- Berkel, Brigitte van; Smedt, Koenrad De (1988). "Triphone analysis: a combined method for correction of orthographical and typographical errors". *Proceedings of the Second Applied Natural Language Processing Conference*, ACL, Austin, pp. 77-83.
- Burillo, Manuel Pujol (1994). "Tècniques de correspondència aproximada de cadenes de text". *Actas del X congreso de Languages Naturales y Languages Formales*, Sevilla, pp. 543-547.
- Carter, David M. (1992). "Lexical Processing in the CLARE System". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 259-260.
- Coch, Jose; David, Raphael (1994). "Representing Knowledge for Planning Multisentential Text". *Proceedings of the 4th Conference on Applied Natural Language Processing*, ACL, Stuttgart, pp. 203-204.
- Costa, J. Almeida; Melo, A. Sampaio (1989). *Dicionário da língua portuguesa, 6ª edição*, Porto Editora, Porto.
- Cuesta, Pilar Vázquez; Luz, Maria Albertina Mendes da (1971). *Gramática da Língua Portuguesa, Edições 70*, Lisboa.
- Cunha, Celso; Cintra, Lindley (1987). *Nova Gramática do Português Contemporâneo*, João Sá da

- Costa, Lisboa.
- Damerau, Fred J. (1964). "A Technique for Computer Detection and Correction of Spelling Errors". *Communications of the ACM*, Março 1964, pp. 171-176.
- Dumitrescu, Cristian (1992). "Lexicon Design Using a Paradigmatic Approach". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 247-248.
- Erbach, Gregor (1992). "Tools for Grammar Engineering". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 243-244.
- Estrela, Edite (1983). *Dúvidas do falar português, vol I*, Editorial Notícias, Lisboa.
- Estrela, Edite; Pinto-Correia, J. David (1988). *Guia Essencial da Língua Portuguesa para a Comunicação Social*, Edição do II Congresso dos Jornalistas Portugueses.
- Fernandes, Carla (1990). "Para uma Tipologia dos Erros do Português Escrito". *Documentação do Grupo Científico IBM-INESC*, Relatório INESC, Lisboa.
- Ferreira, Aurélio Buarque de Holanda (1986). *Novo Dicionário da Língua Portuguesa*, Editora Nova Fronteira, Rio de Janeiro.
- Ferreira, Aurélio Buarque de Holanda (1993). *Aurélio Eletrônico*, Editora Nova Fronteira, Rio de Janeiro.
- Figueiredo, J. M. Nunes; Ferreira, A. Gomes (1974). *Compêndio de Gramática Portuguesa*, Porto Editora, Porto.
- Filho, D' Silvas (1994). *Prontuário – Erros Corrigidos de Português*, Texto Editora, Lisboa.
- Florido, Maria Beatriz; Silva, Maria Emília Duarte (1978). *Novos Caminhos para a Linguagem. Gramática pedagógica do Português*, Porto Editora, Porto.
- Fonseca, Ana C. S. (1993). *Comunicação em Linguagem Natural para um Tutor Inteligente*. Tese de mestrado, Instituto Superior Técnico - Universidade Técnica de Lisboa, Lisboa.
- Galisson, R.; Coste, D. (1976). *Dictionnaire de Didactique des Langues*, Librairie Hachete, Paris.
- IBM (1991). *Display Write 5 – Manual de Referência*, International Business Machine.
- Kerpedjiev, Stephan M. (1992). "Automatic Generation of Multimodal Weather Reports from Datasets". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 48-55.
- Koskenniemi, Kimmo (1983). *Two-Level Morphology: A General Computational Model for Word-form Recognition and Production*, Publications of the Department of General Linguistics, University of Helsinki.
- Leavitt, John R. R. (1992). "MORPHÉ: A Practical Compiler for Reversible Morphology Rules".

- Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 233-234.
- Lima, Vera Lúcia; Kipper, Karin (1992). "Analisador Morfológico para Tratamento de Textos em Português". *Actas do 1º Encontro de Processamento da Língua Portuguesa - EPLP'93*, Lisboa, pp. 39-44.
- Lyons, John (1968). *Introduction to Theoretical Linguistics*, Cambridge University Press, London.
- Marques, Rui (1994-a). "Anotação Contextual do Corpus INESC 1990". In Santos, D. (1994), *Processamento de Corpora de texto no INESC, Vol. 2*, Relatório INESC, Lisboa.
- Marques, Rui (1994-b). *Homografia: Relações Morfológicas e Semânticas*, Relatório INESC, Lisboa.
- Medeiros, José Carlos; Marques, Rui; Santos, Diana (1993). "Português Quantitativo". *Actas do 1º Encontro de Processamento da Língua Portuguesa – EPLP'93*, Lisboa, pp. 33-37.
- Medeiros, José Carlos (1992). "Ferramentas de Processamento de Corpora Usando o Palavroso". In Santos, D. (1992), *Processamento de Corpora de texto no INESC, Vol. 1*, Relatório INESC RT/65-92, Lisboa.
- Medeiros, José Carlos (1994). *Uso de Informação Quantitativa num Analisador Morfológico de Português*. Trabalho da cadeira de mestrado "Introdução à investigação", Instituto Superior Técnico – Universidade Técnica de Lisboa, Lisboa.
- Microsoft Co. (1991). *Microsoft Word for Windows User's Guide*, Microsoft Corporation.
- Nascimento, M. Fernanda Bacelar; Mendes, Amália; Santos, Diana (1993). "O Corpus e a Classificação Sintáctica dos Verbos". *Actas do 1º Encontro de Processamento da Língua Portuguesa – EPLP'93*, Lisboa, pp. 125-129.
- Nijholt, Anton (1992). "Linguistic Engineering: A Survey". *Proceedings of the Second Twente Workshop on Language Technology*, Twente, pp. 1-22.
- Nilsson, Nils (1971). *Problem-Solving Methods in Artificial Intelligence*, McGraw-Hill, New York.
- Oflazer, Kemal (1994). "Spelling Correction in Agglutinative Languages". *Proceedings of the 4th Conference on Applied Natural Language Processing*, ACL, Stuttgart, pp. 194-195.
- Pentheroudakis, Joseph E.; Higinbotham, Dan W. (1991). "Morphogen: A Morphology Grammar Builder and Dictionary Interface Tool". *Proceedings of the 12th Meeting of the Deseret Language and Linguistics Society*, Brigham Young University.
- Pentheroudakis, Joseph E.; Vanderwende, Lucy (1993). "Automatically Identifying Morphological Relations in Machine-Readable Dictionaries". *Proceedings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*.

- Peterson, James L. (1980). "Computer Programs for Detecting and Correcting Spelling Errors". *Communications of the ACM*, Dezembro 1980, pp. 676-687.
- Pollock, Joseph J.; Zamora, Antonio (1984). "Automatic Spelling Correction in Scientific and Scholarly Text". *Communications of the ACM*, Abril 1984, pp. 358-368.
- Prószyński, Gábor (1994). "Industrial Applications of Unification Morphology". *Proceedings of the 4th Conference on Applied Natural Language Processing*, ACL, Stuttgart, pp. 213-214.
- Reis, Regina (1993). "Dicionários de Língua Corrente: Algumas Considerações". *Actas do 1º Encontro de Processamento da Língua Portuguesa – EPLP'93*, Lisboa, pp. 141-146.
- Reiter, Ehud; Mellish, Chris; Levine, John (1992). "Automatic Generation of On-Line Documentation in the IDAS Project". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 64-71.
- Richardson, Stephen; Braden-Harder, Lisa (1988). "The Experience of Developing a Large-Scale Natural Language Text Processing System: Critique". *Proceedings of the Second Conference on Applied Natural Language Processing*, ACL, Austin, pp. 195-202.
- Rösner, Dietmar; Stede, Manfred (1994). "TECHDOC: Multilingual Generation of Online and Offline Instructional Text". *Proceedings of the 4th Conference on Applied Natural Language Processing*, ACL, Stuttgart, pp. 209-210.
- Santos, Diana (1988). *A Fase de Transferência num Sistema de Tradução Automática*. Tese de mestrado, Instituto Superior Técnico - Universidade Técnica de Lisboa, Lisboa.
- Santos, Diana (1989). *Notes on PLNLP grammar writing*, unpublished IBM-INESC Report, Lisboa.
- Santos, Diana; Fernandes, Carla; Marques, Rui; Medeiros, José C. (1992). *Gramática sem Dicionário: Relatório Preliminar*, Relatório INESC RT/15-92, Lisboa.
- Santos, Diana (ed.) (1992). *Processamento de Corpora de Texto no INESC, Vol. 1*, Relatório INESC RT/65-92, Lisboa.
- Santos, Diana (1994). "Português Computacional". *Actas do congresso internacional sobre o Português*, Lisboa.
- Slocum, Jonathan (1988). "Morphological Processing in the Nabu System". *Proceedings of the Second Conference on Applied Natural Language Processing*, ACL, Austin, pp. 228-234.
- Vilela, Mário (1990). *Dicionário do Português Básico*, Edições ASA, Porto.
- Vilela, Mário (1994). *Estudos de Lexicologia do Português*, Livraria Almedina, Coimbra.
- Volk, Martin (1992). "The Role of Testing in Grammar Engineering". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 257-258.

