PORTUGUÊS BRASILEIRO E PORTUGUÊS DE PORTUGAL: ALGUMAS OBSERVAÇÕES

Luzia HelenaWittmann
Tânia Regina Pêgo
Diana Santos
Grupo de Linguagem Natural do INESC

O português de Portugal e o português do Brasil diferem a nível fonológico, lexical, morfológico e sintáctico. O estudo e a quantificação dessas diferenças, contudo, ainda se encontra quase totalmente por fazer, ao contrário do que acontece no caso do inglês, para o qual, citando Johansson, "the literature on the relationship between British (BE) and American English (AE) is vast and varied" (1980:85).

Neste artigo fazemos a apologia desse estudo, propondo uma metodologia baseada tanto em *corpora* quanto em dicionários, e descrevemos alguns primeiros resultados.

I. Da necessidade de uma abordagem científica ao estudo das duas variantes

Embora concordemos com Mateus, quando afirma que "a consciência de falar uma língua — forma privilegiada do comportamento dos homens — está intimamente ligada e até certo ponto dependente das suas convicções e dos seus receios, dos seus desejos e, em última análise, da sua vontade" (Mateus, 1984:303), existem três ordens de razões pelas quais uma afirmação deste tipo não é suficiente para os linguistas portugueses e brasileiros quando se debruçam sobre a sua língua.

1. A necessidade prática

A partir do momento em que é necessário descrever, formalizar, um sistema linguístico para ser abrangido como um todo — como é o caso da engenharia linguística — deparamo-nos com a necessidade de não apenas descrever o que por alguns é chamado um "núcleo comum" ("common core", ou "portugais commun" (Teyssier, 1984)), mas todas e quaisquer manifestações linguísticas do português. Ao reduzir a língua a um núcleo comum, ela deixa de cobrir a linguagem padrão de cada variante.

Do mesmo modo, quando se ensina o português como língua estrangeira, deve-se apresentar as diferenças entre as variantes de forma clara (se não se pretender optar pelo estabelecimento de cursos distintos, de costas viradas entre si, o que é, pelo menos, economicamente desfavorável para a Universidade ou Escola em questão).

Uma opção que não pode ser tomada é a de, por outro lado, aceitar indiscriminadamente sintaxe e/ou léxico das duas variantes. Assim, sustentamos que a seguinte frase, por exemplo,

 # Quando mo deu, ele não tinha se apercebido...

não é português correcto, porque mistura dois fenómenos sintácticos de variantes diferentes: *mo* não é usado em português do Brasil, enquanto que a ordem dos clíticos *não tinha se apercebido* não é aceitável em português europeu.

Ou, no campo do léxico, a frase seguinte

 # Encontrei o banheiro no bonde

não pode ser aceite, porque mistura um termo exclusivo do português de Portugal: *banheiro* (*salva-vidas* em português do Brasil) com outro apenas brasileiro: *bonde* (*eléctrico* no português de Portugal).

Em suma, nem no ensino do português, nem em processamento de linguagem natural, tais aberrações deveriam ser permitidas, o que milita contra a abordagem prática de "apenas ensinar (ou armazenar em computador) aquilo que é comum às duas variantes".

Pela mesma razão, os dicionários de português que queiram cobrir ambas as variantes não se podem limitar a uma soma dos léxicos (e/ou a uma soma das acepções das entradas).

2. O interesse científico

Parece-nos óbvio para a comunidade linguística o interesse de um estudo aprofundado sobre o tema das diferenças entre as duas variantes do português, tão só porque, como Rydén explica claramente, "Linguistic variation is a condition for linguistic change though variation does not necessarily imply change and the study of diachronic change presupposes the study of synchronic variation" (1980:38).

No entanto, listamos aqui algumas das vias que nos parece interessante seguir:

– o estudo de diferentes evoluções e de evoluções paralelas

– o estudo de influências linguísticas distintas

– o estudo da influência da normalização

– o estudo de tendências actuais que resultarão em diferenças futuras prováveis

– e, mesmo, o estudo da influência de uma variante sobre a outra.

3. O interesse político

Se, de facto, como defende entre outros Montes, a diferença entre variante linguística e língua distinta não é linguística, mas sim política: "el problema lengua-dialecto y por tanto de una o varias lenguas no puede resolverse por medios puramente lingüistico-sistémicos (Internos)" (1989:130), como se poderá negar que um estudo conducente a um melhor conhecimento da realidade linguística dos dois povos só pode servir para os aproximar? Pelo contrário, escamotear as diferenças é que não pode ser vantajoso para a cooperação entre as duas comunidades.

Pensamos pois inegável a necessidade de estudar as duas variantes em contraste, restringindo-nos inicialmente à língua escrita[1].

II. Breve descrição do estado da arte

1. O estudo das variantes em geral

A língua inglesa, em particular nas suas variantes britânica e americana, encontra-se numa situação linguística paralela à portuguesa, ainda que de um ponto de vista da descrição das suas diferenças se vá muito mais à frente.

Com efeito, primeiro para o inglês americano, e depois para o inglês britânico, foram constituídos *corpora* informatizados com uma constituição paralela (os chamados Brown corpus (Francis e Kucera, 1979) e Lancaster-Oslo/Bergen corpus, Johansson et al. (1978)). Dada essa situação, virtualmente qualquer estudo feito sobre um destes *corpora* pode ser replicado para a outra variante. No entanto, os compiladores do LOB *corpus* fizeram também investigação específica sobre as diferenças: veja-se Johansson (1979, 1980), Coates and Leech (1980), Krogvig (1979), etc. Hofland e Johansson (1982) apresentam uma comparação das frequências das palavras (de frequência maior do que 10) nos dois *corpora*, afectados de um coeficiente de significância estatística. Essa lista (que é suplementada por todas as concordâncias no *corpus*, fornecidas em microfilme) permite imediatamente alguns estudos globais, nomeadamente a verificação dos contrastes ortográficos e morfológicos, o confronto entre o uso de diferentes palavras gramaticais (como por exemplo as terminadas em *-ward* e *-wards*), o contraste entre o uso de modais e auxiliares, assim como alguns contrastes lexicais e institucionais ou culturais.

