M. Céu Viana
CLUL

Ernesto d'Andrade
CLUL/FLUL

Luís C. Oliveira
INESC/IST

Isabel M. Trancoso
INESC/IST

UMA QUESTÃO DE EQUILÍBRIO

As bases de dados de fala natural ortográfica e foneticamente etiquetadas e gravadas em condições rigorosamente controladas têm vindo a assumir uma importância crescente na área do processamento automático de fala. O recurso sistemático a bases de dados deste tipo é imprescindível não só durante as fases de desenvolvimento dos sistemas de reconhecimento e de síntese, mas também durante as fases de teste e de avaliação de qualidade desses mesmos sistemas. Em consequência, esforços consideráveis têm vindo a ser desenvolvidos para várias línguas tanto no âmbito de programas nacionais (e.g. o British Corpus ou a BDSON) como europeus (e.g. o Projecto SAM)

Os trabalhos conducentes à constituição de uma base de dados de fala para o Português Europeu foram iniciados recentemente no âmbito da cooperação entre o Instituto de Engenharia de Sistemas e Computadores (INESC) e o Centro de Linguística da Universidade de Lisboa (CLUL) e alguns serão continuados no âmbito do programa europeu ESPRIT (Projecto SAM_A). O objectivo é recolher em condições rigorosamente controladas um vasto corpus de língua falada que contemple

um conjunto diversificado de falantes, diferentes situações de comunicação, modos de elocução, estilos e dialectos.

O trabalho de recolha, digitalização, segmentação, etiquetagem ortográfica e fonética de *corpora* de oralidade de grandes dimensões, que possam ser considerados representativos dos principais fenómenos de índole fonética ou fonológica que ocorrem numa dada língua, são, no entanto, extremamente morosos e exigem avultados meios humanos e materiais. Trata-se, por conseguinte, de um programa de trabalho a longo prazo que não se compadece com a urgência de algumas necessidades actuais.

Por outro lado, ainda que os materiais recolhidos sejam muitos e variados, este processo não dispensa a gravação de *corpora* pré-estabelecidos em modo leitura ou em modo entrevista, em que as sequências que se pretendem obter são elicitadas por diferentes meios. De facto, a vastidão de um *corpus* não garante, por si só, a ocorrência de sequências rigorosamente idênticas de falante para falante, nem a cobertura de uma variedade significativa de fenómenos fonéticos e fonológicos para um mesmo falante em diferentes contextos e situações comparáveis.

É fundamental poder-se dispor, a muito curto prazo, de um corpus multi-locutor, foneticamente equilibrado e de dimensões relativamente reduzidas que cubra as diferentes possibilidades fonotácticas do Português.

De um modo geral, diz-se que um corpus é foneticamente equilibrado quando são contemplados todos os sons que ocorrem numa dada língua ou dialecto e respeitada a sua frequência relativa de ocorrência. Este tipo de equilíbrio é fundamental mas não basta, nem para fins de investigação fundamental, nem para o desenvolvimento e avaliação de produtos que envolvam o processamento automático de fala.

Por exemplo, se se quiser estudar o conjunto de propriedades acústicas que caracterizam as vibrantes do Português Europeu, não basta gravar um qualquer conjunto de sílabas, palavras ou frases em que essas consoantes ocorram. É preciso que esse conjunto de materiais sonoros permita descrever as diferenças que existem no comportamento das vibrantes em função do contexto.

De igual modo, se se pretender testar um algoritmo de reconhecimento ou de síntese dessas mesmas vibrantes, ou de qualquer outra classe de sons, é importante poder fazê-lo sobre um corpus em que a percentagem de ocorrência de vibrantes nos diferentes contextos, corresponda à percentagem de ocorrência na língua. Só um equilíbrio fonotáctico poderá permitir uma avaliação realística do desempenho de um sistema ou a comparação efectiva da qualidade de diferentes sistemas.

Surpreendentemente, não existe para o Português Europeu nenhum estudo de índole estatística que, baseado num *corpus* relativamente vasto, trate das possibilidades fonotácticas desta língua[1]. Um estudo deste tipo é necessariamente prévio à constituição de um *corpus* foneticamente equilibrado para fins de investigação e desenvolvimento na área da fala.

