Português Computacional

Diana Santos*
Grupo de Linguagem Natural do INESC

Esta comunicação tem como objectivo descrever algumas das questões que se colocam no desenvolvimento de ferramentas computacionais que tratem o português, assim como alguns estudos que estas ferramentas computacionais permitem.
O campo ilustrado é o do léxico e da morfologia, e os assuntos referidos, a título de exemplo, são a dualidade flexão/derivação e a escolha do léxico.
Os estudos efectuados com base nas ferramentas mencionadas versam a ambiguidade categorial do português, as causas da homografia, e o peso relativo das regras morfológicas.
Finalmente, a comunicação acaba com uma variação sobre o tema do ensino do português.

1. Introdução

O Grupo de Linguagem Natural do INESC acabou há pouco o desenvolvimento de um analisador morfológico e correspondente dicionário, o PALAVROSO, cobrindo a morfologia flexional e alguma derivacional. Além disso, trata correctamente palavras hifenizadas, nomeadamente enclíticos e palavras compostas por justaposição, e fornece uma resposta inteligente quando as palavras não se encontram no dicionário. Neste momento contém 57000 entradas, abarcando mais de um milhão e trezentas mil (1300000) formas. A forma de as especificar e as opções linguísticas tomadas encontram-se descritas em Barreiro et al. (1993).

Além de ser usado em várias aplicações e ser uma peça integrante de uma gramática de português de cobertura vasta em desenvolvimento no Grupo, o PALAVROSO foi desenhado de forma a permitir extensa experimentação linguística[1]. São alguns destes estudos que passamos a descrever.

Esta comunicação contém duas partes. A primeira é sobre alguns problemas que se deparam aos Engenheiros da Linguagem na descrição da nossa língua em computador. A segunda parte descreve alguns estudos que realizámos, para os quais o acesso a um sistema de cobertura vasta era essencial. Em jeito de conclusão, limitamo-nos a descrever algumas aplicações do nosso sistema ao ensino da língua portuguesa.

---

* Este artigo perspectiva e descreve o trabalho de vários investigadores do Grupo de Linguagem Natural do INESC, que por isso merecem menção detalhada: os critérios linguísticos e o desenvolvimento do léxico do Palavroso devem-se principalmente a Maria de Jesus Pereira e a Anabela Barreiro, os estudos sobre homografia e anotação do corpus foram conduzidos por Rui Marques, e a obtenção de estatísticas sobre a aplicação das regras é da exclusiva responsabilidade de José Carlos Medeiros. A autora participou em maior ou menor grau nas várias tarefas aqui descritas, sobretudo ao nível da discussão dos caminhos a seguir. Este artigo não seria possível sem a actual implementação, levada a cabo por José Carlos Medeiros.

[1]Para descrição das aplicações, veja-se Santos et al. (1992), Santos (1992), Medeiros (1992, em prep.), Fonseca (1993), Marques (1994), Carvalho et al. (1994).

2. Cobertura larga: que implicações?

Quando dizemos que estamos interessados em cobrir o português TODO, esta é mais uma perspectiva metodológica que quantitativa. É óbvio que uma língua viva não se pode ter toda -- mas não queremos à partida excluir nada.

Para abranger o léxico e a morfologia do português numa tal perspectiva, os dicionários existentes são insuficientes, assim como as gramáticas. Sobre os primeiros fizemos uma pequena experiência descrita em Reis (1993), sobre as segundas convém não esquecer que são as generalizações e os casos paradigmáticos que importa ao gramático registar[2]. Ao querer processar a língua tal como ela é utilizada, é preciso ter em conta as excepções e os casos menos claros, e qualquer trabalho que se diga de cobertura vasta tem forçosamente de proceder a uma normalização, ainda que por normalização não se entenda prescrição. Veja-se, por exemplo, Engh (1993) sobre este assunto.

Assim, a primeira parte desta comunicação tem como objectivo chamar a atenção para a (grande) zona cinzenta entre o preto e o branco que existe entre a gramática descrita pelos teóricos e teorizadores e as enunciações claramente inaceitáveis na língua portuguesa. Como é óbvio, vamos apenas apresentar problemas que se nos depararam durante o desenvolvimento do PALAVROSO, e que pertencem portanto aos foros lexical e morfológico.