Em relação ao francês, e por nestas actas a comunicação de Blanche-Benveniste se lhes referir de forma inigualável, não trataremos aqui os diversos estudos existentes.

Em relação ao castelhano ibérico e mexicano, foi feito um contraste entre o vocabulário espanhol peninsular, apresentado em (Juilland & Chang-Rodriguez, 1964), e o mexicano, coligido por (Lara, 1992). Contudo, o tamanho dos *corpora* envolvidos era muito diferente, como aponta Biderman (1994). O trabalho desta investigadora, no entanto, contrastando o vocabulário fundamental de Portugal com o que calculou para a variedade brasileira, não desmerecendo o inegável mérito, sofre do mesmo problema, ou seja, dos dois vocabulários se apoiarem sobre *corpora* não comparáveis, a saber: um *corpus* exclusivamente oral espontâneo (suplementado depois com inquéritos de disponibilidade), PF (1984), por oposição a um corpus de língua escrita com uma parte de oral formal (oratória). Também o método de obtenção dos vocabulários foi distinto (baseado exclusivamente num limiar de frequência o brasileiro, e suplementado por outros itens o português, como mencionado acima).

Ora, sabe-se de estudos sobre registos/níveis de língua que as línguas têm comportamentos muito variados em textos de tipo diferente

(e, que, por consequência, "textos" orais espontâneos diferem consideravelmente de textos escritos elaborados[2].) Por isso, não obstante as considerações acertadas feitas em Biderman (1994) sobre as diferenças vocabulares entre as duas variantes, pensamos que esse estudo não é suficiente para estabelecer de forma rigorosa os contrastes que aponta.

Por outro lado, estudos sobre línguas próximas não são muito frequentes numa abordagem tipológica, embora Dahl, tipologista de renome, afirme que "a comparison of several closely related languages may well throw light on the ways in which almost identical grammatical systems may differ in details, and suggest how diachronic processes may influence the grammar" (1985:38).

Também no campo da tradução automática, e embora atraente do ponto de vista prático, tem havido poucos sistemas que tratem línguas próximas ("closely-related languages"), provavelmente por as razões que presidem à escolha das línguas envolvidas terem a ver com factores económicos e não linguísticos. No entanto, os investigadores envolvidos são unânimes em declarar que a complexidade do processo é qualitativamente a mesma, envolvendo apenas um esforço muito menor de um ponto de vista quantitativo (cf. Bémová et al. (1988), Santos e Engh (1992)).

2. O estudo das variantes do português

Embora a especificidade do português brasileiro em relação ao português europeu seja um assunto ciclicamente retomado no Brasil, poucos são os estudos que foquem especialmente o contraste.

No entanto, podemos considerar como indirectamente relacionados com os estudos contrastivos, investigações incidentes sobre a especificidade do português brasileiro. Os artigos reunidos em Roberts e Kato (1993), nomeadamente, contêm informações muito úteis para a definição de alguns tipos de contrastes. Por sua vez, um estudo sincrónico das diferenças entre as duas variantes poderá ser útil na compreensão do processo diacrónico.

Dicionários contrastivos como o de Mauro Vilar (1989), trabalho rico de informações e elaborado com muito rigor, e até mesmo o dicionário humorístico de Mário Prata (1993), constituem os únicos registos voltados especificamente à observação dos contrastes entre PE e PB, além do trabalho de Tereza Biderman, já citado acima.

Entre as gramáticas, é de ressaltar o esforço despendido pelo trabalho conjunto de Lindley Cintra e Celso Cunha (1987), embora considerem, sem discriminação, fenómenos de linguagens particularmente marcadas (como o registo oral popular) com outros que se manifestam de forma generalizada e constante. Algumas gramáticas dirigidas a estudantes e utentes estrangeiros, demonstram particular atenção às diferenças entre estas duas variantes do Português. É o caso de Cuesta e Luz (1971), Paul Teyssier (1989) e Abreu e Murteira (1994). Esta última inclui um livro de exercícios com dupla versão, de PE e PB.

III. Princípios metodológicos

1. Princípios gerais

Além de um levantamento sobre as regras básicas de competência gramatical de um falante, feita através de uma análise detalhada das gramáticas, é essencial recorrer a *corpora* representativos das duas variantes, que contêm a língua em uso, e não apenas em embrião.

Isto porque é essencial ter em mente que se, como notou Jakobson, "Languages differ essentially in what they *must* convey and not in what they *may* convey" (1959:236), por outro lado, na comparação de duas línguas não interessa tanto o que se *pode* dizer, mas o que se *diz*. E, como também notam Hofland e Johansson, referindo-se às diferenças de vocabulário entre o inglês americano e o britânico, "It is reasonable to assume that many differences will be relative rather than absolute" (1982:33).

O uso de *corpora* permitirá assim detectar não só diferenças absolutas, mas sobretudo diferenças relativas, preferências de uma comunidade em relação a outra, que são tão importantes, na nossa opinião, como questões de simples (a)gramaticalidade.

Por outro lado, não podemos esquecer os dicionários que, como repositórios de língua, são sempre o outro lado de uma descrição abrangente de um idioma. Seria por isso impensável que um trabalho sério não fosse portanto apoiado também nas descrições lexicográficas das duas variantes.

Ou seja, como praticamente todos os investigadores em Processamento de Linguagem Natural têm vindo a reconhecer, é tão parcial uma descrição baseada só em *corpora* como uma baseada apenas em dicionários, visto que a informação contida em ambos é mais complementar do que concordante[3] (veja-se, a título de exemplo, Klavans e Tzoukermann, 1995).