Para fazer esse estudo de índole estatística, foram utilizados dois *corpora* de características bastante distintas. O primeiro é um *corpus* de frequência constituído a partir do conjunto de entrevistas coligidas pelo CLUL para o Português Fundamental[2] a que foram retirados os estrangeirismos, as siglas, os acrónimos e ainda um pequeno conjunto de formas não dicionarizadas que não foi possível identificar. Este *corpus*, designado PF_FON, contém cerca de 26000 formas diferentes que correspondem a cerca de 715000 ocorrências de formas de citação e de formas flexionadas, e a mais de 3000000 de símbolos ortográficos. O segundo *corpus*, designado DIC_FON, consiste num vocabulário de cerca de 83000 palavras a que correspondem aproximadamente 740000 símbolos ortográficos. A utilização de DIC_FON permite, por um lado, verificar as tendências observadas em PF_FON e, por outro, encontrar o conjunto de

---

[1] As considerações de ordem estatística de Barbosa (1965: 223-229) baseiam-se num texto extraído de *O Primo Basílio*, de 85 linhas, correspondendo à ocorrência de 2923 fonemas e, como o autor refere, apenas permite dar uma ideia aproximada da frequência dos fonemas no discurso. Os dados fornecidos por Delgado Martins(1975) baseiam-se num corpus especialmente concebido para um estudo de índole fonética e dizem respeito a cerca de 800 ocorrências de segmentos.

[2] cf. Nascimento, Marques e Segura (1987).

combinações fonotácticas que não derivam de processos flexionais e que podendo ocorrer em formas menos frequentes , não estão necessariamente contempladas em PF_FON.

Ambos os *corpora* foram processados com *Ler_PE*[3], um transcritor fonético automático para o Português Europeu, sendo o resultado desse processamento corrigido manualmente, de acordo com o dialecto e o estilo que constituem a opção por defeito do sistema[4]. É importante referir, contudo, que enquanto o tratamento gráfico e fonético de PF_FON se considera já concluido, a correcção fonética de DIC_FON ainda está em curso.

Como a relação existente entre a representação ortográfica e a representação fonológica é relativamente transparente, foram consideradas as correspondências entre a grafia e o som, uma vez que estas se revestem de um interesse mais imediato para o processamento automático de fala e que, ao dispensar a explicitação quer da teoria fonológica subjacente às transcrições, quer do nível da análise a que essa transcrição se situa, tornam as considerações estatísticas aqui apresentadas de uma utilização mais imediata para outros fins, de índole clínica ou pedagógica, por exemplo.

Para descrever os *corpora* foi constituída uma base de dados ortográficos e fonéticos do Português cuja estrutura fundamental é a apresentada em (1). Um campo comum de identificação das formas permite relacionar os registos dos diferentes ficheiros. Torna-se possível, assim, obter valores ponderados da frequência de ocorrência de estruturas silábicas, grafemas, fones, etc. No que diz respeito a DIC_FON não é possível, como é evidente, atribuir diferentes pesos às formas. Trata-se de um dicionário em que todas as entradas têm idêntico valor (=1) de um ponto de vista estatístico.

---

[3] cf. Viana, Andrade, Oliveira e Trancoso (1991)

[4] Trata-se de uma variedade possível da região centro em estilo formal, opção que nos parece adequada para o tratamento fonético de dicionários e que produz resultados aceitáveis na transcrição de textos livres.

(1)

Ficheiro 1	Ficheiro 2	Ficheiro 3
identificação	identificação	identificação
forma do corpus	sílaba ortográfica	grafema
frequência de ocorrência	sílaba fonética	fone
		categoria
		posição na sílaba
		posição na palavra
		grafema anterior
		grafema seguinte
		fone precedente
		fone seguinte

A cada entrada do ficheiro 1 correspondem, por conseguinte, tantas entradas nos ficheiros 2 e 3, quanto o número de sílabas e de grafemas, respectivamente.

A base de dados pode ser interrogada para obter informações que vão desde a frequência absoluta ou relativa de grafemas, fones, correspondências grafema-fone, classes de grafemas e fones ou tipos de sílaba, à comparação de dois corpora diferentes e verificação das respectivas semelhanças e diferenças.