2.1 Derivação ou flexão?

Um primeiro exemplo é o estatuto de alguma sufixação em português: debrucemo-nos sobre os diminutivos (de adjectivos ou nomes) formados com *-inho* ou *-zinho* e sobre os advérbios terminados em *-mente*. No primeiro caso, estamos perante um fenómeno extremamente frequente em português, ainda que mais comum na linguagem oral. Encontramo-nos perante um processo de derivação, extremamente produtivo? Ou o diminutivo em *-inho* é uma flexão da categoria dos nomes e dos adjectivos? Que esta questão não é de fácil resolução prova-o a proposta de Mateus et al. (1989:382ff) de que não seja nem flexão nem derivação (veja-se também Lyons (1968, 1977) sobre as relações complexas entre estes dois processos morfológicos). O problema, contudo, não é teórico - ou pelo menos não é essa a vertente que nos interessa aqui. Ao desenvolver um sistema computacional que analisa automaticamente qualquer forma da língua portuguesa, temos de decidir se aceitamos qualquer diminutivo (desde que de acordo com critérios fonéticos), ou se, ao contrário, só aceitamos um número fixo de palavras com diminutivos (derivadas pois da palavra da mesma classe sem diminutivo[3]). A primeira estratégia corresponde a

[2] Veja-se a esse respeito também Bacelar do Nascimento et al. (1993).

[3] Por exemplo, tanto *mesinha* como *mesa* seriam entradas do dicionário.

tratar os diminutivos como flexão (aceitando embora casos 'defectivos'[4]), a segunda como derivação.

Para aqueles de entre vós que me estão a ouvir, convencidos de que a solução é óbvia, e que é a de tratar os diminutivos como derivação, como aliás foi feito por outros sistemas computacionais, deixem-me apenas apontar dois possíveis argumentos contra: o caso do grau superlativo dos adjectivos, menos comum na prática que o diminutivo, é em geral tratado pelas gramáticas como um caso óbvio de flexão[5]. Por outro lado, reputados gramáticos como Cunha e Cintra (1987:180) falam do grau diminutivo de nomes (substantivos), apontando pois para a flexão também neste caso.

No segundo caso, é indiscutível a existência de um grande número de advérbios relacionados com um dado adjectivo. Como aliás o sufixo *-mente* muda a categoria sintáctica da palavra, parece neste caso não haver dúvida de que se trata de um processo de derivação[6]. De acordo com essa visão, tratámos os advérbios em *-mente* como entradas lexicais independentes dos adjectivos de que derivam. No entanto, gostávamos de salientar que esta abordagem tem os seus problemas: e o principal é que vai contra o uso de alguns dicionários correntes. Com efeito, em Costa e Melo (1992), advérbios em *-mente* derivados de forma regular não constituem entrada, o que significa que temos de ser nós a decidir quais os advérbios em *-mente* que o (nosso) português aceita ou não.

2.2 Cobertura lexical

Na mesma veia, a escolha dos elementos a cobrir pelo léxico e (implicitamente) daqueles que não vamos admitir não é coisa fácil. Esta afirmação é conhecida de todos e prende-se com tarefas tão difíceis como a definição de terminologias, a aceitação de estrangeirismos e neologismos, a cobertura dialectal, ou a escolha de vocabulários básicos.

No entanto, mesmo fora destas áreas "quentes" da lexicologia, apresentam-se-nos casos que não são possivelmente óbvios para aqueles que nunca estiveram envolvidos nesta tarefa. Seleccionámos pois a (re)definição de nomes designando profissões, tradicionalmente descritos nos dicionários como substantivos masculinos, e que decidimos considerar também femininos (*pediatra*, *presidente*). Até aqui, tudo bem, mas o que fazer quando as profissões designam actividades existentes antes da "igualdade" entre os sexos, tal como *gladiadora* ou (no sentido inverso) *carpidor*?

---

[4] Claro que esta afirmação não significa que todas as palavras têm de ter um diminutivo. Por exemplo, há verbos que, segundo alguns autores, não têm algumas formas em alguns tempos (por exemplo, *falir*) ou adjectivos que não aceitam comparação (como, por exemplo, *alegado*). Tal não implica que se ponha em causa a flexão em tempo, pessoa e número para os verbos ou em grau para os adjectivos.