Vosse, Theo (1992). "Detecting and Correcting Morpho-syntactic Errors in Real Texts". *Proceedings of the Third Applied Natural Language Processing Conference*, ACL, Trento, pp. 111-118.

Wittmann, Luzia; Pereira, Maria de Jesus (1994). *European and Brazilian Portuguese: Quantitative Report of Lexical, Syntactical and Ortographic Differences*, Relatório INESC, Lisboa.

## Apêndice A - Os corpos de teste

Neste apêndice faz-se uma descrição dos corpos de teste usados tanto no desenvolvimento do Palavroso e do Correcto, como na elaboração desta dissertação.

Existem, pois, duas razões para a compilação de corpos de teste: por um lado, caracterizar e conhecer os tipos de erros que um corrector ortográfico tem que reconhecer e corrigir; por outro lado, avaliar o desempenho dos correctores.

Para a avaliação completa de um corrector ortográfico, são necessários dois tipos de corpos de teste: um, constituído por listas de erros ortográficos com a respectiva correcção; outro, constituído por palavras sem erro. O primeiro serve para avaliar a qualidade das respostas do corrector e o segundo para avaliar a sua cobertura.

### A.1. Corpo sem erros

Durante o desenvolvimento do analisador morfológico Palavroso, usámos um corpo de teste (CorpoTeste) constituído por cerca 103000 ocorrências de palavras (14000 formas diferentes) extraídas de várias fontes:

- ∇ Corpusinesc, corpo de textos do INESC (Santos, 1992)
- ∇ Texto literário obtido da tradução de um livro de um autor americano
- ∇ Texto literário obtido de um livro de um autor português
- ∇ Palavras obtidas de fontes variadas e dispersas

Para no final avaliar o desempenho do Correcto sobre um corpo não usado no seu desempenho, compilámos um outro corpo (CorpoDic), de cerca de 50000 ocorrências de palavras, correspondendo a 7885 formas diferentes. Este foi obtido a partir de um dicionário (Vilela, 1990), retirando das suas entradas frases que exemplificam o uso das palavras, do género "O Ernesto foi para o exame completamente em branco." Estas frases caracterizam-se por usar um vocabulário bastante acessível, serem curtas e, em geral, usarem muitos nomes próprios.



## A.2. Corpo de palavras com erros

Estamos constantemente a ser confrontados com palavras com erro. Aparecem em jornais, revistas, anúncios publicitários ou até mesmo nas legendas de filmes estrangeiros que passam na televisão. No entanto, quando se trata de construir um corpo de palavras com erro, é extremamente difícil encontrá-los em quantidades apreciáveis.

Na recolha que fizemos, usámos essencialmente três tipos de fontes:

- Erros ortográficos compilados e comentados por autores da especialidade.
- Erros ortográficos encontrados em trabalhos de alunos do ensino secundário.
- Erros ortográficos encontrados em mensagens de correio electrónico trocadas entre os investigadores do INESC.

