Seminários de Linguística 4 (2000-2001): 67-90. Faro:Univ.Algarve - FCHS/CELL

Um filtro para palavras exóticas frequentes em Português ♦

Jorge Baptista
Univ. Algarve – FSHC*
Centro de Automática da UTL**- LabEL

Luís Faísca
Univ. Algarve – FCHS*

Resumo

As formas gráficas (*tokens*) que constituem as palavras de um texto são muitas vezes ambíguas, podendo frequentemente uma mesma forma corresponder a diferentes flexões de duas ou mais entradas lexicais distintas. Algumas dessas formas correspondem a palavras 'exóticas', isto é, palavras pouco frequentes ou até caídas em desuso. O objectivo deste estudo é a determinação, a partir do corpus do *CETEMPúblico*, das formas ambíguas mais frequentes de palavras exóticas do Português, com vista à construção de um filtro que, durante a fase de análise lexical, elimine o 'ruído' provocado por essas formas exóticas e que permita assim reduzir a ambiguidade formal dos textos, simplificando as fases posteriores do seu processamento automático.

Abstract

The words of a text are often ambiguous, corresponding to inflected forms of two or more distinct lexical entries. Some of these entries are 'exotic' words, i.e., infrequent or rarely used words. The purpose of this paper is to determine the most frequent ambiguous forms of exotic words of Portuguese appearing in the *CETEMPúblico* corpus. Building a filter that will make possible to eliminate the 'noise' caused by those forms will reduce the level of formal ambiguity of texts and, at the same time, it will simplify the following steps of their automatic processing.

1. Introdução.

A análise automática de um texto em linguagem natural passa normalmente por uma fase preliminar de etiquetagem (*tagging*) das formas gráficas (*tokens*) de um texto, atribuindo-lhes a sua categoria gramatical e outra informação morfológica pertinente, recorrendo a um dicionário electrónico [1].

♦ Este estudo foi parcialmente financiado pela Fundação para Ciência e Tecnologia, Ministério da Ciência e Tecnologia.

* Faculdade de Ciências Humanas e Sociais, Campus de Gambelas, P-8000 Faro. *jbaptis@ualg.pt;* ; *lfaisca@ualg.pt* .

** Universidade Técnica de Lisboa - Laboratório de Engenharia da Linguagem Instituto Superior Técnico, Av. Rovisco Pais, P-1049-001 Lisboa. *http://label.ist.utl.pt.*

[1] Para uma perspectiva geral sobre a construção de dicionários electrónicos, veja-se Courtois 1990; Silberztein 1993, 2000; para o Português: Eleutério *et al*. 1995, Ranchhod 1999 e 2001.

Neste processo, verifica-se frequentemente que em Português, como em numerosas outras línguas, uma forma gráfica é formalmente (ortograficamente) ambígua, isto é, pode corresponder a diversas flexões de mais do que uma palavra simples do léxico.

Assim, por exemplo, uma forma como *acordo* receberá as seguintes etiquetas[2]:

acordo, *acordo*. N:ms
acordo, *acordar*. V:P1s

Para analisar uma frase em que a forma *acordo* ocorra, será necessário recorrer à informação associada às formas adjacentes para a desambiguar. Esta tarefa é desempenhada por programas específicos (*parsers*) que visam eliminar a(s) análise(s) incorrecta(s). Por exemplo, se a forma *acordo* ocorrer com um pronome clítico:

Acordo-te às 7:30

a análise *N:ms* pode ser eliminada, já que, *grosso modo*, apenas os verbos podem receber clíticos. Evidentemente, tal desambiguação não é possível quando o pronome se encontra em posição proclítica:

Não te acordo antes das 7:30

1.1. Ambiguidade.

A noção de ambiguidade que vamos aqui utilizar é **estritamente formal** e importa, neste ponto, fazer algumas precisões. É considerada ambígua uma forma que corresponde a flexões formalmente idênticas de pelo menos duas entradas lexicais do dicionário electrónico empregue na análise lexical de um texto. Assim, uma palavra como *candidato* é três vezes ambígua na medida em que a ela podem ser atribuídas as seguintes etiquetas:

candidato, *candidato*. A:ms
candidato, *candidato*. N:ms
candidato, *candidatar*. V:P1s

As primeiras duas linhas correspondem a um tipo de ambiguidade muito frequente em português, os chamados adjectivos-nomes (Casteleiro 1981)[3]. A

[2] Utilizamos o conjunto de etiquetas definidas para o Português pela equipa do LabEL (Ranchhod *et al.* 1999) e integrado no sistema *INTEX* (Silberztein 1993, 2000): a primeira palavra corresponde ao *token* do texto, isto é a uma forma flexionada, e a segunda à entrada lexical que lhe está associada (lema); os códigos *N* e *V* designam a categoria gramatical, nome e verbo, respectivamente; *ms* o género (masculino) e o número (singular); *P* o presente do indicativo, *1s* a pessoa-número (primeira pessoa, singular), etc.

[3] Sobre os problemas relativos à desambiguação de nomes e adjectivos em grupos nominais, veja-se P. Carvalho (em preparação).

terceira linha corresponde à primeira pessoa do singular do presente do indicativo do verbo *candidatar*. Decidimos, neste estudo, tratar apenas a *ambiguidade resultante de palavras com lemas diferentes*, como é o caso neste exemplo.

A partir desta definição, cada palavra pode ser classificada quanto ao seu grau de ambiguidade consoante o número de entradas lexicais que lhe são associadas pelo dicionário utilizado na análise lexical. Assim, a ambiguidade das formas do texto depende em grande medida do grau de cobertura lexical e da granularidade desse dicionário

O dicionário electrónico das palavras simples que utilizámos[4] é constituído por um léxico de grandeza real (mais de 100.000 entradas); na sua construção, registaram-se *todas* as palavras que foi possível recensear em dicionários de uso. O nível de ambiguidade formal observado em Português é elevado: muitas vezes, uma mesma forma corresponde a flexões de diferentes entradas lexicais.

Um outro aspecto da ambiguidade que resulta da aplicação de dicionários electrónicos prende-se com o facto de o conjunto de etiquetas utilizado no dicionário poder ser mais ou menos rico (*granularidade*). No caso do dicionário utilizado, toda a informação gramatical pertinente para a descrição morfológica das palavras está explicitada, o que implica uma elevada granularidade. Isto faz com que, por exemplo, uma forma como *afirma*, que corresponde a diferentes flexões da *mesma* entrada lexical (*afirmar*), tenha 4 análises possíveis [5]:

afirma, *afirmar*. V:P2s:P4s:P3s:Y2s

A ambiguidade formal entre diferentes flexões da *mesma* entrada lexical, no entanto, não será tratada neste trabalho, apenas a ambiguidade entre diferentes entradas lexicais.