Cada dicionário é em si uma forma (imperfeita, é certo) de descrever uma língua. Não é pois evidente que ao juntar mais do que um dicionário se obtenha uma descrição melhor (veja-se sobre esse assunto Atkins & Levin, 1994).

Um tipo de obra lexicográfica que nos interessa são os chamados dicionários contrastivos, em que as entradas são apenas aquelas em que há contraste entre duas variantes/línguas próximas, admitindo-se, por omissão, que palavras que lá não se encontrem sejam comuns.

2. Definindo os *corpora*

A definição do tipo de textos que devem constituir os *corpora* está directamente relacionada com os objectivos da investigação. Para o processamento de linguagem natural, pensamos que um primeiro estudo abrangente deve incidir sobre a linguagem corrente escrita de cada variante. Temos os olhos postos em ferramentas como conversores automáticos, sistemas de tradução automática de outras línguas para as duas variantes, adaptação de correctores ortográficos e sintácticos. Pode-se, para tal, recorrer a textos jornalísticos, obras de divulgação científica para leigos, obras literárias contemporâneas nada ou pouco marcadas por regionalismos, etc.

É necessário desenhar e constituir os *corpora* comparativos (ou comparados) e paralelos, posto que ainda não existem para as variantes do Português. Consideramos a utilidade de três tipos de paralelismos:

2.1. *Corpus* paralelo de adaptações

Constitui-se de textos originais de PB, adaptados para PE e vice-versa. Implica, portanto, a recolha de textos originais de uma das variantes e a correspondente adaptação para a outra variante, produzidos e publicados por editoras, jornais etc.

Este tipo de *corpora* permite a captação de contrastes tanto lexicais quanto morfossintácticos, através da observação das substituições

sistemáticas. Permite ainda uma grande precisão no levantamento quantitativo dos contrastes. Esta precisão leva a resultados seguros quanto à frequência de uso das palavras contrastivas. Oferece, no entanto, pouca autenticidade. Uma adaptação (e mesmo uma tradução) sofre normalmente alguma influência do texto original, sobretudo quando se trata de duas variantes tão próximas, onde muitas das diferenças estão fundamentadas apenas na frequência de uso e não na possibilidade/impossibilidade de uso de palavras e construções.

2.2. *Corpus* paralelo de traduções

Compõe-se de textos traduzidos de originais noutras línguas para cada uma das variantes de forma independente.

Embora este tipo de textos, em princípio, ofereça maior autenticidade do que os textos adaptados a partir de uma outra variante, verifica-se alguma influência da língua original, em maior ou menor grau (veja-se por exemplo Santos, 1995). Mas, se, por um lado, temos maior autenticidade, por outro, o paralelismo aqui já é menos exacto, pois cada tradução está sujeita às preferências estilísticas e opções do tradutor.

2.3. *Corpus* comparado de originais

Pode ser constituído por pares de textos originais escritos e publicados em cada variante, de tamanho aproximado, pertencentes à mesma área temática e dirigidos ao mesmo tipo de receptores. De acordo com as autoridades científicas na área, deverá ter uma dimensão mínima de 10 milhões de palavras (5 milhões para cada variante). Se nos *corpora* paralelos de adaptações e de traduções a captação de contrastes tanto de nível sintáctico quanto morfológico e lexical é mais fácil, esses contrastes, uma vez identificados, devem ser submetidos a confirmação e estudo mais aprofundado nos *corpora* comparados de originais.

Embora a utilização dos *corpora* sobretudo os *corpora* paralelos seja essencial na identificação das diferenças, a consulta dos dicionários é indispensável. A informação contida nos dicionários contrastivos constitui já um importante acervo, a ser reaproveitado, desde que devidamente confirmada em *corpora* e classificada segundo a tipologia que agora descrevemos.

3. Para uma tipologia de contrastes

Para garantir a utilidade e eficácia dos dados obtidos no processamento do português, propomos a classificação dos contrastes

(1) quanto ao nível gramatical: sintácticos, morfológicos, e lexicais

(2) quanto à frequência de uso: absolutos e relativos

3.1. Quanto ao nível gramatical

A delimitação dos fenómenos considerados lexicais dos morfológicos, ou a destrinça entre os morfológicos e os sintácticos não é trivial, nem consensual. A prová-lo está o adjectivo "morfo-sintáctico", por um lado, e a progressiva descrição de mais e mais características sintácticas como associadas a itens lexicais.

Com esta separação dos contrastes segundo o nível gramatical não pretendemos portanto resolver o problema da delimitação para o português, mas tão só dividir os vários casos de uma forma pertinente em relação ao objectivo do nosso estudo.

3.1.1. Contrastes de nível sintáctico

Assim, consideramos como contrastes a nível sintáctico aqueles que correspondem a diferentes organizações do texto (ordem na oração, existência ou ausência de palavras gramaticais, diferença ao nível da palavra gramatical empregue, diferença no uso da flexão). Exemplificamos com a questão da colocação dos clíticos, o caso das subcategorizações verbais e nominais, o uso do gerúndio, as contracções, a omissão do artigo, o uso do imperfeito em vez do condicional. Cada um dos casos tem de ser estudado independentemente, combinando as informações disponíveis nas gramaticas normativas, nos estudos linguísticos específicos e nos dados extraídos dos três tipos de *corpora* paralelos descritos acima.

Realizámos um estudo prévio, baseado na recolha de todos os casos contrastivos mencionados em gramáticas normativas e na comparação de dois pequenos *corpora* (de adaptações e de traduções), incluindo estudos quantitativos (ver Wittmann e Pereira, 1994). Esse trabalho permitiu-nos a obtenção de alguns resultados preliminares e, sobretudo, testar a metodologia que agora propomos.

