

# PORTUGUÊS QUANTITATIVO

José Carlos Medeiros

Rui Marques

Diana Santos

INESC

## Resumo

Neste artigo apresentamos algumas medidas quantitativas sobre a língua portuguesa usando uma ferramenta de análise morfológica genérica (o PALAVROSO) e um pequeno *corpus* que coligimos no INESC. Definiremos e mediremos o grau de "preenchimento lexical" do português, assim como o da efectiva ambiguidade entre algumas categorias morfológicas.

Além de descrever a arquitectura do analisador morfológico e a forma como o utilizámos para a construção de dicionários e para a anotação de *corpora*, com esta experiência pretendemos chamar a atenção para os problemas que se colocam, em processamento de linguagem natural, ao estudo do português numa perspectiva de cobertura vasta.

O texto está estruturado do seguinte modo: apresentamos a filosofia e a arquitectura do PALAVROSO, passando a descrever o *corpus* sobre o qual trabalhamos. Definimos e motivamos em seguida as medidas que pretendemos obter. Descrevemos então pormenorizadamente o processo seguido, esclarecendo as opções linguísticas tomadas, concluindo com a apresentação dos resultados obtidos.

## O analisador morfológico

O analisador morfológico PALAVROSO foi concebido para conter informação morfológica independente da sua aplicação, e, em particular, para não depender crucialmente de informação sobre as palavras sobre as quais se deve pronunciar. Este requisito deveu-se a duas razões diferentes: a primeira, de ordem prática, resume-se ao facto de a primeira aplicação do PALAVROSO ter sido a de implementar uma "gramática sem dicionário" [7]; a segunda, e mais importante no contexto deste artigo, prende-se com o nosso interesse na determinação do dicionário "mínimo" para uma dada aplicação.

Vejam os porquê. Em primeiro lugar, e mesmo

que o analisador morfológico cubra todas as categorias gramaticais morfológicamente expressas do português (ou seja, tempo, modo, pessoa, número, género, caso pronominal) e ainda as formas de sufixação produtivas da nossa língua (tais como advérbios em *-mente* ou nomes em *-inho*), tal quantidade de informação não é necessária ou sequer útil como saída para muitas aplicações. Por exemplo: um analisador morfológico usado para um corrector ortográfico simples apenas precisa de dar um veredicto sobre a existência ou não de uma dada palavra, enquanto que um analisador morfológico para uma anotação semi-automática de *corpora* poderá limitar-se a fornecer a categoria morfológica (nome, verbo, etc.). E um analisador morfológico para um analisador sintáctico superficial (usado, por exemplo, em programas de recuperação de informação ou de desambiguação automática recorrendo a dicionários) pode não precisar de distinguir os vários tempos ou mesmo os radicais diferentes correspondentes a uma forma verbal.

Em segundo lugar, pretendemos com esta arquitectura medir objectivamente o salto no desempenho devido à introdução de novas fontes de informação. Assim, o analisador morfológico pode ser invocado parametrizando a informação que nos dá, assim como lhe podemos fornecer diferentes tipos de informação (e, em particular, maior ou menor número de entradas lexicais).

Esta opção, ou seja, a possibilidade de usar o analisador em conjunto com um léxico de tamanho variável, tem duas motivações linguísticas distintas: como sistema dinâmico, nunca é possível cobrir toda a língua; por outro lado, e fora do contexto, um grande número de palavras é realmente ambíguo entre categorias morfológicas distintas, donde a análise morfológica por si só nunca pode fornecer informação sobre a (única) categoria morfológica a que uma determinada palavra pertence: um certo grau de indeterminação tem de ser passado (para tentar ser resolvido) à análise sintáctica... veja-se as palavras *canto*, *vão*, *como*, *amas*, *gema*, por exemplo.