Uma vez que o dicionário electrónico é essencialmente um dicionário morfológico, a ambiguidade semântica de uma palavra não está representada, a menos que tal tenha repercussões na sua flexão, o que permite então considerar tratar-se de duas entradas morfológicas distintas. Assim, por exemplo, os substantivos *canto* (da ave) e *canto* (da sala) são ambos masculinos, mas diferenciam-se morfologicamente pelo facto de apenas o segundo aceitar di-

[4] Trata-se do *Dicionário electrónico das palavras simples do Português – DIGRAS* (versão de 22 de Maio de 1999) – do sistema de *Dicionários e Gramáticas Electrónicos do Português – Digrama* (Eleutério *et al*. 1995, Ranchhod *et al*. 1999).

[5] As etiquetas P3s e Y2s correspondem à terceira pessoa do singular do presente do indicativo e à segunda do singular do imperativo, respectivamente; P4s corresponde à forma de tratamento por 'você', formalmente idêntica à terceira pessoa mas com valor de segunda; finalmente, P2s representa a forma *afirmas* com queda do *–s* final quando seguida de clítico: *tu afirma-lo*

minutivo, *cantinho*. No entanto, a ambiguidade de um verbo como *acordar* (o m. q. *despertar* ou *fazer um acordo*) não acarreta diferenças morfológicas, pelo que só poderá ser resolvida aquando da análise sintáctica do texto.

1.2. Formas 'exóticas'.

Ora, há numerosas palavras simples do Português que, sendo bastante frequentes nos textos, apresentam um tipo de ambiguidade que poderia, sem grande prejuízo para a análise, ser facilmente resolvido, já que corresponde, numa das análises possíveis, a palavras pouco frequentes ou totalmente caídas em desuso [6]. Tomemos como caso paradigmático a forma *campo*, que pode receber as seguintes etiquetas:

campo, *campo*. N:ms
campo, *campar*, V:P1s

Parece-nos óbvio que o verbo *campar*, tal como definido num dicionário[7]:

Campar, *v.intr*. ostentar, brilhar; sair-se bem; lucrar; acampar.

é suficientemente pouco usual para que possamos *arbitrariamente* eliminar essa análise sem que a simplificação assim operada interfira de forma significativa no processamento da maioria dos textos.

O mesmo não sucede com a forma *acordo* que vimos acima, em que quer o nome quer o verbo são ambos palavras frequentes da língua.

1.3. Reduzir o 'ruído' controlando o 'silêncio'.

A eliminação do 'ruído' provocado por estas palavras 'exóticas' poderá contribuir significativamente para a simplificação da tarefa de construção de *parsers*, isto é, de programas para a redução de ambiguidade para a análise sintáctica.

A noção de 'ruído' que aqui utilizamos pode ser definida como uma consequência do grau de ambiguidade formal resultante da aplicação do dicionário às formas do texto. Apesar de uma mesma forma poder corresponder a mais do que uma entrada lexical, geralmente só uma ou parte das suas análises é a correcta numa dada frase de um texto. As restantes análises consti-

[6] Para uma discussão e apresentação da metodologia de hierarquização do léxico numa base não estatística, veja-se Garriges (1992), que seguimos de perto neste estudo.

[7] Almeida Costa, J. e A. Sampaio e Mello. 1998. *Dicionário da Língua Portuguesa* (8ª. ed.). Porto: Porto Editora (doravante abreviado *DLP8*). Utilizámos sobretudo o dicionário em suporte magnético *Dicionário Electrónico da Língua Portuguesa – PROfissional*. Porto/Lisboa: Porto Editora/Priberam (a partir de agora, DELP), que se baseia na 6ª edição em suporte papel, tendo-o confrontado com a última edição (ainda não disponível em CD-Rom), quando tal se mostrou necessário.

tuem o ruído. A identificação dessa análise correcta é feita por *parsers* tendo em conta a informação associada à palavra em jogo e às palavras que ocorrem na sua vizinhança (Laporte 2001).

Um *parser* eficiente deverá reduzir esse ruído gerado durante a consulta do dicionário, verificando, numa primeira fase, a informação gramatical associada às formas adjacentes, por forma a eliminar as análises incorrectas (ver exemplo com *acordo*, acima).

Esta noção opõe-se à noção de 'silêncio'. Um *parser* eficiente **não** deverá, em princípio, produzir silêncio, isto é, não deverá eliminar **nenhuma** das análises **possíveis** de uma dada forma num determinado contexto.

A abordagem que aqui apresentamos pretende *reduzir o ruído gerado pelo dicionário produzindo um nível baixo e controlado de silêncio*, na medida em que remove liminarmente análises possíveis, com base na sua inverosimilhança lexical, mas não se baseia na informação disponível para as formas adjacentes. A eliminação dessas análises 'parasitas' passa pela construção de um filtro que, operando antes do dicionário geral, elimine as análises exóticas das formas ambíguas (pelo menos as mais frequentes). Este filtro terá precedência sobre o dicionário geral, pelo que formas já etiquetadas pelo filtro não serão posteriormente analisadas. O emprego de um filtro desta natureza é *facultativo* e depende em grande medida dos fins (e da profundidade/complexidade de análise) a atingir.

O objectivo principal deste trabalho é, pois, o de determinar as formas ambíguas mais frequentes num corpus de grandes dimensões do Português com vista à construção de um filtro que elimine as análises correspondentes a palavras exóticas. Para o efeito, procurámos caracterizar os diferentes níveis de ambiguidade das palavras de um corpus de texto em língua natural (Português) a fim de avaliar o impacto de um filtro de palavras exóticas como processo de redução da ambiguidade.

2. Metodologia.

2.1. Caracterização do corpus e análise lexical.

Para a determinação das formas ambíguas mais frequentes em Português, utilizámos, nesta fase, apenas o primeiro fragmento do corpus *CETEMPúblico* (versão 1.1.)[8]. Este consiste num único ficheiro de cerca de 57,4 Mb, contendo 12.887.079 *tokens* (os quais incluem não só as palavras, mas também os separadores, os números e outros caracteres especiais)

Este corpus foi analisado utilizando o sistema *INTEX* 4.21 (Silberztein 1993, 2000) com o dicionário electrónico das palavras simples do Português (*DIGRAS*) do sistema *DIGRAMA* (Eleutério *et al.* 1995, Ranchhod *et al*. 1999).

O quadro seguinte resume a informação que serviu de base para o presente estudo:

	Ocorrências	Formas diferentes	Entradas lexicais (dicionário do texto)
Tokens	12.887.079		
Palavras gráficas	9.705.674	179.156	
Npr.fst [9] (eliminadas da análise)		- 79.811	
Não reconhecidas (eliminadas da análise)		- 15.969	
Reconhecidas	8.964.411	89.376	102.468

Quadro 1. Caracterização do corpus e análise lexical.

Após a análise formal preliminar, verifica-se que o texto é constituído por 9.705.674 palavras gráficas (sequências de caracteres do alfabeto entre separadores). Destas, 179.156 são formas diferentes.

Há 15.969 formas que não foram reconhecidas pelos dicionários[10]. Trata-se essencialmente de erros ortográficos ou de dactilografia (*abetura*, *aape-*

[8] http://cgi.portugues.mct.pt/cetempublico. Esta versão, obtida pela aplicação de um *patch* à versão 1.0, distribuída pela comunidade científica, foi por nós manualmente corrigida para eliminar os caracteres '{ }', que impediam a sua utilização no *INTEX*. Ao mesmo tempo, eliminámos um pequeno número de extractos ilegíveis. Foram entretanto disponibilizados outros *patch*s que, estando já bastante adiantado este trabalho, não utilizámos.