Os trabalhos de Estrela (1983) e Fernandes (1990), embora tratem de erros ortográficos, tendem a reportar-se mais à incorrecta utilização das palavras, do que à incorrecta grafia das mesmas. Uma vez que é este último tipo de erros que nos interessa neste trabalho, estas duas fontes revelaram-se pouco produtivas. Contudo, não deixou de ser uma fonte interessante, e, acima de tudo, formativa.

Em Estrela e Pinto-Correia (1988) encontrámos um levantamento dos principais erros de ortografia, mas foi de Filho (1994) que extraímos a maior parte dos erros ortográficos que constituem o nosso corpo de erros de origem linguística.

Embora as duas primeiras fontes (autores da especialidade e alunos do secundário) tenham fornecido os exemplos mais gritantes (e divertidos), trazem algumas dificuldades. Nomeadamente,

- Só aparecem erros de origem linguística; os erros tipográficos não são utilizados nos estudos dos autores da especialidade e os trabalhos dos alunos eram manuscritos e não dactilografados.
- Só dão um exemplo de cada; não dizem quantas vezes cada erro ocorre. Assim, embora se tenha uma grande diversidade de erros, não sabemos a representatividade de cada um.
- Há uma certa tendência para não fornecerem os casos mais simples e mais correntes, que são os casos de falta de acentos. No caso dos trabalhos dos alunos do secundário, os professores comentaram que se fossem a recolher também esses casos, então acabariam por transcrever os trabalhos...

A terceira fonte, mensagens de correio electrónico, forneceu-nos erros dos três tipos: linguísticos, tipográficos e de transmissão.

Embora tenhamos obtido desta fonte erros em quantidades apreciáveis, verificámos que não é isenta de problemas, com protagonismo para o "problema dos sete bits".

O sistema de correio electrónico mais elementar de que dispomos só processa mensagens escritas em código ASCII de sete bits. A consequência imediata é a impossibilidade de escrever caracteres acentuados. Uma das soluções para os utilizadores é escrever o acento a seguir à vogal. O problema desta solução é que se torna mais aborrecido de escrever e a mensagem perde legibilidade. Entre este tipo de escrita e a escrita de português sem acentos, a maior parte das pessoas opta por esta segunda via, incluindo os acentos apenas em casos de ambiguidade (ex: *e* vs. *e'*).

Concluindo, a maior parte dos erros encontrados nesta fonte consiste em falta de acentos e cedilhas. Se a lista de erros for usada para testar o desempenho dos correctores, há que ter em conta que não reflecte a realidade dos erros ortográficos, a não ser que o corrector seja usado exactamente nesse ambiente.

A vantagem do "problema dos sete bits" é que assim também obtivemos uma lista de erros de transmissão.

Embora todas as palavras com erro possam ser usadas para testar o desempenho dos correctores, em termos de análise e caracterização de erros temos de ter muito cuidado. Por exemplo, não podemos incluir no repertório de erros linguísticos as palavras sem acento que foram obtidas das mensagens de correio electrónico, pois não sabemos se isso se deveu a preguiça ou a ignorância.

Para que não incorrêssemos neste tipo de erro, dividimos as palavras com erro em várias listas, conforme o tipo de erro que tinham. Obtivemos as listas que vêm descritas no quadro A.1.

Os valores do quadro A.1 referentes ao número de palavras reportam-se a ocorrências únicas. A opção por esta via deve-se principalmente a dois factores:

1. Por uma questão de uniformidade. Grande parte dos erros que dispúnhamos consistiam em ocorrências únicas; só os erros obtidos das mensagens de correio electrónico podiam vir expressos em ocorrências reais. Para não estarmos a trabalhar com dois tipos diferentes de dados, optámos por um formato uniforme.
2. As palavras com erro tendem a repetir-se pouco, donde, a diferença entre uma abordagem e outra possa não ser significativa. Repare-se que no que diz respeito a erros tipográficos, é extremamente reduzida a probabilidade de na mesma palavra ocorrer exactamente o mesmo erro tipográfico em alturas diferentes. Adicionalmente, estes aparecem numa proporção muito superior em relação aos erros de origem linguística.