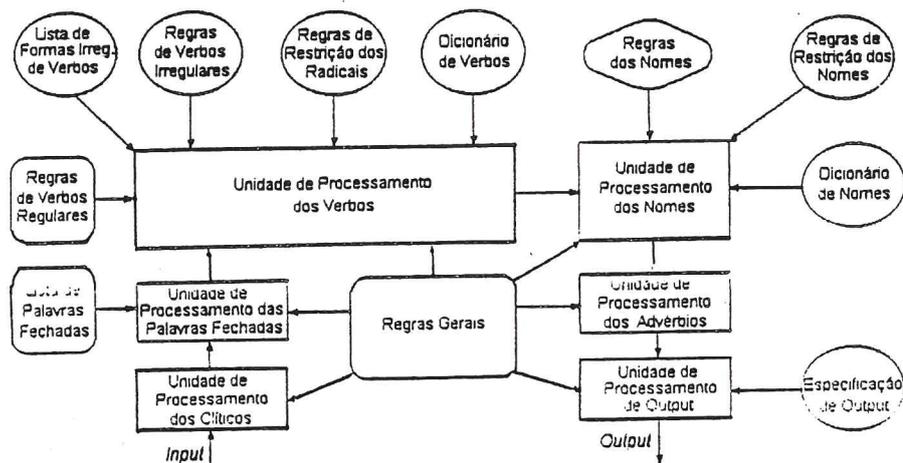


Figura 1: Arquitectura do PALAVROSO

## Arquitectura

Importa realçar que o PALAVROSO não consiste numa aplicação autónoma, que possa ser executada de forma independente, com o seu próprio interface com o utilizador. É sim uma ferramenta para ser inserida noutras aplicações que de alguma forma necessitem efectuar análise morfológica em português.

A sua filosofia genérica está bem evidenciada (veja-se [5]) pelo facto de já ter sido integrado em dois analisadores sintácticos de características diferentes, num "varredor" (browser) inteligente de *corpora*, e no anotador de *corpora* e no produtor de estatísticas referentes a um *corpus* que foram usados para a escrita deste artigo. Em curso encontra-se também a sua utilização para geração.

O sistema está dividido em vários módulos, cada um deles responsável por um tipo de processamento distinto. Associado a cada módulo de processamento podem existir vários ficheiros contendo dados ou regras. A correcta manipulação desses ficheiros por parte do utilizador é que permite otimizar o analisador para um uso específico (veremos adiante alguns exemplos desta manipulação). O esquema da Figura 1 representa os vários módulos e ficheiros a eles associados (para mais detalhes, consulte-se [7, capítulo 4]).

## O corpus

Estando convencidos de que, sem uma apreciação baseada em *corpora*, não existe medida objectiva para decidir incluir determinada informação num programa de processamento de linguagem natural, usámos um corpus de frases<sup>1</sup> escritas soltas recolhidas de texto literário, didáctico, jornalístico, técnico e informal (panfletos publicitários e correspondência

<sup>1</sup>Frase aqui é definida como texto entre dois sinais de pontuação de fim de frase.

electrónica). Informação detalhada encontra-se em [6].

Se bem que a dimensão do *corpus* seja relativamente pequena (14875 ocorrências, correspondendo a 4846 formas distintas), é importante salientar que a verificação e anotação manual (descritas abaixo) seriam impraticáveis em *corpora* maiores. Pensamos portanto que as medidas efectuadas, se não consideradas suficientemente significativas, poderão ao menos servir de padrão para contagens baseadas em *corpora* de dimensão significativamente superior, e também como coadjuvante "dinâmico" das medidas apresentadas em [2].

## Motivação e objectivo

Constatámos que, usando apenas um dicionário mínimo (palavras gramaticais e formas verbais irregulares frequentes), obtínhamos com o PALAVROSO uma determinada classificação morfológica (em excesso), que, em certo número de casos (que nos pareceu ser grande), estava correcta, apesar das palavras testadas não se encontrarem cobertas pelo léxico.

Por isso, e para um texto com uma dimensão razoável, resolvemos introduzir no dicionário apenas informação sobre as palavras que estavam incorrectamente classificadas (por excesso), de forma a produzir uma anotação correcta do *corpus*, e ver qual a diferença de tamanho de dicionário que precisaríamos comparando com a introdução no dicionário de todas as palavras constantes no *corpus*.