3.1.2. Contrastes de nível morfológico

Como contrastes a nível morfológico consideramos apenas aqueles contrastes que dizem respeito à flexão, ou à classificação morfológica da própria palavra. As características da flexão podem constituir contrastes em si, por exemplo, variação em género ou número, ou podem corresponder a diferentes formas particulares de flexão, tal como diferenças na forma do particípio passado (*aceite* e *aceito*) ou diferente conjugação do pretérito perfeito do indicativo dos verbos em *-ar*.

Além disso, considerámos como uma questão morfológica a classificação morfológica de uma dada palavra (em inglês, "part of speech"), o que engloba, por exemplo, a oscilação entre a classificação de nome ou adjectivo, ou adjectivo e particípio passado. Este assunto será retomado no item V.

Contrastes referentes a diferente formação por derivação, por outro lado, ainda que se originem num processo morfológico, dão origem a itens lexicais distintos (por exemplo, *doutoramento* e *doutorado*), o que nos levou a considerá-los como contrastes lexicais.

3.1.3. Contrastes de nível lexical

Os contrastes de nível lexical, finalmente, são aqueles associados às palavras sozinhas, e que vão desde a sua mera ortografia até ao seu sentido e conotações. Informação sintáctica associada a uma palavra deverá também ser considerada nesta rubrica. Por exemplo, a reflexividade dos verbos pode constituir um contraste. Há verbos que, sendo reflexivos numa das variantes, não o são na outra, como por exemplo *reunir* e *reunir-se*.

3.2. Quanto à frequência de uso

Estando a investigação circunscrita à linguagem de uso corrente, ou seja, uma linguagem não marcada, a frequência de uso impõe-se como um factor determinante na descrição contrastiva. Há que ter presente o facto de a linguagem corrente caracterizar-se pela familiaridade de uma larga fatia da população com as escolhas linguísticas actualizadas num texto, em oposição à linguagem artística, por exemplo, que tem como um dos seus méritos causar estranhamento[4], ou ao jargão técnico-científico, cuja compreensão está limitada a grupos

restritos. Assim, consideramos como contrastes tanto formas ou construções não comuns às duas variantes, quanto formas e construções comuns, mas que apresentem grande disparidade quanto à frequência de uso.

Esta tipologia foi elaborada para/durante a construção de um léxico contrastivo (ver item IV). Estamos convencidos de que as distinções básicas propostas com base na frequência de uso para os contrastes lexicais também são aplicáveis aos níveis sintáctico e morfológico. No entanto, como foram estabelecidas a partir da análise dos contrastes lexicais, apresentamo-las apenas para o nível lexical, deixando prudentemente para um novo estudo as complexidades tipológicas dos níveis sintáctico e morfológico.

3.2.1. Contrastes absolutos

Entendemos por absolutos os contrastes constituídos por palavras usadas em exclusivo numa das variantes, ou seja, palavras cujo correspondente ou não existe ou é diferente na outra variante.

Dentre os contrastes absolutos distinguimos ainda (a) palavras diferentes para o mesmo referente, (b) palavras sem equivalência, ou seja, cujo referente (objecto ou conceito) não existe na cultura do país da outra variante e (c) contrastes institucionais.

3.2.1.1. Palavras diferentes para o mesmo referente. Este campo inclui as seguintes nuances:

i) pares contrastivos do tipo *autocarro* (PE) e *ônibus* (PB), ou seja, envolvendo palavras exclusivas de uma das variantes *(autocarro)*, cujo correspondente na outra variante é uma palavra distinta e também exclusiva *(ônibus)*;

ii) pares contrastivos nos quais pelo menos uma das palavras é usada em ambas as variantes, mas com significados diferentes, como por exemplo *banheiro* (PE), cujo correspondente em PB é *salva-vidas*, enquanto o correspondente para *banheiro* (PB) em PE é *casa-de-banho*;

iii) pares contrastivos envolvendo palavras com pelo menos um significado diferente nas duas variantes, mas que também têm pelo menos um significado comum: *alcatrão* (PE) e *asfalto* (PB) ou *cartão* (PE) e *papelão* (PB);

iv) palavras compostas em que apenas uma das componentes é contrastiva: ex: *gira-discos* (PE) e *toca-disco* (PB).

3.2.1.2. Palavras sem equivalência. São palavras que constituem contrastes por, além de não serem usadas na outra variante, não haver nenhuma palavra equivalente. Em geral, são nomes vulgares (não científicos) de certas plantas, frutas ou animais não pertencentes à linguagem corrente da outra variante, embora não exista nenhum equivalente, como por exemplo *azinheira* (PE) ou *sapoti* (PB).

3.2.1.3. Contrastes institucionais. Cobrem palavras e expressões relacionadas com diferenças a nível organizacional entre Portugal e Brasil, como é o caso, por exemplo, do sistema educacional (*liceu, primeiro grau*) das regiões administrativas (*distrito, estado*), de instituições oficiais, etc. Este tipo de contrastes compõe um conjunto à parte por representar uma realidade equivalente mas não igual no âmbito cultural dos dois países.

3.2.2. Contrastes preferenciais

Palavras que, embora existam ou estejam atestadas em dicionários de ambas as variantes com o mesmo significado, têm uma frequência de uso diferentes. Em outros termos, a palavra classificada como *contraste preferencial*, existe nas duas variantes com o mesmo significado, mas torna-se contrastiva, do ponto de vista da linguagem corrente, por ser usada com maior frequência, ou seja, por ser preferencial.

Ambas as palavras que compõem o par contrastivo podem ser preferenciais como por exemplo o par *chávena* (PE) e *xícara* (PB). Quando apenas uma das palavras do par é preferencial para a sua variante, subentende-se que o seu equivalente não é usado nessa variante, como no par *talho* (PE) e *açougue* (PB). Em PE *talho* é preferencial a *açougue*, enquanto em PB a palavra *talho* não é usada com esse significado.