[9] A utilização de um transdutor (Ranchhod *et al*. 1999) que, após a aplicação do dicionário das palavras simples, etiqueta grosseiramente as palavras com maiúscula inicial como nomes próprios (incluindo topónimos e siglas) permite detectar 79.811 formas que não foram consideradas nas palavras desconhecidas e foram excluídas deste estudo.

[10] O dicionário das palavras simples ignora a oposição entre maiúsculas e minúsculas, pelo que diferentes grafias da mesma palavra (*DE*, *de* e *De*) correspondem a uma entrada lexical: *de*.*PREP*.

nas), neologismos (*abismalmente*), estrangeirismos ou palavras noutras línguas (*abnormally*) e alguns diminutivos estilísticos (*abebiazinha*). Esta listagem de formas desconhecidas será objecto de descrição noutra ocasião.

As 83.376 palavras diferentes reconhecidas pelo dicionário correspondem a 102.648 entradas lexicais (que constituem o dicionário do texto do *corpus*) e cobrem ao todo 8.964.411 formas (92,4 %) do texto. Todas as palavras reconhecidas pelo dicionário são etiquetadas com a informação linguística pertinente: o lema, a categoria gramatical e a indicação dos respectivos valores gramaticais (género, número, tempo, modo, etc.).

2.2. Constituição do filtro de formas exóticas.

Com vista à identificação das palavras exóticas mais frequentes no corpus em estudo, construímos uma tabela a partir do dicionário do texto (tabela **DLF**, v. Anexo), em que indicamos para cada forma reconhecida do texto o seu grau de ambiguidade (AMB), a sua frequência (FRQ) e as etiquetas da categoria gramatical e da informação morfológica que lhe foram atribuídas pelo dicionário. De seguida, percorremos 'manualmente' esta tabela, marcando as formas consideradas exóticas.

3. Resultados.

3.1. Caracterização dos níveis de ambiguidade.

O facto de, às 83.376 formas reconhecidas pelo dicionário, corresponderem 102.648 entradas lexicais no dicionário do texto dá uma primeira medida de *ambiguidade média* de **1,231** por cada forma. No entanto, o grau de ambiguidade varia de forma para forma.

O Quadro 2 caracteriza o grau de ambiguidade das palavras do corpus:

Nº de análises lexicais (Classe de ambiguidade)	Número de palavras	Percentagem
1 (não-ambíguas)	66.316	79,5385%
2	15.109	18,1215%
3	1.731	2,0761%
4	181	0,2171%
5	37	0,0444%
6	2	0,0024%
Total	83.376	100,0000%

Quadro 2. Classes de ambiguidade.

Como se pode verificar, o número de palavras não ambíguas (66.316) representa uma significativa percentagem (79,5 %) das formas diferentes do texto. Há 17.060 palavras, isto é, cerca de 20,5 % de palavras do texto que apresentam algum grau de ambiguidade. O número de palavras com um elevado grau de ambiguidade lexical (quatro, cinco ou seis análises diferentes) é muito reduzido (menos de 0,3 % das palavras diferentes do texto).

Deve-se no entanto salientar que o grau de ambiguidade de uma palavra não está directamente relacionado com o **impacto** que a sua ambiguidade têm sobre a análise do texto. De facto, um pequeno número de palavras com um grau de ambiguidade elevado pode apresentar um número elevado de ocorrências no texto, fazendo aumentar a ambiguidade total.

Vejamos, por exemplo as primeiras cinco palavras mais frequentes no corpus:

Palavra (Tokens)	**Classe de Ambiguidade**	**Frequência**	**% do texto**
de (de, De, DE)	1	466.747	4,81 %
a (a, A)	5	403.265	4,15 %
que (que, Que, QUE)	4	246.418	2,54 %
o (o, O)	4	269.136	2,77 %
e (e,E)	2	226.346	2,33 %

Quadro 3. Cinco palavras mais frequentes do corpus.

Como se pode ver pelo quadro acima, a palavra mais frequente no texto (*de*), embora represente 4,81 % do total das palavras do texto, não é ambígua. Em contrapartida, a palavra *a*, cujo número de ocorrências no texto é de uma grandeza comparável (4,15%) tem um elevado grau de ambiguidade.

Neste sentido, uma caracterização da ambiguidade das palavras de um texto deve ter em conta a frequência com que estas ocorrem. Uma possível medida do *impacto da ambiguidade de uma palavra sobre a análise do texto* poderia ser o produto do seu grau de ambiguidade (AMB) pela sua frequência absoluta (FRQ). À falta de melhor termo, chamamos simplesmente a essa medida produto (PROD).

Na coluna PROD do Quadro 4 apresenta-se o somatório dessa medida de ambiguidade calculado para todos os itens lexicais de cada uma das classes encontradas:

Classe de Ambiguidade (AMB)	**Nº Palavras**	**Frequência total de cada classe (FRQ)**	**PROD (FRQ * AMB)**	**% PROD**
1	66.316	3.948.195	3.948.195	22,2454
2	15.109	2.866.005	5.732.010	32,2960
3	1.731	977.279	2.931.837	16,5189
4	181	729.906	2.919.624	16,4501
5	37	441.464	2.207.320	12,4368
6	2	1.562	9.372	0,0005
Total	83.376	8.964.411	17.748.358	100,0000

Quadro 4. Impacto da ambiguidade das palavras.

Se se levar em conta a frequência das palavras não ambíguas verifica-se que o seu impacto para a ambiguidade total do texto (PROD) representa apenas cerca de 22%. As palavras com ambiguidade 3, 4 e 5, embora constituam conjuntos de dimensões bastante diferentes, têm, pela sua frequência, um impacto comparável na ambiguidade que introduzem na análise do texto. Porém, em cada classe de ambiguidade encontramos diversas situações que podem merecer uma abordagem diferenciada.

Vejamos, por exemplo, a palavra *rés* (66 ocorrências), uma das duas palavras da classe 6 (a outra é *certo*, com 1.496 ocorrências).

Uma rápida verificação dos contextos em que *rés* ocorre permite verificar que em 63 casos se trata do nome composto não-ambíguo *rés-do-chão*. Seria, pois desejável que este composto fosse incluído num dicionário de compostos não-ambíguos. A aplicação deste dicionário durante a fase de pré-processamento do texto resolveria parcialmente a sua ambiguidade (apenas as restantes três ocorrências deveriam ser objecto de desambiguação). A construção de um tal dicionário não é, porém, o nosso objectivo neste momento.