Simultaneamente, obteríamos uma primeira medida do grau de ambiguidade lexical (em termos de categorias morfológicas) do português.

Além disso, poderíamos também medir o grau de "aproveitamento" lexical da nossa língua. Este conceito explica-se facilmente recorrendo ao seguinte

exemplo: ainda que uma dada palavra possa ser forma de verbo e de nome, nem todas as formas possíveis correspondem a palavras reais da língua: assim, não existe o verbo *janelar*, donde *janela* é só nome, nem existe o nome *assusta*, fazendo com que esta palavra só seja interpretável como forma do verbo *assustar*.

Formalmente, então, queremos medir

**grau de aproveitamento** dada uma determinada palavra existente no português, qual a probabilidade de ela ser também outra palavra (no sentido estritamente morfosintático, e não semântico), se tal for permitido pelas regras da língua.

**grau de ambiguidade** dado um texto em português, a probabilidade média de uma palavra qualquer ser efectivamente ambígua entre mais do que uma categoria gramatical (separadas em N, V, Adv, e palavras fechadas).

e notar que estas são medidas não do léxico como tal, mas da língua como é utilizada: se uma palavra (ambígua) aparece frequentemente, conta mais para a percentagem do que uma palavra rara não ambígua.

De notar que estas duas medidas são complementares: a primeira, compara o desconhecimento total do léxico com a realidade fora do contexto; a outra compara um conhecimento total do léxico com o seu efectivo uso (em que cada ocorrência tem exactamente uma classificação).

## Descrição do processo

### A classificação inicial

Utilizou-se o PALAVROSO sobre o *corpus* sem adicionar ao dicionário qualquer informação referente às palavras que o continham. O programa criou assim a primeira versão do *corpus* anotado. (Convém lembrar que essa classificação diz respeito apenas a propriedades morfológicas, e que, além disso, não é tomado em conta para a classificação automática o contexto sintáctico em que cada forma ocorre.) Pedimos ao PALAVROSO as seguintes etiquetas (que podem coocorrer):

v a palavra é uma forma verbal diferente de participípio passado.

n a palavra é um nome ou um adjetivo<sup>2</sup>.

vpp a palavra é um participípio passado.

<sup>2</sup>Dado que a maioria dos adjectivos podem também funcionar como nomes e não há, praticamente, critérios de distinção a nível morfológico, decidimos, para o presente trabalho, não fazer uma distinção entre estas duas classes de palavras.

pf a palavra pertence a uma classe fechada. Denominamos os elementos desta classe "palavras fechadas".

adv a palavra é um advérbio<sup>3</sup>.

cl a palavra é um clítico (pronome pessoal átono)<sup>4</sup>.

Nesta classificação inicial, as formas que não pudessem ser classificadas como *verbo* segundo as regras, só eram catalogadas de *nome/adjectivo*. Quando uma palavra tinha uma terminação possível de forma verbal, o programa classificava-a como *verbo*. Obviamente, muitas palavras classificadas como *verbo* não o são de facto. No entanto, pelo menos nenhuma forma verbal deixa de ser classificada como tal.

Nesta primeira fase, o sistema apenas conhecia a lista de "palavras fechadas" do português (preposições, conjunções, pronomes, artigos, contrações de preposição com artigo ou com pronome e advérbios sem a terminação *mente*: 320 entradas, 38 das quais homónimas de palavras de classes produtivas) e uma lista de 156 formas verbais irregulares, tais como *é, foi*.

Além das regras genéricas de formação de tempos verbais, de plurais, etc., o sistema possuía 54 regras de restrição a radicais nominais e ficheiros de regras de restrição de radicais verbais com a seguinte dimensão: **Impossiv**: 84; **Possiv1**: 29 e **Possiv2**: 228 (veja-se [7, capítulo 3]). A explicação do funcionamento destes ficheiros será dada mais à frente.