3.2.3 Contrastes opcionais

Palavras de uso exclusivo da sua própria variante, mas cuja palavra equivalente na outra variante também é usada e preferida.

Mesmo que as *palavras contrastivas opcionais* sejam menos usadas do que seus sinónimos comuns às duas variantes, podem ainda pertencer à linguagem corrente. Veja-se por exemplo a palavra *sebo* (PB), cujo par contrastivo é *alfarrabista* (PE). Ora, *alfarrabista* também se usa em PB e é preferencial, não constituindo, em si, contraste. *Sebo,* portanto, será marcada como opcional em relação a *alfarrabista.*

IV. Dados quantitativos globais

Aqui apresentamos alguns dados quantitativos de que dispomos e que permitem uma primeira medição da profundidade da diferença entre as duas variantes, resultando dos seguintes estudos preliminares:

(1) a comparação dos dois léxicos, com base em *corpora* paralelos (Wittmann & Pereira, 1994);

(2) a comparação dos dois léxicos com base em dicionários computacionais de linguagem corrente (Barreiro et al., 1995);

(3) a comparação dos dois léxicos com base em dicionários bilingues inglês-português de linguagem técnica (Barreiro et al., 1995).

1. Comparação dos dois léxicos com base em *corpora*

Dois tipos de *corpora* foram examinados nesta experiência. Por um lado, um corpus paralelo de cerca de 5.000 palavras formado por textos originais em PB e adaptados para PE, dos quais foram extraídos dois léxicos de cerca de 2.270 palavras plenas cada. Por outro lado, um corpus paralelo de textos originais em inglês, traduzidos para PE e para PB cerca de 5.800 palavras cada, dando origem a dois léxicos com cerca de 3.100 palavras cada.

Os resultados, extraídos de Wittmann & Pereira (1994), encontram-se na Tabela 1.

	nº de palavras plenas	nº de contrastes	nº de palavras plenas distintas	nº de contrastes distintos
corpus convertido de PE→PB	2.270	144 6,3%	1.007	70 7%
corpus traduzido do inglês para PE e PB	PB: 3.120 PE: 3.107	209 6,7%	PB: 753 PE: 942	79 PB: 10,5% PE: 8,4%

Tabela 1

2. Comparação dos dois léxicos com base em dicionários computacionais

Esta investigação, descrita detalhadamente em Barreiro et al. (1995), foi efectuada através da análise minuciosa e exaustiva de partes de um léxico computacional existente de PE[5] e de um léxico computacional existente de PB[6], por investigadores brasileiros e portugueses, respectivamente. Recorreu-se à ajuda de dicionários (Aurélio, Luft, Figueiredo e Porto Editora), assim como ao conhecimento empírico dos investigadores. Entradas pertencentes ao registo popular ou regional foram excluídas, assim como entradas desconhecidas e não encontradas nos dicionários consultados. As palavras contrastivas, ao serem localizadas, eram analisadas, propunha-se-lhe um equivalente na outra variante e atribuía-se uma classificação segundo a tipologia anteriormente definida.

Um resumo dos resultados apresentados em Barreiro et al. (1995) encontra-se na Tabela 2.

	nº de palavras analisadas	nº.de contrastes	número percentual
léxico de PE (nomes e adj.)	3.550	417 absolutos: 214 (94 c/equiv. em PB 120 s/equiv. em PB) preferenc: 67 (44 no PB; 6 no PE 17 em ambos) opcionais: 73 (36 no PB; 37 no PE) ortográfic: 114	11,74%
léxico de PB (todas as categ.)	6.393	639 absolutos: 400 (76 c/equiv. em PB 323 s/equiv. em PB) preferenc: 18 (10 no PE; 5 no PB 3 em ambos) opcionais: 115 (no PB) ortográfic: 114	9,99%

Tabela 2

Note-se que é possível que alguns dos contrastes assim coligidos e, sobretudo, as classificações quanto à frequência de uso (*absolutos*,

preferenciais e *opcionais*), não resistam à confrontação com os dados de frequência a serem extraídos de largos *corpora*. Estamos convencidos, no entanto, de que a comparação de léxicos, com o auxílio de dicionários descritivos, é uma tarefa complementar à comparação de *corpora* paralelos no estabelecimento e classificação dos contrastes lexicais entre variantes de uma mesma língua.

2.1. Profundidade das diferenças ortográficas antes do novo Acordo Ortográfico

Quanto aos contrastes ortográficos, convém lembrar que, mesmo depois de o Acordo Ortográfico da Língua Portuguesa (Decreto nº 43/91) ser posto em prática, permanecerão algumas diferenças, pois em todos os casos de admissão de dupla grafia, será o uso a definir a permanência ou não do contraste. Analise-se, por exemplo, os casos de *comitê* e *comité*, *fato* e *facto* ou mesmo *aspecto* e *aspeto*, que são diversamente pronunciadas em cada variante.

Do léxico de PE, contendo 48.019 lemas, constatámos que 2,35% das palavras diferem do Português Brasileiro a nível ortográfico. Com esse dado poder-se-á, por exemplo, inferir a extensão interventiva do novo Acordo Ortográfico.

Grande parte dos contrastes ortográficos foram detectados automaticamente no léxico de PE, a partir de sequências de consoantes como *cc, ct, pc, pç, pt, mpt, bd, bt, mn, mm* e *nn*, onde o fenómeno ocorre com maior frequência, permitindo a captação de palavras como *facto, adopção* etc. O mesmo foi feito a partir do exame de todas as palavras terminadas em *é* (*bebé*) e todas as que contêm as sequências *ém, én, óm* e *ón* (ex. *académico, biénio, atómico, bónus*). Todas as palavras assim extraídas foram conferidas manualmente, para eliminar palavras não contrastivas como *pacto*, que tem a mesma grafia nas duas variantes.