Por outro lado, se se observar o conjunto de etiquetas que *rés* recebe no dicionário:

rés, ré.N:fp
rés, ré.N:mp

rés, rés.A:mfs
rés, rés.ADV

rés, réu.A:fp rés, réu.N:fp

parece-nos evidente que algumas destas entradas (marcadas a cheio) deverão talvez ser repensadas, pois não é evidente que *rés* tenha de facto um emprego quer como adjectivo uniforme, quer como a forma do feminino plural de um adjectivo *réu*, quer ainda como um advérbio, apesar de um dicionário de referência[11] considerar estes valores:

rés: *adjectivo de dois géneros: rente; raso;*
réu: *adjectivo : criminoso; culpado; malévolo;*
rés: *advérbio: cerce; rente a*

Talvez não fosse de todo inadequado actualizar o dicionário eliminando algumas ou mesmo todas estas palavras, mantendo apenas o nome feminino *ré* (parte do barco e feminino do nome *réu*) e o nome masculino *ré* (nota musical). Como evidente, tal tarefa não faz parte do escopo deste trabalho, embora os casos aqui identificados possam contribuir para tais decisões. O filtro que nos propomos construir elimina aquelas análises deixando as outras, que intuitivamente se podem considerar de uso frequente, como pudemos verificar empiricamente.

A importância de levar em conta a medida do impacto da ambiguidade de uma palavra na ambiguidade global do texto (PROD) pode ainda ser ilustrada pelo exemplo seguinte: Na classe 5, com 37 palavras, só a palavra *a* apresenta 403.265 ocorrências, o que corresponde a 91,3% das ocorrências das palavras dessa classe. Qualquer procedimento que se adoptasse no sentido de uma diminuição da ambiguidade de *a* teria um impacto muito maior para a análise do texto do que para qualquer outra das palavras dessa classe de ambiguidade.

3.2. Identificação das palavras exóticas.

Como referimos na metodologia, para a identificação das palavras exóticas com maior impacto na ambiguidade global do texto, organizámos a informação para cada forma reconhecida do texto numa tabela (tabela **DLF**), onde se indicava o seu grau de ambiguidade (AMB), a sua frequência (FRQ), a medida do seu impacto na ambiguidade global do texto (PROD) e as etiquetas da categoria gramatical e da informação morfológica que lhe foram atribuídas pelo dicionário. Seleccionámos apenas as entradas do dicionário em que AMB ≥ 2 (36.332) e ordenámos essa selecção por ordem decrescente da medida de ambiguidade PROD, acima referida. Apresentamos em anexo

[11] Dicionário da Língua Portuguesa, (8ª ed.). Porto: Porto Editora. Edição *on-line: http://www.portoeditora.pt/dol* (11.05.2001)

um extracto das primeiras linhas da tabela **DLF**. A partir dessa tabela é possível identificar as palavras com maior impacto na ambiguidade no texto.

De seguida, percorremos 'manualmente' as primeiras entradas dessa tabela até às formas com PROD = 200, o que corresponde a 4.144 linhas de dicionário, identificando as palavras exóticas mais frequentes no corpus com vista à construção de um filtro que elimine essas análises durante a fase de pré-processamento do texto.

No processo de identificação de palavras exóticas a partir do dicionário das palavras do texto encontrámos algumas regularidades, que descreveremos em seguida, e tomámos certas decisões que cumpre aqui explicitar. Algumas destas formas exóticas são-no sobretudo em função da natureza jornalística do corpus utilizado, pelo que não é linear que a lista obtida possa ser sempre adequada para textos de outra natureza, como, de resto, também não foi nossa intenção. Seria fastidioso explicitar caso a caso as razões que nos levaram a marcar como exótica cada uma das formas que identificámos. Faremos, pois, apenas algumas observações pontuais sobre os aspectos que apresentam alguma generalidade[12]. Descrevemos alguns dos casos mais frequentes e o conjunto de decisões tomadas para a sua selecção.

A maioria das entradas que decidimos marcar como exóticas são-no de facto e não oferecem dúvidas. A título de exemplo, citemos os verbos *antar* (com a forma *antes*), *bispar* (*bispo*), *cincar* (*cinco*), *demasiar* (*demasiado*), *erar* (*era*, *eras*), *estilar* (*estilo*), *estradar* (*estrada*, *estradas*), *farar* (*faro*), *futurar* (*futuro*, *futura*, *futuras*), *machadar* (*machado*), *madeirar* (*madeira*, *madeiras*), *mear* (*meio*, *meia*, *meias*), *padrar* (*padre*, *padres*), *pistar* (*pista*, *pistas*), *positivar* (*positiva*, *positivas*, *positivo*), *respostar* (*resposta*, *respostas*), *surpresar* (*surpresa*, *surpresas*), *trintar* (*trinta*), *umar* (*uma*, *umas*), *vezar* (*vezes*), etc.

Um caso particularmente interessante porque bem ilustrativo dos critérios de constituição das nomenclaturas dos dicionários de uso é o da forma *li*, que no dicionário de referência utilizado é descrita como equivalente ao advérbio *ali*, e como um nome masculino que designa uma medida itinerária chinesa, uma forma de tratamento chinesa e uma pequena moeda chinesa[13]. Como é evidente, nenhum destes valores pode ser considerado de emprego usual e apenas se observa, de facto, a forma do verbo *ler*.

[12] Um relatório técnico (Baptista e Faísca 2001) apresenta a lista integral das formas 'exóticas' e das entradas lexicais que lhes estão associadas. Essa lista estará igualmente disponível na internet em: http://www.ualg.pt/fchs/dlcm/cell/ rt/ling2.htm (Set/2001).

[13] Uma vez que se trata sempre de um nome masculino, o dicionário apenas considera uma entrada morfológica.

Ainda nos advérbios, temos também os exemplos de *i* e de *u* que são formas arcaicas equivalentes aos advérbios *aí* e *onde*, respectivamente.

Ao eliminarmos estas entradas, estamos seguros de que o silêncio que produziremos será ínfimo e perfeitamente negligenciável.

Nos casos em que se levantavam dúvidas quanto à possibilidade de um dada entrada ocorrer efectivamente no texto verificámos as respectivas concordâncias. Quando o fizemos, pudemos verificar que apenas as entradas correspondentes a palavras usuais são de facto as palavras empregues no corpus.

Marcámos ainda como exóticas:

a) várias formas nominais tais como as entradas que designam o nome das letras:

a,a.N:ms, e,e.N:ms, h,h.N:ms, i,i.N:ms, o,o.N:ms, u,u.N:ms

De um modo geral, eliminámos a ambiguidade gerada por palavras de outras categorias que podem ser empregues nominalmente, como é o caso, por exemplo, o pronome *tu* em expressões do tipo:

No poema, o tu a quem o poeta se dirige <...>

Verificámos, nestes casos, que o pronome é efectivamente usado como tal e nunca como um substantivo, embora *tu* esteja registado no dicionário como tal.

Do mesmo modo, eliminámos os substantivos homógrafos de numerais. Efectivamente, qualquer numeral pode ser empregue substantivalmente:

O treze é um número problemático

mas não só estes empregos são raros, como não têm um significado (usual) autónomo do que apresentam enquanto numeral. Trata-se de um processo sintáctico corrente, que Z. Harris (1976, 1991) considera ter subjacente uma frase classificadora, de natureza metalinguística (*O treze é um número* → *O número treze* → *O treze*), não vendo nós razões de fundo para se considerar que estas formas tenham de constituir uma entrada nominal independente[14].