[5] Com excepção de Mateus et al. (1989).

[6] Alina Vilalva chamou-me a atenção para o facto de que uma análise mais correcta destes processos involve a noção de composição, como é proposto em Vilalva (1992). Contudo, parece-me que o problema discutido no artigo, ou seja, o tratamento destas palavras por regra ou por incorporação no dicionário se mantém, razão porque escolhi não alterar significativamente o texto.

Não se trata aqui de deitar fora do léxico da língua portuguesa palavras que caíram em desuso ou que sempre foram raras, problema esse que também se coloca no desenho de um sistema computacional que tenha, por exemplo, a tarefa de correcção ortográfica. *Carpideira* e *gladiador* são palavras, diríamos, de uso generalizado e que ninguém se lembraria de decretar não pertencerem à nossa língua.

2.3 Problema geral

De facto, o problema mais genérico que se coloca em relação à cobertura vasta de uma língua é: qual a opção a tomar nos casos (muitos) em que ou não há fontes ou elas se contradizem? Tomemos como último exemplo a conjugação dos verbos: nos dois dicionários de verbos conjugados que consultámos, a forma da primeira pessoa do plural do presente do conjuntivo do verbo *ler* é *leamos* (Nogueira, sem data) ou *leiamos* (Silva e Tavares, 1988). Decidimos (ao contrário da grande maioria dos sistemas existentes[7]) aceitar as duas. O mesmo para os particípios passados duplos.

Este problema é não só importante para os engenheiros da linguagem, mas também para os professores de português, que ensinam a nossa língua. Algumas perguntas que, mais do que a elas ter resposta, gostaríamos de lançar como ilustração: É *candeeirinho* ou *candeeirozinho*[8]? Pode-se dizer *atrasadamente*? Qual é a forma da primeira pessoa do presente do conjuntivo do verbo *colorar*? É *peixe-espadazinho* ou *peixe-espadinha* ou *peixinho-espada*? Qual o superlativo absoluto simples do adjectivo *malévolo*?

3. Estudos quantitativos

Os problemas descritos acima põem-se ao tentar formalizar a descrição da língua num computador. Em menor grau - ou pelo menos em teoria - os mesmos problemas põem-se aos lexicólogos, lexicógrafos e gramáticos, ainda que a validação e verificação da consistência das suas teorias não seja fácil sem recurso ao auxílio de um computador.

Contudo, há certos estudos, impossíveis de realizar sem meios informáticos, que estão mesmo a pedir para serem feitos logo que se possua um instrumento como o PALAVROSO. Dentre o enorme número daqueles que se sugerem, e dada a escassez de tempo e recursos, realizámos apenas alguns que descrevemos em seguida, e apelamos a todos aqueles que queiram aprofundar questões desta natureza que nos contactem.

3.1 Ambiguidade morfológica do português

---

[7] Veja-se por exemplo o Aurélio Eletrônico (1993).

[8] Esta pergunta pode não fazer sentido... Ambas as formas são válidas, mas então como explicar as preferências, facilmente descortináveis na nossa correcção de crianças e estrangeiros? E veja-se *dorzinha* e *dorinha*, a segunda das quais inaceitável, mas não *florzinha* e *florinha*.

Em Medeiros et al. (1993), apresentámos as primeiras experiências que levámos a cabo ainda durante o desenvolvimento do PALAVROSO. Em particular, procedemos a uma primeira medida da ambiguidade morfológica do português, entre as categorias **n/a**, **pf**, **v**, **vpp** e **adv** (nome/adjectivo, palavra gramatical, forma verbal excepto particípio passado, particípio passado e advérbio), sem entrar em conta com ambiguidade intracategorial (por exemplo, o facto da forma *vendo* ser uma forma verbal de três verbos diferentes não foi levada em conta). Esta medida deve portanto ser considerada por defeito.

Neste artigo, apresentamos alguns dados que não tínhamos disponíveis na altura, ou seja, o resultado da anotação em contexto do corpus utilizado, em que além disso se adicionou a etiqueta **a** (adjectivo).