Finalmente, o sistema continha um dicionário muito reduzido, desenvolvido antes desta experiência e sem recurso ao *corpus*, com 19 nomes e 52 verbos.

Os resultados obtidos encontram-se na tabela da última página, colunas A e B. A primeira contagem refere-se à utilização do sistema sem ligar às listas de nomes e verbos, enquanto que a segunda coluna reflecte o segundo modo de invocar o anotador, em que, no caso de uma palavra se encontrar no dicionário com uma dada classificação, não apresenta as outras opções morfológicamente possíveis.

### O preenchimento do dicionário

As palavras listadas no dicionário foram seleccionadas tendo em conta este *corpus* anotado. De facto, nem todas as palavras que segundo as regras

<sup>3</sup>Ainda que alguns advérbios sejam considerados por nós palavras fechadas, neste artigo, nas tabelas e na discussão subsequente consideramos a classe dos advérbios como separada da classe das palavras fechadas.

<sup>4</sup>Esta classificação apenas é atribuída aos pronomes que ocorrem ligados ao verbo por hífen. Convém dizer que neste caso esta informação é suficiente para classificar inequivocamente a palavra precedente como verbo e o clítico como pronome em oposição a artigo. Por exemplo: em *livro-o*, *livro* não é classificado como nome, nem *o* como artigo.

morfológicas são ambíguas entre, por exemplo, *verbo* e *nome* ou *nome* e *particípio passado* o são na realidade. Para se obter automaticamente uma classificação morfológica correcta do corpus é, então, necessário usar mais informação que as regras da morfologia.

Assim sendo, procurámos as palavras para as quais o programa tinha fornecido mais que uma classificação e verificámos se eram de facto ambíguas. Se o não fossem, adicionaríamos ao sistema, sob a forma de regra, ou apenas à lista de palavras, o necessário para alterar a sua classificação. Vejamos como:

### Ambiguidade V N

Quanto às palavras que, no primeiro corpus anotado, foram classificadas como (v n) (isto é, podendo ser uma forma verbal ou uma forma nominal), há três hipóteses: ou de facto a palavra é morfologicamente ambígua entre N e V (ex: *telefone*); ou a palavra só pode ser uma forma nominal, embora tenha uma terminação exibida por algumas formas verbais (ex: *cadeira*); ou a palavra só é uma forma verbal, embora a sua terminação seja partilhada por formas nominais (ex: *ezisto*).

Na primeira hipótese, não precisamos de listar nada no dicionário, visto que a classificação existente é a pretendida.<sup>5</sup>

Na segunda hipótese, em que uma forma nominal (nome ou adjectivo) é classificada também como verbo, há duas maneiras de solucionar o problema. A primeira delas consiste em listar a terminação do nome, num dos três ficheiros que contêm as regras de restrição dos radicais, indicando as conjugações verbais para as quais essa regra não é válida (i.e., que não têm essa terminação no radical). Esse ficheiro (de agora em diante *Impossiv*) contém, para cada conjugação, as terminações que não ocorrem em nenhum radical de verbo. Este é o processo ideal para os casos em que essa terminação é partilhada por mais que um nome. Com uma única regra pode-se dar conta de vários casos. Um exemplo é fornecido pela terminação "hament". Nos dicionários que consultámos, não está dicionarizado qualquer verbo cujo radical tenha esta terminação. Por outro lado, ela faz parte do radical de vários nomes (*alinhamento*, *definhamento*). Não precisamos, no entanto, de listar todos esses nomes no dicionário, basta listar a impossibilidade de "hament" ser uma terminação de radical verbal em português.

Paralelamente ao ficheiro *Impossiv* funcionam dois ficheiros (*Possiv1* e *Possiv2*), que também

<sup>5</sup> Assumimos, nesta exposição, que o programa produz correctamente os radicais a que as formas pertencem, neste caso: *telefonar* e *telefone*.

contêm regras gerais. Nesses ficheiros são listadas as terminações que podem constituir o fim de um radical verbal. Nos casos em que uma dada terminação só ocorre em poucas formas verbais, é mais económico listarem-se essas formas em *Possiv1* e *Possiv2*. Tomemos como exemplo a terminação de radical *ogi*. Apenas o verbo *elogiar* tem esta terminação. Se isso for indicado nestes ficheiros, precisamos de uma única regra para excluir formas de verbo como *biologia*, *arqueologia*. Se optarmos por listar em *Impossiv* todas as sequências que terminem em *ogi* com excepção de *logi*, precisamos, obviamente, de um número maior de regras.