Mesmo outros empregos de numerais, como por exemplo (*tirar* + *perder*) *os três*, *fazer o quatro*, *arranjar um trinta e um*, etc., também seriam mais adequadamente tratados no quadro da descrição das expressões fixas idiomáticas (M. Gross 1982), uma vez que muito do seu valor enquanto entradas lexicais autónomas já se terá perdido.

[14] Para uma discussão deste tipo de fenómenos, veja-se também M. Gross (1999).

Já o mesmo não se passa com certos numerais que se autonomizaram de tal forma do seu emprego original, que devem, de facto, ser considerados novas entradas lexicais, como é o caso de *oitava* (Mús.) e de *nonas* (hora canónica).

b) verbos formados sobre adjectivos/nomes gentílicos: *japonesar* (*japonesa*, *japonesas*, *japoneses*), *inglesar* (*inglesa*, *inglesas*, *ingleses*), *portuguesar* [15] (*portuguesa*, *portuguesas*, *portugueses*); ainda nesta linha, certas variantes de adjectivos/nomes pátrios e gentílicos, tais como *angola*, *suécio* (*suécia*), para alguns dos quais existe uma forma mais usual (*sueco*).

c) certos verbos (geralmente com uma construção causativa, Baptista 2001) que apresentam um forma alternativa, geralmente precedidos do morfema *a-*, como é o caso de *portuguesar* (o m. que *aportuguesar*).

d) certas flexões de verbos que não podem, com propriedade, ser classificados como exóticos, foram por nós marcadas como tal, na medida em que, apesar de serem homógrafas de outras entradas lexicais, são de emprego improvável. É o caso de *para* (conjuntivo de *parir*), *bala* (conjuntivo de *balir*), *intervalo* (indicativo de *intervalar*), *bolsas* (indicativo de *bolsar*), etc. Indicadores desta situação são, por exemplo, as formas da segunda pessoa do plural (*industriais*), de emprego raro na norma-padrão e que correspondem praticamente sempre a outras entradas lexicais. Dado não corresponder exactamente à noção de palavra 'exótica', pelo menos de forma tão clara como os exemplos que demos acima, verificámos nestes casos as concordâncias das ditas formas, a fim de nos certificarmos de que estas entradas não são realmente empregues no corpus, como de facto pudemos confirmar.

e) várias palavras do léxico comum homógrafas de nomes próprios (*morais*) e topónimos (*barreiro*) que foram considerados 'exóticas' quer pela raridade do seu emprego (*ramalho*, *barroso*) quer pelo facto de, no corpus, aparecerem sempre como nomes próprios (*helena*, *isabel*, *saraiva*), como pudemos verificar. Dois casos interessantes de ambiguidade entre nomes de letras e nomes próprios é o do nome *Gama* e o da sigla *ETA*, homógrafos dos nomes das letras do alfabeto grego *gama* e *eta*; pudemos confirmar tratar-se sempre de ocorrências do nome próprio e do nome da organização terrorista.

f) vários prefixos que entram na formação de palavras compostas (*contra-*, *ibero-*, *luso-*, *mal-*, etc.), cuja natureza afixal pode ser identificada como tal no texto recorrendo a uma gramática de desambiguação (cf. Silber-

[15] Obviamente, o verbo *aportuguesar* existe e está bem atestado no corpus e as suas formas flexionadas não são ambíguas. Apenas esta variante levanta ambiguidade com o gentílico.

ztein 1993, 2000; Laporte 2001; M. Gross 2001) relativamente simples, de resto aplicável a outros casos idênticos, na medida em que estes elementos ocorrem sempre ligados por hífen à palavra seguinte.

g) várias interjeições, quando homógrafas de outras entradas lexicais. Na sequência de M. Gross (1982), consideramos que as interjeições devem ser tratadas no quadro das frases fixas. Na medida em que elas formam uma expressão completa e aparecem sempre acompanhadas de um sinal de pontuação específico, como por exemplo, o ponto de exclamação (*Livra!*) ou o ponto de interrogação (*Como?*), uma gramática de tipo idêntico ao referido atrás poderia ser empregue para as desambiguar e etiquetar como tal nos textos em que ocorrem. Algumas, porém, são suficientemente arcaicas para as considerar como exóticas (*Santiago!*).

3.3. Avaliação do impacto do filtro.

A lista de formas marcadas como 'exóticas' tem, neste momento, 560 palavras, que correspondem a 653 entradas lexicais exóticas (v. Anexo). O Quadro 5 apresenta a distribuição das palavras exóticas identificadas, segundo a classe de ambiguidade a que essas formas pertencem.

Classes de Ambiguidade	**Número de análises exóticas eliminadas**				Subtotal
	1	**2**	**3**	**4**	
Classe 2	298	35	0	0	333
Classe 3	139	26	2	0	167
Classe 4	31	10	6	0	47
Classe 5	9	1	1	0	11
Classe 6	1	0	0	1	2
Subtotal	478	72	9	1	Total de Palavras: 560
	478*1=478	72*2=144	9*3=27	1*4=4	Total de Entradas: 653

Quadro 5. Distribuição das palavras exóticas identificadas, segundo a classe de ambiguidade original.

Para a grande maioria das entradas exóticas identificadas (85,4%) verifica-se uma redução da ambiguidade das respectivas formas em apenas um dos seus valores potenciais. A lista permite ainda desambiguar totalmente 330 palavras, que passam a receber apenas uma etiqueta.

Para avaliar o impacto do filtro que elimina estas entradas na desambiguação das formas do texto, temos de considerar a diferença entre a ambiguidade total das formas reconhecidas do texto (somatório do produto da FRQ por AMB para todas as palavras, ou seja 17.748.358) e o valor que representaria o texto totalmente desambiguado e que é equivalente ao somatório da

frequência de todas as palavras reconhecidas, ou seja, 8.964.411. Obtemos pela aplicação do filtro uma redução de 1.497.399 análises 'exóticas', o que representa 17,047 % da ambiguidade que (teoricamente) seria possível reduzir.

O gráfico seguinte apresenta a redução da ambiguidade global do texto à medida que são eliminadas as palavras exóticas, aqui dispostas por ordem decrescente de PROD.