3. Comparação dos dois léxicos com base em dicionários bilingues de linguagem técnica

A partir de um corpus em inglês contendo textos técnicos em seis áreas diferentes, descrito em Barreiro et al. (1995), uma lista de 2.435 termos técnicos ingleses foi escolhida, incluindo apenas nomes, adjectivos e verbos.

Através de consulta aos dicionários, as áreas a que os termos pertenciam, assim como uma primeira tradução, foram obtidas. Fixada a área, a tradução para a outra variante era procurada. Somente em 1.376 casos se obteve tradução para a outra variante. Nos casos em que foi encontrada tradução para as duas variantes, o número de contrastes é o apresentado na Tabela 3.

nº de termos em inglês	nº de contrastes	percentagem
1.376	469	32,77%

Tabela 3

Convém, no entanto, notar que, embora os termos ingleses fossem expressos por uma palavra só, em muitos casos os termos correspondentes em português envolviam mais do que uma palavra, e, portanto, alguns destes contrastes englobam questões sintácticas também.

V. Contrastes ao nível da classificação morfológica

No decorrer dos trabalhos de análise dos léxicos para a detecção de contrastes lexicais, referido acima, apercebemo-nos da existência de contrastes, ou melhor dizendo, de diferenças a nível das características morfológicas. Ora como uma informação fundamental, quer em dicionários tradicionais quer em léxicos computacionais, é a categoria a que a palavra pertence, tal forçou-nos a iniciar um estudo mais específico sobre a questão da classificação morfológica.

1. Selecção do léxico

Partimos do léxico em PE do nosso analisador morfológico, Palavroso, que associa a cada lema a sua classificação gramatical, e extraímos todos os substantivos, adjectivos (incluindo os compostos) e verbos, num total de 49.134 palavras.

Uma vez que utilizámos material retirado de uma só variante (PE), era necessário, antes de mais, eliminar os contrastes lexicais e nos atermos às palavras que pertencessem às duas variantes. Seleccionado esse núcleo de palavras comuns (39.693), é que se iniciou o estudo morfológico contrastivo.

2. Procedimentos

Pretendemos desenvolver este estudo em duas fases, as quais chamámos fase de detecção, cujas dificuldades comentamos aqui, e fase de confirmação, ainda por realizar.

O objectivo da fase de detecção era detectar as diferenças de classificação, quer dentro de uma mesma variante, quer entre as duas variantes em estudo, com base em dicionários portugueses: Cândido de Figueiredo (1986) e Porto Editora (Costa & Melo, 1994), e brasileiros: Aurélio (Ferreira, 1993) e Luft (s/d), recorrendo a gramáticas quando necessário.

Na fase de confirmação pretende-se confirmar e/ou rectificar os contrastes detectados durante a primeira fase, através de pesquisa em *corpora*.

3. Dificuldades

Não vamos aqui enumerar todas as dificuldades inerentes a um trabalho próprio de confronto morfológico entre duas variantes de uma língua, mas convém ressaltar aquelas que consideramos mais problemáticas.

3.1. Dicionários

É notória a disparidade na classificação morfológica, e por vezes até a omissão do vocábulo, apresentada em dicionários de uma mesma variante. Quando tal acontece, é necessário conjugar essas mesmas classificações, e por vezes adiar uma conclusão, por ser imprescindível a sua confirmação em *corpora*, a fim de se apurar a mais abrangente e correcta.

A título de constatação, apresentamos alguns casos de divergência na classificação morfológica encontrados quer em PE quer em PB:

entrada	*C.de Figueiredo*	*Porto Editora*
beta	nome/masculino e nome/feminino	nome/feminino
jurássico	adjectivo/masculino	nome/masculino

entrada	*Aurélio*	*Luft*
almejado	verbo no particípio	verbo no particípio adjectivo
paciente	nome e adjectivo/inv. nome/masculino	nome e adj./inv. nome e adj./masc.

Por vezes, esbarramos numa divergência ainda mais difícil de ser tratada, a da metodologia empregada por um dado autor. Nesse caso, não podemos recorrer a outros dicionários nem mesmo a *corpora*, por se tratar de uma opção do próprio autor. Citamos, como exemplo, dois verbetes extraídos do Aurélio, com a mesma classificação morfológica (adjectivo), mas apresentados com características diferentes.

paciente – adjectivo/invariável
significante – adjectivo/masculino

3.2. Adjectivos e/ou particípios

Convém ainda mencionar uma questão que há algum tempo vem suscitando o interesse de linguistas de renome, seja no Brasil ou em Portugal. Trata-se da adjectivação dos particípios, ou, como alguns preferem designar, do adjectivo verbal, ou seja, do particípio com valor de adjectivo, como é o caso de *abarcado*, *abolido*, *contemplado* e outros. Muito já foi/tem sido dito e escrito sobre esse tema, mas não temos conhecimento de um estudo que pudéssemos chamar conclusivo, sobretudo porque, também aí e uma vez mais, não se pode dizer que haja concordância de opiniões.

Enquanto não podemos desfrutar de uma definição normalizada, continuamos a esbarrar na já citada divergência entre os dicionários.

4. Resultados preliminares

Considerando que as diferenças morfológicas estão intimamente relacionadas com o(s) significado(s) próprio(s) da palavra e com aquele(s) que pode adquirir em contexto, apresentamos alguns exemplos dos tipos de diferenças encontrados e os resultados quantitativos preliminares.