O nosso corpus continha então 14870 ocorrências (4846 formas distintas). A sua distribuição por categorias é pois a seguinte: 6211 palavras gramaticais[9], 1956 verbos, 5540 nomes e adjectivos, 835 advérbios e 328 particípios passados. Da comparação entre a anotação manual e a classificação automática, obtivemos os seguintes resultados:

Tipo		Mais freq.		Menos freq.
n/a v	917	658 n	118 a	141 v
n/a adv	105	83 adv	12 n	10 a
n/a pf	37	37 pf		
n/a vpp	167	80 n	50 a	37 vpp
v adv	9	8 adv	1 v	
v pf	231	227 pf	4 v	
adv pf	42	28 pf	14 adv	
v vpp	5	3 vpp	2 v	
n/a v vpp	3	2 n	1 vpp	
n/a v pf	35	35 pf		
n/a v adv pf	8	8 pf		

Grosso modo, e debruçando-nos sobre a dicotomia nominal verbal (não contando portanto com as palavras de classes fechadas nem com advérbios) verificámos que existiam aproximadamente duas vezes e meia mais palavras do campo nominal do que do verbal. Do mesmo modo, uma palavra ambígua entre nome/adjectivo por um lado e verbo por outro é nome/adjectivo com probabilidade 84%. De facto, neste corpus obtiveram-se as probabilidades 68% para nome, 16% para verbo e 15% para adjectivo.

3.2 Causas da homografia

[9] Ou seja, conjunções, preposições, pronomes, contracções, artigos, etc..

Ainda que o estudo anterior tenha interesse de forma a prever, em traços largos, quais as características do português para o processamento automático, é demasiado grosseiro para que se possam tirar conclusões sobre a estrutura do português. Procedemos pois a um estudo mais fino da homografia, de forma a elucidar possíveis regras de derivação ou apenas relações sistemáticas entre as várias análises de uma mesma forma. Inspirando-nos no estudo de Ostler e Atkins (1991) para o inglês, e utilizando as capacidades do PALAVROSO ao nível da escolha do resultado da anotação, procedemos a uma nova anotação do corpus, obtendo 1609 palavras distintas que tinham mais do que uma interpretação categorial.

Analisando-as uma a uma, separámo-las nas seguintes classes:

Origem comum: as palavras pertencem a uma mesma família de palavras, mas uma não deriva da outra. Exemplos: *futuras*, *ambientais*, *triangular*.

Coincidência: Não há qualquer relação entre as palavras exceptuando o facto de serem homógrafas. Exemplos: *pias*, *saudáveis*, *vogais*.

Derivação (forma original verbal; forma original nominal; forma original adjectival): Exemplos: *procura* (de *procurar*), *vidrar* (de *vidro*), *café* (da cor do *café*), *precisar* (de tornar *preciso*).

Não entrando em conta com homografia com as palavras gramaticais, correspondendo a 29 casos distintos, todos sem qualquer relação entre si, temos 1457 formas distintas que se distribuem entre os seguintes casos:

Tipo de homografia	Origem comum	Deriva de verbo	Deriva de nome	Deriva de adj	Coinci-dência	Outras
n a	70		7	45	68	35
vpp (n ou[10] a)	48	18		5	47	572
vpp v (n ou a)	3				4	10
v (n ou a)	52	64	141	28	152	88
Total	**173**	**82**	**148**	**78**	**271**	**705**

A primeira observação é a de que nem todas as relações entre as palavras homógrafas foi possível englobar nas categorias, à primeira vista exaustivas, que usámos. Várias são as causas para a falta de classificação (a coluna "Outras" na tabela acima):

1. Nos casos em que existe mais do que uma interpretação para a categoria em estudo, como é o caso de *canto* como nome[11], que tanto pode ser derivado do verbo *cantar* como denotar um conceito completamente não relacionado (*canto* de uma divisão), optou-se pela sua não classificação, visto que estaríamos, caso contrário, a

[10] A palavra *ou* significa aqui disjunção inclusiva, ou seja, sob esta rubrica contamos os casos de (vpp n), de (vpp a) e de (vpp n a).