<b>Tipo de erro</b>	<b>Num. de palavras</b>	<b>Denominação do corpo</b>	<b>Descrição</b>
Falta de acento	1235	CorpoCento	Conjunto de palavras cujo único erro consiste na falta de um acento. Foram extraídas de mensagens de correio electrónico, donde não sabemos se a omissão se deve a falha ou a preguiça.
Terminações em <i>ção</i>	505	CorpoÇão	Palavras terminadas em <i>ção</i> ou <i>ções</i> . É o mesmo problema da falta de acento. Cremos que este tipo de erro não é normal na escrita corrente, sendo motivado única e exclusivamente pela preguiça de quem escreve (num ambiente de edição electrónica).
Transmissão	110	CorpoTrans	Conjunto de palavras que contêm erros devidos à conversão das mensagens de correio electrónico de código de oito bits para código de sete bits. Isto redundava em erros de substituição de caracteres.
Tipográficos	312	CorpoTipo	Palavras que contêm algum erro tipográfico, simples ou complexo.
Linguísticos	1581	CorpoLing	Palavras que contêm um erro que cremos ser de origem linguística e não tipográfica.
Total	3743		

Quadro A.1 - Descrição das várias listas de erros.

## Apêndice B - Rótulos usados no Palavroso

Neste apêndice descrevem-se todos os rótulos usados nas tabelas do Palavroso.

<b>Rótulo</b>	<b>Descrição</b>	<b>Rótulo</b>	<b>Descrição</b>
abrev	Abreviatura	int	Pronome interrogativo
adj	Adjectivo	loc	Locução
adjdem	Adjectivo/Demonstrativo	mult	Múltiplo (numeral)
adjind	Adjectivo/Indefinido	nomadj	Nome/Adjectivo
adjrel	Adjectivo/Relativo	nome	Nome
adv	Advérbio	num	Numeral
art	Artigo	obliq	Pronome obliquo
card	Cardinal (numeral)	ord	Ordinal (numeral)
cjs	Conjunção	pes	Pronome pessoal
conj	Conjunção	pont	Sinal de pontuação
cont	Contractão	pos	Pronome possessivo

def	Definido (artigo)	prep	Preposição
dem	Pronome demonstrativo	prop	Próprio (nome)
ij	Interjeição	rel	Pronome relativo
ind	Pronome indefinido	sigla	Sigla ou acrónimo
indef	Indefinido (artigo)	verbo	Verbo

---

Quadro B.1 - Rótulos para identificar as classes e subclasses gramaticais do Palavroso.

<b>Rótulo</b>	<b>Descrição</b>
M	Masculino
F	Feminino
I	Invariável

---

Quadro B.2 - Rótulos usados para identificar o género

<b>Rótulo</b>	<b>Descrição</b>
S	Singular
P	Plural
I	Invariável

Quadro B.3 - Rótulos usados para identificar o número.

<b>Rótulo</b>	<b>Descrição</b>
1	1ª pessoa
2	2ª pessoa
3	3ª pessoa

Quadro B.4 - Rótulos usados para identificar a pessoa (verbos)

<b>Rótulo</b>	<b>Descrição</b>
pi	Presente do Indicativo
pii	Pretérito Imperfeito do Indicativo

ppi	<b>Pretérito Perfeito do Indicativo</b>
pmp	<b>Pretérito-Mais-que-Perfeito</b>
fi	<b>Futuro do Indicativo</b>
pc	<b>Presente do Conjuntivo</b>
pic	<b>Pretérito Imperfeito do Conjuntivo</b>
fc	<b>Futuro do Conjuntivo</b>
i	<b>Imperativo</b>
c	<b>Condicional</b>
ips	<b>Infinitivo Pessoal</b>
ii	<b>Infinitivo Impessoal</b>
p	<b>Particípio Passado</b>
g	<b>Gerúndio</b>

---

Quadro B.5 - Rótulos usados para identificar os tempos dos verbos.

<b>Rótulo</b>	<b>Descrição</b>
---------------	------------------

---

---

Aum	Aumentativo
Dim	Diminutivo
Sup	Superlativo

---

Quadro B.6 - Rótulos usados nos graus dos nomes e adjectivos.