Um outro caso em que é útil usar os ficheiros *Possiv1* e *Possiv2* é quando uma dada terminação só ocorre num contexto particular. Por exemplo a terminação *l* com verbos da segunda conjugação só ocorre precedida por *a* (como em *valer*) ou por *e* (como em *reler*). É mais prático listar a terminação *l* em *Possiv1* e as terminações *al* e *el* em *Possiv2* do que indicar em *Impossiv* todas as outras sequências que possam preceder *l* em palavras que à partida podiam ser classificadas como formas verbais da segunda conjugação (por exemplo: *aula*, *viola*, *orta*).

A outra maneira de conseguir uma classificação correcta consiste em listar no dicionário de nomes a palavra em causa. Esta metodologia é preferível nos casos em que existe um único nome com essa terminação. Veja-se, por exemplo a palavra *clarão*. Não existe outro nome que termine em *larão*. Por outro lado a terminação *larão* não é exclusiva da classe dos nomes (veja-se *salarão*, *calarão*). Ou seja, não se pode listar a forma *l* em *Impossiv* (visto que isso iria provocar subgeração) e se se listar *cl* essa regra cobrirá um único caso. É preferível, então optar por listar a palavra *clarão* no dicionário de nomes, reservando o *Impossiv* para regras mais gerais.

No caso de a palavra classificada como ambígua entre nome e verbo ser uma forma verbal, a solução é análoga. Pode-se listar no dicionário de verbos o infinitivo da forma verbal em causa, ou adicionar regras de restrição a formas nominais.

Em resumo, podemos dizer que existem dois ficheiros que funcionam como qualquer dicionário tradicional (i.e. onde são listadas as palavras completas), o dicionário de nomes e o dicionário de verbos, e outros ficheiros, onde se listam regras de cobertura mais larga, organizadas por conjugação quando se referem a verbos (i.e., as regras explicitam a que conjugação se referem — *ar*, *er* ou *ir*).

### Ambiguidade V VPP N

As palavras com a terminação *ado*, sem informação lexical, são classificadas como podendo pertencer a três classes diferentes: uma forma verbal de um verbo cujo radical termine em *ad*; um particípio pas-

sado ou uma palavra da classe dos nomes e adjetivos.

Como não há muitos verbos cujo radical termine em *ad*, optámos por listar esse pequeno número com o auxílio dos ficheiros *Possiv1* e *Possiv2*. Isso exclui a hipótese de uma forma que termine em *ado*, cujo radical não esteja dicionarizado nesse ficheiro, seja classificada como forma verbal. A sua classificação resume-se, assim a VPP N. Ou seja, são classificadas como podendo ser um particípio passado ou uma palavra da classe dos nomes e adjetivos. Se, na bibliografia consultada ([8,4]), essa palavra estiver dicionarizada como nome ou adjetivo e se o verbo de que ela pode ser particípio passado também estiver dicionarizado, a classificação é considerada correcta. É o caso de formas como *estado*, *passado*, *pesado*.

Se, porém, o infinitivo do verbo correspondente a essa forma não constituir uma entrada de dicionário, pretendemos que apenas a etiqueta N seja atribuída a essa forma. Tal objectivo atinge-se listando essa palavra no dicionário de nomes. É o caso de *supermercado*, *lado*, *proletariado*.

No caso inverso, encontramos nos dicionários consultados o verbo cujo particípio passado a palavra possa ser, mas não encontramos essa palavra listada como nome ou adjetivo. Neste caso, pretendemos que seja atribuída a etiqueta VPP, mas não N à forma em causa. Para se conseguir isso, listamos no dicionário de verbos o infinitivo do verbo em causa. É o caso de formas como *comprado*, *empurrado*, *voltado*.