[11] Outro exemplo é a palavra *revista*, que, além de nome, é forma de verbo de *revistar* e *rever*.

debruçarmo-nos sobre a homografia intracategorial, menos interessante de um ponto de vista computacional, e mais complexa de um ponto de vista teórico.

2. Os casos em que uma palavra tem três acepções categoriais e em que as relações par a par são distintas, também não foram contados. Por exemplo, *sereno* é nome, adjectivo e verbo, mas enquanto que o verbo *serenar* deriva do adjectivo, a relação entre o nome e o adjectivo não é trivial. Outros exemplos são as palavras *rota* e *gelado*.

3. Finalmente, todos os casos de palavras homógrafas que, embora possam ser apenas ambíguas entre duas categorias, não se ajustam às etiquetas consideradas, não foram evidentemente etiquetados. Como exemplos apresentamos *certificado*, *conhecido*, *coberta*. Claramente, existe uma relação entre os sentidos das palavras, mas essa relação é mais difusa e difícil de formalizar, sendo que não podemos definir a palavra numa categoria em termos da outra, ainda que tal tenha sido provavelmente a origem. Visto que nos colocamos num ponto de vista estritamente sincrónico, e já que o significado da palavra na acepção que inicialmente era derivada sofreu uma alteração significativa, estas palavras ficam para um estudo futuro mais aprofundado.

Em conclusão, no corpus considerado obtivemos em 20% dos casos distintos nenhuma relação entre as palavras homógrafas, e em apenas 21% dos casos também se observou um caso claro de derivação. Em 12% dos casos foi clara a existência de uma origem comum, e obtivemos um excedente de 47% de casos em que havia quer uma multiplicidade de situações quer uma relação entre sentidos menos fácil de definir com rigor.

Em curso está um estudo que pretende averiguar da existência de relação entre a informação morfológica e o tipo de homografia, visto que, a existir, seria uma ajuda preciosa para a análise computacional da homografia. Não temos no entanto ainda resultados definitivos.

3.3 Frequência de recurso às regras morfológicas

O trabalho referente a este ponto, que se encontra ainda numa fase preliminar, tem como principal objectivo melhorar o desempenho do (analisador morfológico do) PALAVROSO, ao fornecer-lhe informação estatística sobre a frequência das regras usadas na língua. Como consequência, ressaltam algumas propriedades estatísticas sobre o português, que tentaremos apresentar aqui.

Essa informação poderá melhorar o sistema a dois níveis:

1. O da velocidade de processamento: Tal como sugerido em Wirth (1986), se as regras da análise morfológica estiverem organizadas seguindo uma motivação estatística, a velocidade de acesso a essas regras poderá ser substancialmente superior. Por exemplo, podemos colocar as regras em árvores binárias (ou outra estrutura de dados semelhante) e colocar junto da raiz da árvore as regras que são mais usadas.

2. do fornecimento dos resultados por ordem decrescente de probabilidade (e, nos casos em que não exista informação no dicionário sobre uma dada palavra, fornecer a resposta correcta primeiro, também).

Para o primeiro efeito, organização das regras para melhorar a velocidade, aquilo que nos interessa é saber o número de vezes que cada regra é usada, independentemente de originar um resultado correcto ou não. Para tal, a contagem deve ser feita sobre corpora. Assim, reflectirá a utilização real das regras. Sobre este ponto, que é mais relacionado com a arquitectura do programa em si do que com a língua que descreve, apenas diremos que se verifica que cerca de apenas 15% das regras são responsáveis por 90% do processamento (baseado em regras) do PALAVROSO.

Para o segundo efeito, interessa saber, principalmente, o grau de certeza com que as regras são aplicadas. Idealmente, deveríamos comparar a aplicabilidade das regras com a resposta verdadeira, mas para tal precisaríamos de corpora anotados de grandes dimensões, que não existem. Decidimos assim usar a seguinte aproximação (tabelada como "Grau de Certeza 1"): Em vez de testarmos se a regra originou a resposta correcta, verificámos se originou uma palavra que exista no dicionário. Corre-se o risco de a classificação obtida não ser a correcta na situação específica, apesar da palavra existir no dicionário. No entanto, é uma primeira aproximação ao grau de confiança que a regra oferece. Parece um bom princípio darmos maior credibilidade a uma regra que 80% das vezes que é utilizada origina uma palavra existente na língua (e que se encontra portanto no dicionário), do que a uma regra que só o faz 40% das vezes.