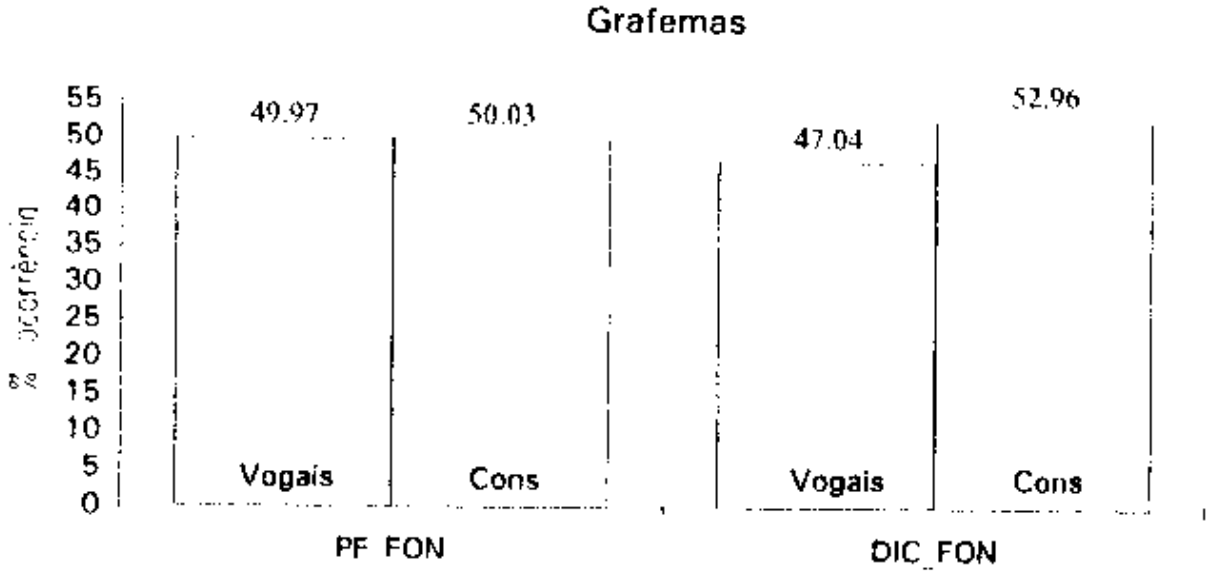
Uma descrição exaustiva dos resultados de índole estatística já obtidos através da interrogação desta base de dados não cabe, no entanto, no âmbito desta comunicação. Os exemplos e as considerações que se apresentam em seguida limitam-se, por conseguinte, a ilustrar apenas algumas das suas potencialidades. À selecção dos exemplos presidiu, contudo, o objectivo de disponibilizar alguns dados sobre o português[5].

Como se pode observar em (2), a percentagem de ocorrência de grafemas consonânticos é ligeiramente superior à dos grafemas vocálicos.

---

[5] Por falta de espaço não são aqui analisados e descritos todos os dados apresentados oralmente no Encontro da APL.

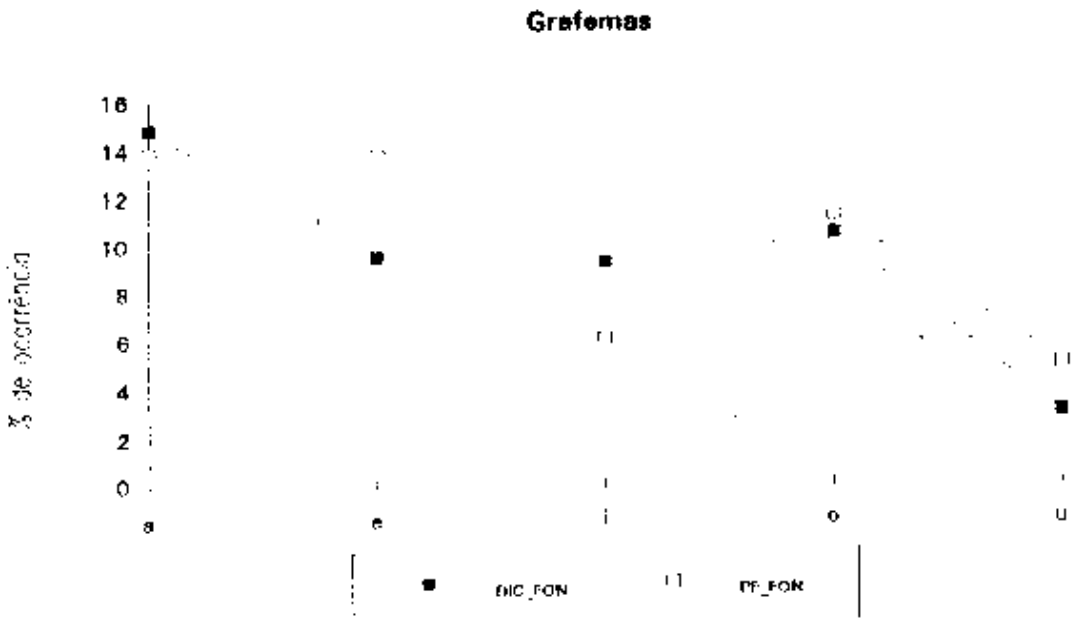
(2)