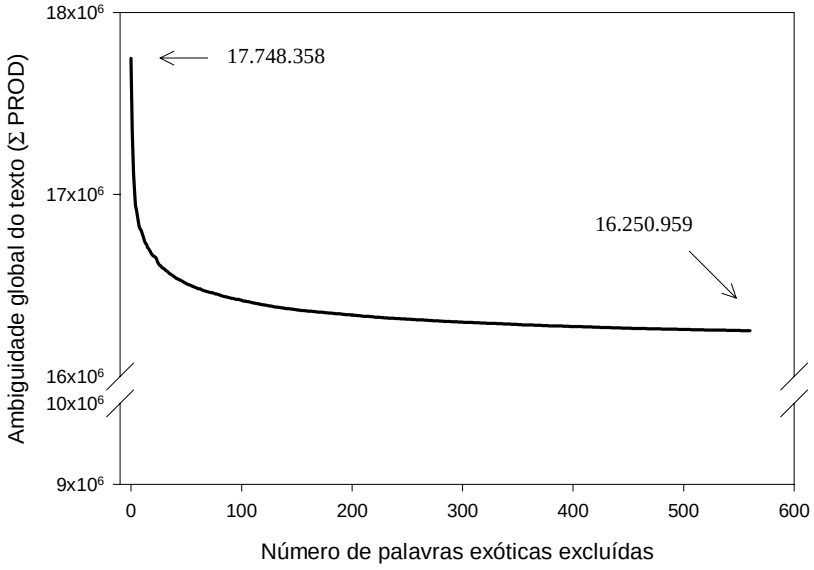

Gráfico 1. Redução da ambiguidade global do texto, mediante a eliminação sucessiva das palavras exóticas, por ordem decrescente de PROD.

Como se pode verificar, a maior quebra a ambiguidade dá-se nas primeiras 100 formas. A partir daí, a taxa redução da ambiguidade diminui significativamente.

Deixamos para outro momento, as restantes 32.188 entradas do dicionário do *corpus* (PROD ≤ 200). Em anexo apresentamos a tabela **DLF (restante)** que apresenta de forma muito simplificada a distribuição das entradas remanescente pelas classes de ambiguidade, com indicação da respectiva frequência no corpus. Todas elas apresentam frequência igual ou inferior a 100 e 6.317 (19,6 %) só ocorrem uma única vez (*hapax*).

Ajustando um modelo de regressão não linear aos níveis de redução observados para as diferentes palavras exóticas que eliminámos[16], é possível predizer qual a redução conseguida pelo eliminação de um número progressivamente maior de formas exóticas. Como podemos constatar pelo gráfico seguinte, eliminar 1000 ou 10000 palavras terá um impacto pouco relevante, pois dificilmente se consegue ultrapassar uma taxa de redução de ambiguidade de 18%.

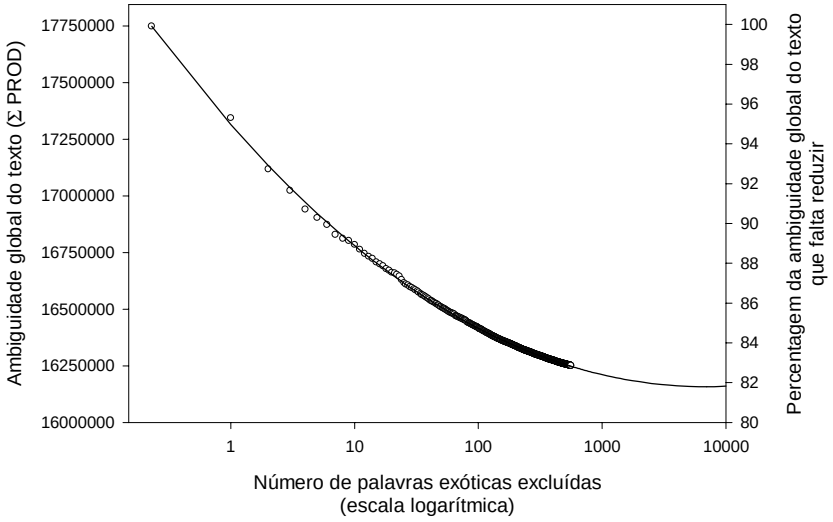

Gráfico 2. Projecção do nível de ambiguidade do texto após excluir 10000 palavras exóticas.

5. Observações finais

Este estudo debruçou-se apenas sobre as primeiras 4.144 entradas do dicionário do texto, por ordem decrescente do produto da ambiguidade pela frequência. A continuação da identificação de palavras 'exóticas' nas restantes 32.188 entradas do dicionário do texto (v. Anexo) que não chegámos a analisar deverá, com certeza, aumentar este valor, não sendo, porém, previsível que se chegue a alcançar uma taxa de redução do ruído muito superior a 18 %, dado que a frequência destas formas vai sendo progressivamente menor.

[16] O modelo de regressão ajustado é um modelo logarítmico quadrático, com a seguinte expressão funcional: logPROD = 16,67-0,01566*logW + 0,0008873*$\log^2$W, onde W corresponde ao número de palavras excluídas. Este modelo apresenta um R^2 = 0,9994, valor que indica um ajustamento óptimo aos dados.

Outros processos de desambiguação ou de redução de ambiguidade (Laporte 2001, Garrigues 1997, Carvalho (em preparação)) têm necessariamente de ser empregues, ou em simultâneo ou em alternativa, mas somos de opinião que o contributo deste tipo de estudos poderá constituir um instrumento útil para a identificação de formas problemáticas do léxico com vista: a) à constituição de métodos de desambiguação ou de redução de ambiguidade mais eficientes e b) ao estabelecimento de subléxicos de natureza menos geral, mas mais adequados à natureza dos textos a processar.

Têm sido dados alguns passos nesse sentido, nomeadamente com a estruturação de dicionários electrónicos de grandeza real em diferentes níveis (Garrigues 1992). A marcação de um dado verbo ou nome como pertencendo a um nível lexical associado à sua raridade, tecnicidade, arcaísmo, etc. é, porém, um trabalho de natureza diferente do que aqui se apresenta (embora talvez possa beneficiar das observações que aqui fizemos).

De facto, as palavras exóticas correspondem por vezes a um subconjunto bastante restrito das formas flexionadas associadas a um lema, que são formalmente ambíguas porque homógrafas de outras entradas lexicais. Em geral, as restantes flexões de uma palavras não são ambíguas pelo que não se torna necessário, se se utilizar um filtro como o que propomos aqui, eliminar essas entradas do dicionário ou restringir o seu uso durante a fase de análise lexical de um texto.

Referências

Almeida Costa, J. e A. Sampaio e Mello. 1998. *Dicionário da Língua Portuguesa* (8ª. ed.). Porto: Porto Editora (*DLP*[8]).

Baptista, J. 1995. *Estabelecimento e formalização de classes de nomes compostos* (Tese de Mestrado). Lisboa: FLUL (242 pp., policopiado).

Baptista, J. 2001. *Sintaxe dos Predicados Nominais construídos com o verbo-suporte SER DE*. Tese de Doutoramento. Faro: Universidade do Algarve (375 pp., policopiado).

Baptista, J. e Faísca, L. 2001. *Um filtro para palavras exóticas frequentes do Português*. Relatório Técnico do CELL/RT/LING/2 (policopiado).

Carvalho, P. (em preparação). *Gramáticas de resolução de ambiguidades no interior de grupos nominais*. Tese de Mestrado. FLUL

Casteleiro, J. 1981. *Sintaxe transformacional do adjectivo*. Lisboa : INIC

Courtois, B. 1990. Le dictionnaire électronique des mots simples. *Langue Française* 87: 11-22. Paris: Larousse.