Contudo, e embora este grau de confiança seja importante para o uso do sistema como fornecedor de toda a informação sobre uma palavra isoladamente, é claramente deficiente se estivermos interessados em, pelo contrário, obter resultados sobre as palavras no seu uso na língua, em que cada ocorrência de uma palavra gráfica corresponde a uma e apenas uma acepção. Nesse caso, o facto de uma dada regra produzir uma palavra existente na língua não pode ser contabilizado como sucesso da regra. Como referido atrás, sucesso seria a palavra gerada ter sido a usada no texto. Como segunda aproximação (apresentada na tabela como "Grau de Certeza 2"), ponderámos a contagem de sucesso através da fracção de alternativas (constantes no dicionário) que cada ocorrência produziu. Por exemplo, para contar o grau de certeza da(s) regra(s) que analisam o Presente do Indicativo, a palavra *vendo* contaria como 2/3 (ao invés de 1, para a primeira aproximação descrita), visto que o PALAVROSO devolveria três classificações, duas das quais como Presente do Indicativo (dos verbos *vendar* e *vender*) e apenas uma não reforçando o sucesso da regra, como Gerúndio (do verbo *ver*).

Vejamos os primeiros resultados, aplicados aos tempos:

Tempo	Regras Aplicadas	Verbos Existentes	Verbos Ponderados	Grau de certeza 2 (%)	Grau de certeza 1 (%)
Presente do Indicativo	78465	11797	7720.4	9.46	15.03
Presente do Conjuntivo	69638	4062	1645.0	2.36	5.83
Imperfeito do Indicativo	9805	1566	898.5	9.00	15.97
Imperativo	23975	5114	2207.9	8.81	21.33
Perfeito do Indicativo	15654	2560	1806.7	11.04	16.35
Pretérito mais que Perfeito	1876	445	354.2	14.2	23.72
Imperfeito do Conjuntivo	542	335	215.8	30.1	61.81
Futuro do Conjuntivo	7719	4871	1046.3	13.25	63.10
Infinitivo Impessoal	3738	3135	695.6	18.61	83.87
Infinitivo Pessoal	7894	6572	1571.6	19.91	83.25
Futuro do Indicativo	436	415	400.8	82.14	95.18
Condicional	527	440	215.0	40.80	83.49
Gerúndio	666	652	577.8	86.76	97.90
Particípio Passado	5322	3692	1733.8	31.41	69.37

A primeira conclusão que se evidencia é a de que alguns tempos (nomeadamente os futuros, o condicional, os infinitivos e o gerúndio) têm graus de confiança bastante elevados (superiores a 80%), o que era de esperar visto que têm desinências mais longas e características. Ao invés, os presentes, com graus de confiança inferiores a 20%, mostram que se encontram entre as palavras menos marcadas do sistema verbal, não só pela singeleza das desinências como pelo facto de partilharem as marcas morfológicas com o sistema nominal.

Um caso interessante é o do particípio passado. Seria de esperar que este tempo verbal se comportasse em termos quantitativos de forma semelhante, por exemplo, ao gerúndio, mas, no entanto, afasta-se, e, a meu ver, por duas razões: pela emergência das novas formas (irregulares) deste tempo, muito menos marcadas morfologicamente, e pela alta ambiguidade entre adjectivo e particípio passado na nossa língua (veja-se Casteleiro, 1981, como referência essencial, e Marques, 1994a, para os problemas que daí advêm na anotação do nosso corpus).

Os dados apresentados na tabela foram obtidos sobre um corpus de 103300 palavras, que invocaram as regras de conjugação dos verbos 232722 vezes, originando 45656 formas verbais existentes na língua. Considerando que as palavras categorialmente ambíguas, em média, têm uma probabilidade igual por categoria, isso correspondia a apenas 21090 verbos reais. De notar que os graus de certeza foram calculados dividindo o número de

formas correctas pelo número de vezes que a regra foi aplicada, e que portanto não medem de forma alguma a frequência de ocorrência dos tempos na nossa língua[12].