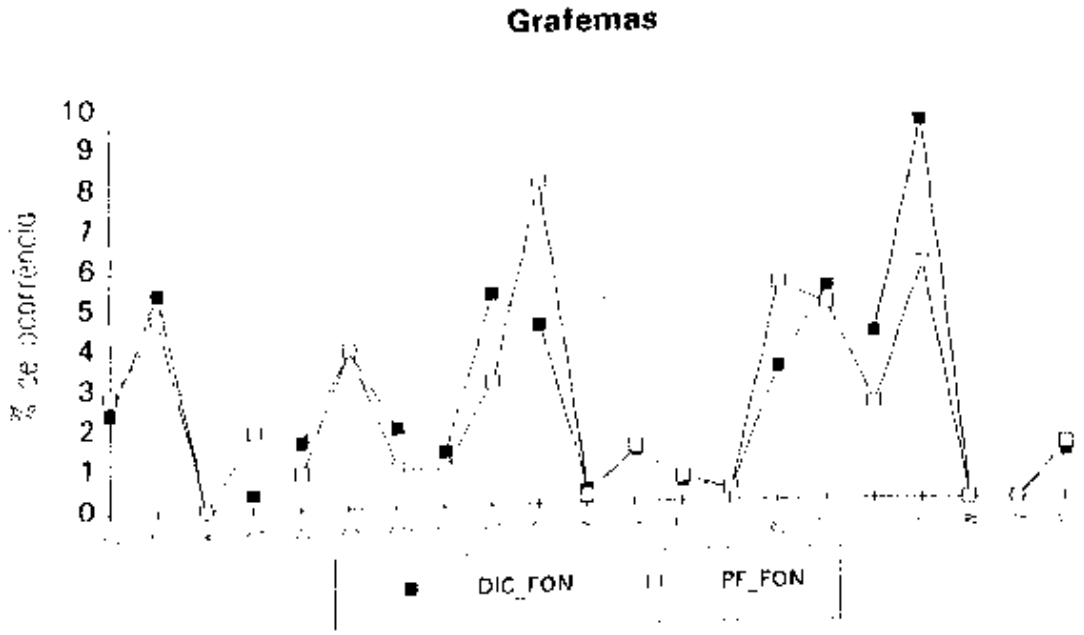


Repare-se, contudo, que enquanto em PF_FON a diferença de ocorrência das duas classes de grafemas é de 0.06%, em DIC_FON esta é já da ordem de 5,92%. Os *corpora* envolvidos não são, no entanto, do mesmo tipo: no primeiro, ocorrem formas em contexto, necessariamente flexionadas; no segundo, trata-se de um conjunto de formas de citação e, como tal, a maior parte das entradas de nomes e adjectivos está no masculino singular e todos os verbos se encontram no infinitivo.

Comparando estes resultados com os de (3) e (4) relativos à frequência de ocorrência de cada grafema em ambos os *corpora*, podem observar-se, efectivamente, algumas diferenças significativas. Em DIC_FON o grafema r é de longe o símbolo consonântico mais frequente e a sua percentagem de ocorrência é francamente superior à de PF_FON. Tal facto, é concerteza devido à existência de mais de 11000 verbos no infinitivo. Pelo contrário, o grafema m, que ocorre frequentemente em terminações verbais, assim como o grafema s, que ocorre nessas terminações e nos plurais de nomes e adjectivos, são mais frequentes em PF_FON.