Dicionário Electrónico da Língua Portuguesa – PROfissional. Porto/Lisboa: Porto Editora/Priberam (*DELP*)

Eleutério, S., E. Ranchhod, H. Freire e J. Baptista. 1995. A System of Electronic Dictionaries of Portuguese. *Linguisticae Investigationes* XIX-1:57-82. Amsterdam: John Benjamins Publishing Company.

Garrigues, M. 1997. Une méthode de désambiguation locale nom/adjectif pour l'analyse automatique de textes. *Langages* 126: 60-70. Paris: Larousse.

Garrigues, M. 1992. Dictionnaires hiérarchiques du français. Principes et méthode d'extraction. *Langue Française* 96: 88-100. Paris: Larousse.

Gross, M. 1982. Une classification des phrase figées du français. *Révue Québecoise de Linguistique* 11-2:151-185. Montréal: Presses de l'Université du Québec à Montréal.

Gross, M. 1999. Nouvelles applications des graphes d'automates finis à la description linguistique. *Linguisticae Investigationes* XXII :249-262. Amsterdam : John Benjamins, B.V.

Harris, Z.S. 1976. *Notes du Cours de Syntaxe*. Paris : Seuil.

Harris, Z. S. 1991. *A Theory of Language and Information. A Mathematical Approach*. Oxford: Clarendon Press.

Laporte, E. Resolução de ambiguidades. *in* Ranchhod, E. (Org.) 2001: 49-90.

Marrafa, P. e M. A. Mota (Orgs.). 1999. *Linguística Computacional. Investigação Fundamental e Aplicações*. Lisboa: APL/Ed. Colibri.

Ranchhod, E. (Org.). 2001. *Tratamento das Línguas por Computador. Uma introdução à Linguística Computacional e suas aplicações*. Lisboa: Caminho.

Ranchhod, E. 1999. Dicionários Electrónicos e Análise Lexical Automática. *in* Marrafa, P. e M. A. Mota (Orgs.). 1999: 221-233.

Ranchhod, E. 2001. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. *in* Ranchhod, E. (Org.) 2001: 13-48.

Ranchhod, E., C. Mota e J. Baptista. 1999. A Computational Lexicon of Portuguese for Automatic Text Parsing. *SIGLEX'99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, College Park, Maryland, USA, pp. 74-81.

Silberztein 1993. *Dictionnaires electronique et analyse automatique de textes. Le système INTEX*. Paris: Masson.

Silberztein, M. 2000. Intex Manual. Paris: LADL/ASSTRIL. http://ladl.univ-mlv.fr/INTEX/intex.html (23-07-2001)

ANEXOS

DLF (Extracto das primeiras entradas)

Extracto das primeiras linhas da lista das entradas do DLF com indicação para cada uma da classe de ambiguidade a que pertence (AMB), da frequência (FRQ) com que ocorre no texto e do produto (PROD) de AMB por FRQ. As entradas estão dispostas por ordem decrescente do valor de PROD.

AMB	FRQ	PROD	DLF	Lema.Categ:InfoGram

5	403265	2016325	***a***	o.PRO+Dem:fs
5	403265	2016325	***a***	eu.PRO+Pes:A4fs:A3fs
5	403265	2016325	***a***	a.PREP
5	403265	2016325	***a***	o.DET+Art+Def:fs
5	403265	2016325	***a***	a.N:ms
4	269136	1076544	***o***	o.N:ms
4	269136	1076544	***o***	o.DET+Art+Def:ms
4	269136	1076544	***o***	o.PRO+Dem:ms
4	269136	1076544	***o***	eu.PRO+Pes:A4ms:A3ms
4	246418	985672	***que***	que.CONJ
4	246418	985672	***que***	que.PRO+Exc
4	246418	985672	***que***	que.PRO+Int
4	246418	985672	***que***	que.PRO+Rel
2	226346	452692	***e***	e.CONJ
2	226346	452692	***e***	e.N:ms
2	168493	336986	***do***	do.PREPXPRO+Dem:ms
2	168493	336986	***do***	do.PREPXDET+Art+Def:ms
2	164513	329026	***da***	do.PREPXPRO+Dem:fs
2	164513	329026	***da***	do.PREPXDET+Art+Def:fs
3	101511	304533	***os***	o.PRO+Dem:mp
3	101511	304533	***os***	eu.PRO+Pes:A4mp:A3mp
3	101511	304533	***os***	o.DET+Art+Def:mp
3	95530	286590	***para***	parar.V:P2s
3	95530	286590	***para***	para.PREP
3	95530	286590	***para***	parir.V:S1s:S4s:S3s:Y4s

DLF (**restante**)

Distribuição das entradas (e correspondente número de formas) no dicionário do texto (i.e. das formas reconhecidas do corpus) que não foram analisadas neste estudo.

	FRQ = 1 (hapax)		**FRQ = 2 - 50**		**FRQ = 51 - 100**		
	Entradas	Formas	Entradas	Formas	Entradas	Formas	Totais Entradas /Formas
AMB = **2**	5.610	2.805	18.284	9.142	2.030	1.015	25.924 12.962
AMB = **3**	633	211	2.685	895	243 (<66)	81	3.561 1.187
AMB = **4**	64	16	320 (<46) [17]	80	0	0	384 96
AMB = **5**	10	2	55 (<17)	11	0	0	65 13
AMB = **6**	0	0	0	0	0	0	0 0
Totais Entradas /Formas	6.317	3.034	23.598	11.209	2.273	1.096	32.188 15.339

[17] Os valores entre parênteses indicam a frequência máxima encontrada para a classe de frequência/ambiguidade considerada.

PALAVRAS EXÓTICAS – LISTA 1

Listagem das formas exóticas identificadas [18]

a	*bala*	*cabe*	*comerciais*	*diplomas*
abertos	*bala*	*cabelo*	*como*	*disco*
academia	*balas*	*cabrita*	*computadores*	*disto*
acidente	*balas*	*cabrita*	*concerto*	*dita*
acidentes	*banda*	*cabrita*	*conde*	*divisas*
aço	*bandas*	*caldeira*	*conjunta*	*documento*
açores	*bandeira*	*calma*	*conjunto*	*dois*
áfrica	*bandeira*	*calmo*	*contínuo*	*dose*
áfrica	*bandeiras*	*cambiais*	*contra*	*doutrina*
africana	*barata*	*camões*	*conversas*	*durante*
africanas	*barra*	*camões*	*copo*	*e*
africano	*barreira*	*campo*	*cordas*	*editora*
agostinho	*barreiras*	*capa*	*corredor*	*editores*
agostinho	*barreiro*	*capucho*	*correia*	*efectiva*
agosto	*barroso*	*capucho*	*correio*	*efectivo*
ala	*barroso*	*carne*	*couto*	*embora*
albufeira	*barulho*	*carneiro*	*crises*	*empresa*
aldeia	*basta*	*carta*	*crises*	*empresas*
alfa	*basto*	*cartas*	*cuba*	*engenharia*
aliança	*basto*	*casamento*	*cuba*	*era*
alma	*bastos*	*cavaco*	*curso*	*eram*
aluno	*bastos*	*cavaleiro*	*checa*	*escócia*
amante	*bate*	*centrais*	*cheque*	*escócia*
amantes	*beira*	*centro*	*cheques*	*escultura*
amaro	*belenenses*	*certa*	*china*	*espanhola*
amaro	*belenenses*	*certas*	*dê*	*espanholas*
amiga	*bento*	*certo*	*decerto*	*esperança*
angola	*bento*	*cia*	*décima*	*esperanças*
angola	*bispo*	*cia*	*décima*	*espinho*
antas	*boca*	*cimento*	*décimo*	*esquadra*
antes	*bocado*	*cinco*	*demasiado*	*esquerda*
areia	*bola*	*cinquenta*	*deputado*	*estilo*
armando	*bolas*	*circuito*	*desfecho*	*estive*
asiática	*bolsas*	*cita*	*dessas*	*estrada*
asilo	*bolso*	*cita*	*desse*	*estradas*
assessores	*bordo*	*cita*	*desses*	*estrangeira*
assessores	*bruno*	*côa*	*devem*	*estrangeiras*
ave	*bruno*	*comandante*	*diferença*	*estrangeiro*
aves	*ca*	*comboio*	*diferenças*	*estrela*
azeite	*ca*	*começa*	*diploma*	*estruturais*