4. Possíveis aplicações ao ensino do português

O tipo de ferramenta (o PALAVROSO) e o conhecimento linguístico que engloba podem ser utilizados para a descoberta da nossa língua com o apoio do computador. Tanto para crianças da escola como para estrangeiros a aprender a nossa língua. A prová-lo, basta observar como as pessoas se entusiasmam com o analisador morfológico quando têm o primeiro contacto com ele: "Deixa ver se ele conhece esta palavra? E esta? E esta...".

Em primeiro lugar, o conhecimento que o PALAVROSO possui torna-o um valiosíssimo auxiliar de consulta. Através da sua simples invocação pode, por exemplo, conhecer-se o género de uma palavra, a ortografia correcta de outra, ou ainda a existência (ou não) de palavras derivadas de uma terceira.

Fácil de utilizar também em programas educativos sobre a morfologia do português, em que o programa, após a exposição da matéria, gera problemas aleatórios e corrige as respostas fornecidas pelo aluno. Finalmente, a capacidade de usar o conhecimento contido no PALAVROSO para geração, em desenvolvimento neste momento, abre as portas ainda a mais jogos educativos em que o sistema pode ser integrado.

De facto, numa perspectiva mais larga, convém salientar que de um ponto de vista de motivação, é essencial fazer compreender ao aluno as faculdades do computador como máquina de palavras e não só de números, o que por outro lado torna a aprendizagem da língua mais fácil e sobretudo muito mais fascinante para o aluno da era da informática.

Por outro lado, o PALAVROSO revela-se, dada a sua capacidade de analisar palavras não existentes na língua, como um auxiliar valioso, sobretudo no ensino de português como língua segunda, para corrigir inteligentemente a excessiva utilização de regras nos casos irregulares, do tipo *fazeu* em vez de *fez*, etc.

Não é demais tornar a sublinhar o contraste entre o conhecido e o desconhecido na nossa ligação à língua materna, e a utilidade que um sistema de descrição abrangente pode ter quer no aprefeiçoamento desse conhecimento quer na compreensão dos seus limites.

5. Conclusão

[12] Além disso, nestes números não estão incluídos os casos das formas verbais francamente irregulares, que, obviamente, não são tratados por regras. Visto que correspondem às formas de verbo mais frequentes da língua, não se podem utilizar os números apresentados para medir qualquer frequência de ocorrência de tempos, a não ser como um limite inferior porque acontece que os tempos menos frequentes e mais marcados são também aqueles em que existe um menor número de excepções.

Esta comunicação teve como objectivo, por um lado, exemplificar a semelhança dos problemas que se deparam à Linguística Computacional com aqueles que os lexicógrafos e os professores da língua enfrentam. Por outro lado, quis descrever alguns estudos quantitativos em curso no INESC e desafiar os ouvintes para também participarem.

O fio condutor que liga as três experiências descritas é a observação das regularidades existentes em corpora (fiéis representantes da nossa língua como é utilizada, ao invés de ser intuída), e a procura de capacidade de previsão para a análise computacional do português. Igualmente interessados estamos na explicação dos fenómenos que observamos, mas se pouco explicámos nesta comunicação, tal deve-se a que primeiro precisamos de observar.

Agradecimento

Quero agradecer aos meus colegas no INESC o trabalho que têm realizado, e em particular a Maria de Jesus Pereira, a Jan Engh e a José Carlos Medeiros a ajuda que me deram na redacção deste artigo.

Referências

AURÉLIO Eletrônico, Editora Nova Fronteira, 1993, baseado em *Novo Dicionário da Língua Portuguesa*, de Aurélio Buarque de Holanda Ferreira, e desenvolvido por Márcio E. Girão Barroso.

BACELAR do Nascimento, Maria Fernanda, Amália Mendes e Diana Santos, "O corpus e a classificação sintáctica dos verbos", *Actas do 1º. Encontro de Processamento de Língua Portuguesa - EPLP'93* (Lisboa, 25-26 Fevereiro 1993), pp.125-9.

BARREIRO, Anabela, Maria de Jesus Pereira e Diana Santos, "Critérios e opções linguísticas no desenvolvimento do Palavroso, um sistema computacional para a descrição morfológica do Português", Relatório INESC RT/54-93, Dezembro de 1993.