(3)


Grafemas
% de ocorrência
16
14
12
10
8
6
4
2
0
DIC_FON
PF_FON


Grafemas
% de ocorrência
10
9
8
7
6
5
4
3
2
1
0
DIC_FON
PF_FON

É possível consultar a base de dados para pesquisar um qualquer conjunto de formas em função da sua frequência de ocorrência no corpus. Por exemplo, se essa consulta for efectuada para o, sobre PF_FON obtêm-se os seguintes resultados:

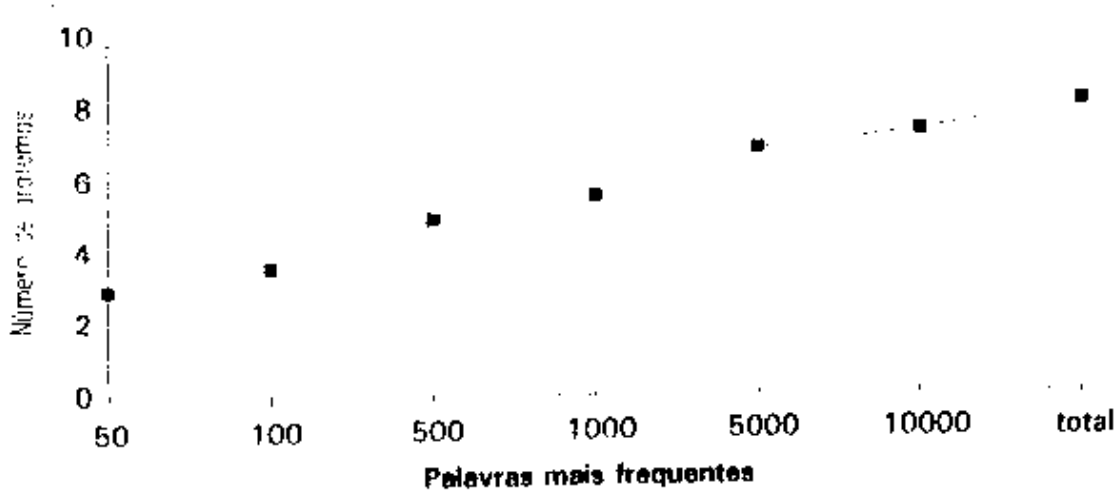
(5)

		Total de ocorrências
Número de formas com *o*:	14880	279621
10 formas mais frequentes em que ocorre pelo menos um *o*, (ordenação decrescente):	não	23033
	o	15516
	com	5380
	muito	5179
	os	5105
	por	5044
	porque	4831
	do	4451
	no	3550
	depois	3531

As palavras gramaticais são as principais responsáveis pelos elevados valores de frequência deste grafema. É natural que sejam algumas palavras gramaticais as que apresentam valores mais elevados de frequência, facto que provavelmente também explica algumas das diferenças observadas entre os dois corpora. Note-se que é também a frequência de ocorrência de formas como *que* ou *quando* que explicam as diferenças observadas entre os dois corpora para o grafema *q*.

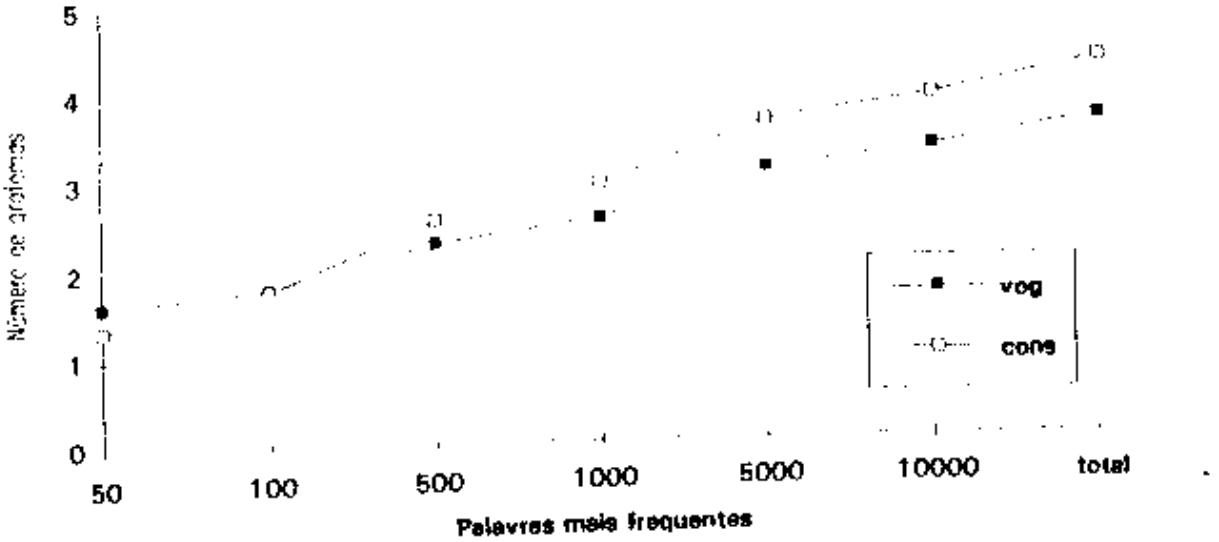
Apesar de o conjunto de factos apresentados justificar pelo menos uma parte das diferenças observadas na proporção de vogais e consoantes nos dois *corpora*, uma análise mais detalhada de (3) e (4) sugere claramente que outros factores devem ser tomados em conta. Uma vez que a frequência de ocorrência das formas no discurso traz algumas modificações significativas, pode supor-se que a estrutura dos vocábulos mais frequentes não deve ser idêntica à dos menos frequentes. De facto, a maior parte dos monossílabos com uma estrutura V, VC ou CV (32 casos) encontram-se entre as 50 palavras mais frequentes e constituem 24,1% da totalidade de PF_FON. Como (6) mostra, as palavras mais frequentes são, em média, as mais curtas.