[18] O número de formas repetidas corresponde ao número de entradas do dicionário do texto que pudemos eliminar.

eta
eta
excepto
expo
externas
externo
extrema
fado
faixa
faixas
fala
falam
falo
falsa
falsa
falsas
falso
falso
falso
fará
faro
fatia
fecha
feira
feiras
feita
feriado
ferreira
ferreira
festas
fila
filas
filha
filhas
filho
filipinas
finais
fino
fio
flores
floresta
florestais
florestas
folha
folhas
for
for
fosse
França
França
francisco
francisco
fraude
fronteira
fruto
futura
futuro
gala
galo
gama
garante
general
gera
geral
golfe
golfo
gramas
gramas
grupo
guerrilha
h
h
há
habituais
helena
helena
hora
horta
huambo
huambo
humana
humanas
humano
i
i
ibero
ibero
ideais
ideia
ilha
ilhas
impacte
impacto
imprensa
individuais
industriais
inferno
inglesa
ingleses
instrumento
intervalo
inverno
irá
iria
iriam
isabel
isabel
japonesa
japonesa
japonesas
japonesas
japoneses
jerónimo
jerónimo
jesus
judiciais
justa
justiça
justo
laranja
lata
lata
latas
latas
latina
legais
lenta
lento
li
li
lia
lia
liberais
libra
libras
lido
linda
livre
locais
longa
lote
lotes
lucas
lucas
lusa
lusa
luxo
luzes
machado
madeira
madeiras
madeiras
maia
mais
mal
mal
marginais
marido
marinha
marinho
marinho
marroquino
mas
más
matias
matias
matias
medalha
mediante
mediante
mediante
medicina
medo
mega
meia
meias
meio
melo
melo
mercado
metais
metais
milhar
missa
miúdo
moita
montantes
morais
motivo
moura
muro
musicais
nada
narciso
necessária
negativa
negativas
negativo
neles
norte
nota
nove
objectivo
obra
obras
obriga
oferta
oficiais
oitava
oitavo
oitavo
onde
onze
opa
opa
orçamentais
ordem
ouro
outono
padre
padres
palmas
pano
papa
para
pára
paradoxo
paradoxo
parcela
pares
parlamentares
parlamento
passiva
passivo
pedra
pedras
pela
pelas
pele
pelo
penas
penas
penas
perigo
peseta
pesetas
piano
pina
pina
pinto
pista
placa
placas
planeta
plano
pode
podemos
poeta
poetas
poetas

polaca
policiais
polónia
polónia
ponta
ponte
pontes
ponto
porto
portuguesa
portuguesas
portuguesas
portugueses
positiva
positivas
positivo
postais
postura
presa
presas
presente
presentes
preso
preta
pretexto
primavera
primo
professora
professores
profunda
profundas
profundo
propinas
quadrado
quantia
quarenta
quarenta
quarta
quarto
quatro
queda
queda
quedas
quedas
queijo
quente
quentes
quintas
quis
rainha
rainha
ramalho
ramalho
rapaz
raposo
raposo
rato
realista
rede
redes
regata
regra
relevo
renda
repúblicas
repúblicas
rés
rés
rés
rés
resposta
respostas
restelo
restelo
réus
ritmo
roque
rosa
rota
rotas
roupa
roupas
rua
ruas
sagres
sagres
sai
salariais
santiago
santiago
são
saraiva
secreto
sede
segue
seguem
seguida
segunda
segundo
seis
seres
seria
seriam
sete
sétima
sétima
sétima
sétimo
sétimo
seus
sexto
silva
silveira
silveira
simples
sindicais
sinistro
só
sobre
subscrito
sudoeste
suécia
suécia
suécia
sul
surpresa
surpresas
tabela
tais
tais
tal
tal
tal
tarefa
tarefas
tarifas
telhado
terça
terceira
terceiro
terço
terços
tese
teses
tesouro
timor
topo
torna
torres
transparente
travessa
três
treze
trigo
trigo
trinta
tu
u
u
uma
umas
vacas
vale
valores
vara
varzim
varzim
vasta
vasto
vazio
venda
vendas
vende
vento
vermelha
vermelho
vezes
vice
vida
vidas
vidro
vira
viva
vivas
vive
vivem
vivemos
vivo
vizinha
vizinho
volume
volumes
vulgar

PALAVRAS EXÓTICAS – LISTA 2

Primeiras linhas da lista de palavras exóticas
com a respectiva informação gramatical[19]

DLF	Lema . Categ : InfoGram	AMB	PROD
a	a.N:ms	5	2016325
e	e.N:ms	2	452692
para	parir.V:S1s:S4s:S3s:Y4s	3	286590
uma	umar.V:P4s:P3s:Y2s	3	246906
como	como.INTERJ	4	145568
mas	mas.N:msp	4	129556
mais	mais.N:msp	2	86998
são	são.N:ms	3	51060
segundo	segundar.V:P1s	5	50585
sobre	sobre.N:ms	3	47394
pelo	pelar.V:P1s	2	41786
pela	pelar.V:P2s:P4s:P3s:Y2s	2	38638
só	só.N:mfs	3	35157
onde	onde.ADV	3	31728
há	há.N:ms	2	30772
era	erar.V:P2s:P4s:P3s:Y2s	3	26073
seus	seu.N:mp	3	26037
dois	dois.A:msp	2	25460
contra	contra.PFX	3	23013
durante	durante.N:ms	3	21369
nada	nado.A:fs	5	19780
portuguesa	portuguesar.V:P2s:P4s:P3s:Y2s	4	18568
três	três.N:msp	2	17890
h	h.N:ms	2	16468
h	h.N:msp	2	16468
tal	tal.A:mfs	4	14876
tal	tal.ADV	4	14876
tal	tal.N:ms	4	14876
pode	podar.V:S1s:S4s:S3s:Y4s	2	14822

[19] As entradas encontram-se dispostas por ordem decrescente de PROD.