CARVALHO, Pedro, Paulo Lopes, Isabel Trancoso e Luís Oliveira, "E-Mail to Voice-Mail Conversion Using a Portuguese Text-to-Speech System", 1994, a publicar.

CASTELEIRO, João Malaca, *Sintaxe transformacional do adjectivo*, INIC, Lisboa, 1981.

COSTA, J. Almeida, e A. Sampaio Melo, *Dicionário da Língua Portuguesa*, 6ª edição corrigida e aumentada, Porto Editora, Porto, 1992.

CUNHA, Celso e Lindley Cintra, *Nova Gramática do Português Contemporâneo*, Lisboa: João Sá da Costa, 4ª ed., 1987.

ENGH, Jan, "Normalisation in Language Industry: Some Normative and Descriptive Aspects of Dictionary Development", *Hermes: Journal of Linguistics*, Vol 10, February 1993.

FONSECA, Ana Cristina de Sena Raposo Paiva, *Comunicação em Linguagem Natural para um Tutor Inteligente*, Tese de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Junho de 1993.

FIGUEIREDO, Cândido de, *Grande Dicionário da Língua Portuguesa*, Bertrand Editora, 23ª ed., 1987.

KOOGAN LAROUSSE, *Dicionário Enciclopédico*, Selecções do Reader's Digest, Lisboa, 1981.

LYONS, John, *Introduction to Theoretical Linguistics*, London, Cambridge University Press, 1968.

LYONS, John, *Semantics*, Volume 2, London, Cambridge University Press, 1977.

MACHADO, José Pedro, *Grande Dicionário da Língua Portuguesa*, Amigos do Livro Editores, 1980.

MARQUES, Rui, "Anotação contextual do corpus INESC 1990", Março de 1994, a publicar em Diana Santos (ed.), *Processamento de Corpora de Texto no INESC*, Vol. 2, Relatório INESC, 1994.

MARQUES, Rui, "Homografia: relações morfológicas e semânticas", Relatório INESC, 1994.

MEDEIROS, José Carlos, "Ferramentas de Processamento de corpora usando o Palavroso", em Diana Santos (ed.), *Processamento de Corpora de Texto no INESC*, Vol. 1, Relatório INESC RT-65/92, 1992.

MEDEIROS, José Carlos, Diana Santos e Rui Marques, "Português Quantitativo", *Actas do 1. Encontro de Processamento de Língua Portuguesa - EPLP'93* (Lisboa, 25-26 Fevereiro 1993), pp.33-8.

MEDEIROS, José Carlos, "Correcção ortográfica e sua implementação", Tese de mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, em preparação.

OSTLER, Nicholas e B.T.S. Atkins, "Predictable Meaning Shifts: Some Linguistic Properties of Lexical Implication Rules", James Pustejowsky and Sabine Bergler (eds.), *Lexical Semantics and Knowledge Representation, Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics*, (Berkeley, 17 June 1991), ACL, pp.76-87.

NOGUEIRA, Rodrigo de Sá, *Dicionário de Verbos Portugueses Conjugados*, 8ª edição, Clássica Editora, Lisboa.

REIS, Regina, "Dicionários de língua corrente: Algumas considerações", *Actas do 1. Encontro de Processamento de Língua Portuguesa - EPLP'93* (Lisboa, 25-26 de Fevereiro de 1993), pp.141-6.

SANTOS, Diana, Carla Fernandes, Rui Marques e José Carlos Medeiros. "Gramática sem dicionário: relatório preliminar", Relatório INESC No. RT/15-92, Maio de 1992.

SANTOS, Diana (ed.), *Processamento de Corpora de texto no INESC*, Vol. 1., Relatório INESC RT-65/92, 1992, Vol 2, a publicar em 1994.

SILVA, Emídio e António Tavares, *Dicionário dos Verbos Portugueses: Conjugação e Regências*, Dicionários Editora, Porto Editora, Porto, 1988.

VILLALVA, Alina, "Compounding in Portuguese", *Rivista di Linguistica* 4, I (1992), pp. 201-219.

WIRTH, Niklaus, *Algorithms & Data Structures*. Prentice-Hall International Editions, 1986.