4.1. Quando a carga semântica do vocábulo não sofre alteração significativa, apenas verificámos diferenças a nível do género.

entrada	*PE*	*PB*
jurisprudente	nome/masculino	nome/invariável
invariante	nome/invariável	nome/masculino

4.2. Quando o significado do vocábulo é mais alargado numa das variantes, observamos alterações na classe e no género.

4.2.1. variação na classe

entrada	*PE*	*PB*
luminoso	adjectivo/masculino	adjectivo/masculino nome/masculino
sinistrado	adjectivo/masculino	adjectivo/masculino nome/masculino

4.2.2. variação na classe e no género

entrada	*PE*	*PB*
juvenil	adjectivo/invariável	adjectivo/invariável nome/masculino
servente	nome e adjectivo/inv.	nome e adjectivo/inv. nome/masculino

4.2.3. Resultados quantitativos globais

Entre as 39.693 palavras de PE observadas e seguindo apenas as informações disponíveis nos dicionários, captámos 791 (1,13%) palavras com classificações divergentes: 563 (1,41%) adicionais para PE e 731 (1,84%) para PB.

Conclusão

Um dos principais objectivos desta comunicação foi evidenciar a necessidade de estabelecer as diferenças entre as variantes do português. Embora o interesse científico o justificasse por si só, uma nova necessidade emergiu da convergência de vários factores históricos interligados, entre os quais se destacam a criação da Comunidade dos Países de Língua Oficial Portuguesa e o desenvolvimento dos novos meios de comunicação/transmissão de informação computacionais.

Mais do que nunca é necessário defender a língua portuguesa como um todo, projectando-a para o futuro como uma língua de informação científica e profissional, de modo a evitar a sua marginalização e restrição a um âmbito literário e familiar.

Uma das medidas mais importantes para a defesa da nossa língua face ao avanço do inglês é o desenvolvimento de programas de processamento de texto. Uma vez sabido que na linguagem corrente as variantes de PE e PB diferem em cerca de dez por cento, é possível juntar esforços para a criação de novos programas, contendo o "núcleo comum" mais as especificidades de cada variante. Deste modo a mesma ferramenta poderá ser usada tanto para a variante brasileira quanto para a europeia e mais tarde também para as demais variantes sem esbarrar na forma peculiar com que cada povo actualiza o português.

Por essa razão desenvolvemos uma metodologia adaptada às necessidades da engenharia linguística. Introduzimos o conceito de contrastes absolutos e relativos e considerámos a necessidade do uso de largas quantidades de *corpora* paralelos e comparados, para atingir a linguagem em uso e não uma linguagem ideal do ponto de vista do especialista.

Notas

[1] Esta restrição é motivada por numerosos factores, de índole científica uns, de ordem prática outros: maior homogeneidade e estandartização, mais facilidade em obter materiais de estudo, e maior número de aplicações (em termos de sistemas de processamento de linguagem natural) para os resultados obtidos.

[2] De facto, na sua obra sobre as diferenças entre o oral e a língua escrita, Biber (1988) sustenta que o contraste é maior em relação ao tipo de texto ("text type"), do que ao meio (oral/escrito), e que, por consequência, há mais parecenças entre o género epistolar familiar e o oral espontâneo, por um lado, e o género oral formal e escrito, por outro, do que entre textos elaborados vs. textos espontâneos.

[3] Pese embora a nova moda de compilar dicionários através do recurso a corpora informatizados, que substituem as abonações em obras literárias de autores consagrados, é preciso insistir que um dicionário é muito mais do que uma concordância gigantesca – é um ente distinto de um corpus, devido à introdução do critério e da análise do lexicógrafo.

[4] Termo empregado aqui conforme a sua definição na retórica (cf. Lausberg, 1967).

[5] O léxico do Palavroso, desenvolvido pelo Grupo de Linguagem Natural do INESC. Veja-se Medeiros (1995) para uma descrição completa do sistema, e Barreiro et al. (1993) e Santos (1994) para a descrição das suas bases linguísticas e discussão.

[6] Uma lista de cerca de 67.000 palavras, pertencente à SMD Informática, e que muito agradecemos ter sido posta à nossa disposição para efeitos desta investigação.

Referências

— Acordo Ortográfico da Língua Portuguesa. *Decreto do Presidente da República nº 43/91; Resolução da Assembleia da República nº 26/91, Imprensa Nacional-Casa da moeda, Lisboa, 1991.*

ABREU, HELENA & BENAMOR, RITA, *adapt. do livro de exercícios para Português Brasileiro por Wittmann, Luzia Helena.* Gramatica del Portughese Moderno, *Zanichelli Editore, Bologna, 1994.*

ATKINS, BERYL T. & LEVIN, BETH. *"Admitting Impediments", in Uri Zerni (ed.),* Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, *Lawrence Erlbaum Associates, Publishers, New Jersey Hove and London*

BARREIRO, ANABELA; PEREIRA, MARIA DE JESUS; SANTOS, DIANA. Critérios e Opções Linguísticas no Desenvolvimento do Palavroso, um Sistema Computacional de Descrição Morfológica do Português, *Relatório INESC RT/54-93, 1993.*

BARREIRO, ANABELA; WITTMANN, LUZIA HELENA; PEREIRA, MARIA DE JESUS. *"Lexical differences between European and Brazilian Portuguese", in* The INESC Journal of Research and Development, *no prelo.*

BEMOVÁ, ALLA; OLIVA, KAREL; PANEVOVÁ, JARMILLA. *"Some Problems of Machine Translation Between Closely Related Languages",* Proceedings of COLING'88 *(Budapest, 22-27 August 1988), 1988, pp. 46-48.*

BIBER, DOUGLAS. *"Textual Comparison of British and American Writing" in* American Speech *2:99-119, 1987.*

BIBER, DOUGLAS. Variation Across Speech and Writing, *Cambridge University Press, 1988.*

BIDERMAN, MARIA TEREZA C. Vocabulário Fundamental: Cultura e Sociedade, *UNESP, Araraquara, SP, exemplar policopiado.*