### Ambiguidade na mesma classe

Nos casos abordados até agora, as palavras homónimas pertenciam a diferentes classes de palavras. Pode, porém dar-se o caso de uma dada ocorrência ser homónima de uma outra que faz parte da mesma classe de palavras. Um exemplo paradigmático é a forma *vendo*. Pode-se tratar de uma forma do verbo *vendar*, de uma forma do verbo *vender* ou ainda de uma forma do verbo *ver*. Em qualquer dos casos será sempre uma forma verbal.

Este tipo de ambiguidades não foi contemplado no trabalho actual. Assim, aos vocábulos que estão nestas condições é atribuída uma única etiqueta. No corpus anotado, não há, portanto, indicação do número de paradigmas a que a ocorrência possa pertencer. No caso concreto, à forma *vendo* o programa atribui apenas a saída V.

Em trabalho futuro poder-se-á ter em conta este tipo de ambiguidade. Pode-se alterar o programa de forma a que, por exemplo à palavra *vendo* seja atribuída a classificação (V V V), indicando que essa forma pode ser a realização de três formas de verbos diferentes.

### Palavras Fechadas

Ao contrário dos nomes, verbos e particípios passados, a classe de palavras fechadas forma um conjunto limitado. Essas palavras foram, por isso, exaustivamente listadas num de dois ficheiros, consoante tenham ou não uma palavra homónima numa outra classe. Assim, por exemplo à forma *como* o programa atribui a etiqueta (PF V), indicando que se pode tratar de uma ocorrência da preposição *como* ou de uma forma do verbo *comer*. Já a outras palavras (por exemplo *eles*) a etiqueta que o programa atribui é simplesmente (PF). Tais ocorrências não são (morfologicamente) ambíguas entre várias classes de palavras.

### Advérbios

Da classe dos advérbios, apenas os que apresentam a terminação *mente* constituem um tipo produtivo. Os outros advérbios são, pois, tratados como as (restantes) palavras fechadas. Os advérbios em *mente* são, contudo, em número (pelo menos teoricamente) ilimitado. Não se podem, por isso, listar juntamente com os outros advérbios. Por outro lado a terminação *mente* não é exclusiva dos advérbios, sendo partilhada pela classe dos nomes e adjetivos (ex: *semente*, *demente*), bem como pela classe dos verbos (ex: *desmente*, *comente*).

A solução adoptada foi a de listar as palavras da classe dos nomes/adjectivos e da classe dos verbos, visto estas serem em número reduzido. A todas as outras formas que terminem por *mente* o programa atribui apenas a etiqueta (ADV).

### A classificação final

Com o dicionário preenchido da forma acima descrita, o programa fez uma outra classificação do mesmo corpus, a que chamamos o segundo corpus anotado. Diferenciamos na tabela seguinte a classificação proveniente de regras apenas (C) daquela que também faz uso do dicionário (D).

Neste momento, a dimensão das nossas regras de restrição de radicais tinha aumentado para: *Impossiv*: 424, *Possiv1*: 50, *Possiv2*: 289, e os dicionários de nomes e verbos continham exactamente 859 e 575 entradas respectivamente. 176 clícticos não foram contabilizados na tabela.

### Conclusão

Medimos a ambiguidade em termos de partes de oração (v, n/a, vpp, adv e pf), verificando que 12097 ocorrências apenas têm uma interpretação, 1503 têm duas, 37 têm três, e 8 têm quatro, o que em média dá 1.02494 classificações por ocorrência.