COATES, JENNIFER & LEECH, GEOFFREY N. *"The Meanings of the Modals in British and American English",* York Papers in Linguistics 8, *1980, pp. 23-24.*

COSTA, J. A. & MELO, A. S. Dicionário da Língua Portuguesa, *Porto Editora, 7ª ed revista e ampliada, 1994.*

CUESTA, PILAR VÁZQUES & LUZ, MARIA ALBERTINA MENDES DA. Gramática da Língua Portuguesa. *Edições 70, Lisboa, 1971.*

CUNHA, CELSO & CINTRA, LINDLEY. Nova Gramática do Português Contemporâneo, *Edições João Sá da Costa, Lisboa, 1987.*

DAHL, ÖSTEN. Tense and Aspect Systems, *Basil Blackwell, 1985.*

FERREIRA, A. B. H. Dicionário Aurélio Electrônico, *Editora Nova Fronteira, Rio de Janeiro, 1993.*

FIGUEIREDO, CÂNDIDO DE. Grande Dicionário da Língua Portuguesa, *Bertrand Editora, Venda Nova, 23ª ed., 1986.*

FRANCIS, W.N. & H. KUCERA. Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers, *3rd edition, 1979 (first edition, 1964).*

HOFLAND, KNUT & JOHANSSON, STIG. Word Frequencies in British and American English, *Bergen and London, 1982*

JAKOBSON, ROMAN. *"On Linguistic Aspects of Translation", in Brower, R. (ed.),* On Translation, *Haward University Press, 1959, pp. 232-239.*

JOHANSSON, S., G. LEECH & H. GOODLUCK. Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers, *Oslo, 1978.*

JOHANSSON, STIG. *"American and British English Grammar: An Elicitation Experiment",* English Studies 60, *1979, pp. 195-215.*

JOHANSSON, STIG. *"Corpus-based Studies of British and American English",* Papers from the Scandinavian Symposium on Syntactic Variation *(Stockholm, May 18-19, 1979), Almqvist & Wiksell International, 1980.*

JUILLAND, A. & CHANG-RODRIGUEZ, E. Frequency Dictionary of Spanish Words, *Haia, Mouton, 1964.*

KLAVANS & TZOUKERMANN, *"Dictionaries and Corpora: Combining corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons" 1995.*

KROGVIG, INGER. *"Shall, Will, Should, and Would in Present-Day American and British English. With Special reference Shall and Should in British English", Hovedfag thesis, University of Oslo, 1980.*

LARA, L. F. *"La Cuantificación en el Dicionario del Español de México", in* Computatioanl Lexicology and Lexicography. Special issue dedicated to Bernard Quémada, *Giardini Editori e Stampatori, Pisa, 1992, II, Vol. VII, pp. 1-27*

LAUSBERG, HEINRICH. Elementos de Retórica Literária, *trad. R.M. Rosado Fernandes, Fundação Calouste Gulbenkian, Lisboa, 1982, 3ª ed. (ed. original em alemão, 1967).*

LUFT, CELSO PEDRO. Mini Dicionário Luft, *Ed. Ática e Scitioni, 7ª ed, revista e ampliada por Francisco de Assis Barbosa, s/d.*

MATEUS, MARIA HELENA MIRA, ET AL. Gramática da Língua Portuguesa, *3ª ed. refundida, 1989 (1ª edição, 1971).*

MATEUS, MARIA HELENA MIRA. Actas do I Congresso Internacional da Língua Galego-Portuguesa na Galiza, *Ourense, 20-24 Setembro 1984, pp.297-303.*

MEDEIROS, JOSÉ CARLOS. Processamento Morfológico e Correcção Ortográfica do Português, *Tese de Mestrado, Instituto Superior Técnico, Lisboa, 1995.*

MONTES, JOSÉ JOAQUIM. *"La Delimitación de Lenguas: Cuestión Lingüística o Idiomática?", in Alfa, São Paulo, 1989, 33: 129-135.*

PRATA, MÁRIO. Dicionário de Português – Schifaizfavoire, *Editora Globo S.A., São Paulo (SP), 1993.*

ROBERTS, IAN & KATO MARY A. *(orgs.)*, Português Brasileiro, Uma Viagem Diacrônca, *Ed. UNICAMP, Campinas, SP, 1993.*

RYDÉN, MATS. *"Syntactic Variation in a Historical Perspective", in Sven Jacobson (ed.)*, Papers from the Scandinavian Symposium on Sintactic Variation *(Stockholm, May 18-19, 1979). Almqvist & Wikrell Internacional, 1979, pp. 37-45.*

SANTOS, DIANA. *"Português Computacional"*, Actas do Congresso Internacional sobre o Português, *Lisboa, 1994.*

SANTOS, DIANA "On grammatical translationese", Kimmo Koskenniemi (org.), Short Papers presented at NODALIDA'95, (Helsinki, May 1995).

SANTOS, DIANA & ENGH, JAN. "Appendix to Chapter 9: Use of PORTUGA for the two Norvegian Written Standards", in K. Jensen, G. Heidorn & S. Richardson, Natural Language Processing: the PLNLP Approach, Kluwer Academic Press, 1992, pp. 115-118.

TEYSSIER, PAUL. Manuel de Langue Portugaise (Portugal-Brésil), *deuxième édition révue et corrigée, Editions Klincksieck, Paris, 1984.*

VILAR, MAURO. Dicionário Contrastivo Luso-Brasileiro, *Editora Guanabara, Rio de Janeiro, 1989.*

WITTMANN, LUZIA HELENA E PEREIRA, MARIA DE JESUS. *"Português Europeu e Português Brasileiro: alguns contrastes", in* Actas do X Encontro da Associação Portuguesa de Linguística, *Évora, 1994.*