Tipo	A	B	C	D
v	693	1169	693	1808
n	2638	2638	3283	4610
vpp	0	53	0	287
adv	730	730	730	730
pf	5863	5882	5863	5876
n v	3982	3506	3321	917
n adv	23	105	55	105
n pf	26	12	26	37
n vpp	108	101	447	167
v adv	0	9	0	9
v pf	0	262	0	231
adv pf	42	42	42	42
v vpp	8	2	8	5
n v vpp	368	328	45	3
n v adv	91	0	59	0
n v pf	290	23	290	35
n v adv pf	8	8	8	8

No entanto, se descontarmos a contribuição das palavras fechadas, já que qualquer sistema de processamento de linguagem natural as deverá conter, com dicionário reduzido ou não, e se considerarmos que o sistema também classifica correctamente os advérbios, obtemos que cada ocorrência (que não seja ou possa ser uma palavra fechada) tem em média em português 1.1398 classificações (relembrando que agora o máximo seriam 3).

Notemos que esta medida é um limite inferior, visto que poderíamos ter tomado em conta, como foi dito atrás, os casos de ambiguidade dentro de cada classe (em que *vendo* pode ser forma dos verbos *ver*, *vender*, *vendar*, por exemplo) ou mesmo por classificação (as formas das primeira e terceira pessoas do singular do imperfeito do indicativo do verbo *vender* contariam por dois...).

Em termos do preenchimento lexical, é claro que as nossas regras de restrição lexical já em si englobam implicitamente uma fatia do conhecimento do que é um verbo típico do português, um nome típico, etc. Mesmo assim, pensamos que é interessante medir a diferença entre o estado inicial descrito neste artigo, em que o conhecimento provinha apenas da consulta a gramáticas e a dicionários inversos, e a classificação de facto encontrada no *corpus*. Mais uma vez ignorando as palavras fechadas e os advérbios, verificamos que o número total de verbos possíveis (por oposição aos reais) é de 5440 (contra 3016), de nomes possíveis 7534 (5882 reais), verbos no particípio passado 484 (462 reais), o que resulta em 55.44%, 78.07% e 95.45% para verbos, nomes e particípios passados respectivamente. Procedendo à mesma experiência mas agora sem qualquer informação de restrição quer de nomes quer de verbos, os números obtidos foram: 52.98%

(verbos), 78.07% (nomes) e 94.67% (part. pass.).

A tarefa que se segue é evidentemente a anotação total do *corpus*, entrando em conta com o contexto sintáctico, de forma a contabilizar qual a "direcção" da ambiguidade observada (ou seja, não contar apenas, no caso (v n pf), "3 para 1", mas contar sim "v pode ser mais 2").

Pensamos que estes estudos preliminares podem ser vistos como o preâmbulo do desenvolvimento, para o português, de sistemas semelhantes aos desenvolvidos por Church [1] e Ejerhed [3].

Finalmente, pensamos que valia a pena investigar a forma de verificar estas medidas usando *corpora* maiores, usando este *corpus* e dicionário à medida como padrão.

## Referências

- [1] Kenneth Ward Church. "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the 2nd Conference on Applied NLP* (Austin, 9-12 February 1988), pp.136-43.
- [2] Fernanda Bacelar et al., "Ambiguidade morfológica no Português Fundamental", este volume.
- [3] Eva I. Ejerhed. "Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods", *Proceedings of the 2nd Conference on Applied NLP*, pp.219-27.
- [4] Cândido de Figueiredo. *Grande dicionário da língua portuguesa*, Bertrand Editora, 23 ed., Lisboa, 1986.
- [5] José Carlos Medeiros. "Ferramentas de processamento de corpora usando o PALAVROSO", [6].
- [6] Diana Santos (ed.). *Processamento de corpora no INESC*. Relatório INESC.
- [7] Diana Santos, Carla Fernandes, Rui Marques & José Carlos Medeiros. "Gramática sem dicionário: Relatório Preliminar". Relatório INESC num. RT/15-92.
- [8] E.M. Wolf et al. *Dicionário inverso da língua portuguesa*. Moscovo, 1971.

## Agradecimento

Queremos agradecer a Fernanda Bacelar e a Fernando Pereira as suas valiosas sugestões e críticas construtivas. Agradecemos também a ajuda prestada, na classificação do *corpus*, por Ana Teresa Alves e Maria do Céu Novais.