(6)



É interessante notar, também, que a proporção entre o número médio de consoantes e de vogais por palavra, também se altera com a frequência, como se pode obervar em (7).

(7)

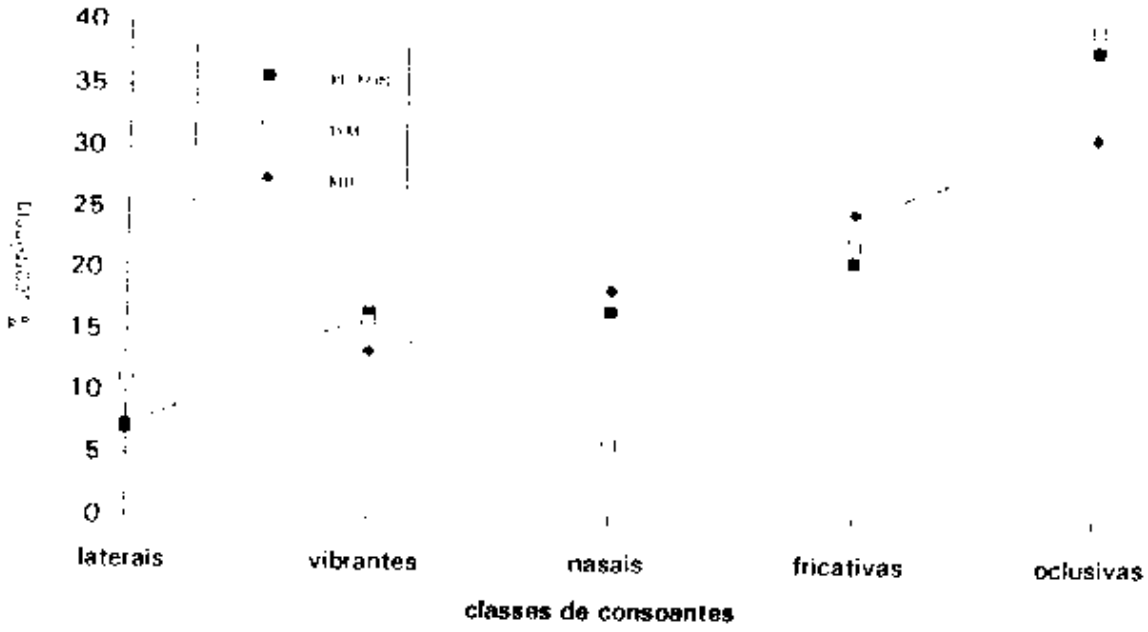


Tem-se vindo a insistir sobre as diferenças, mas as semelhanças que se podem encontrar entre PF_FON e DIC_FON são tanto ou mais importantes. Em qualquer dos corpora é manifesta a tendência para um equilíbrio entre o número de grafemas que representam vogais e consoantes e não se encontram grandes disparidades em termos de frequência relativa de ocorrência de grafemas. Uma vez que a ortografia do

português é de base fundamentalmente fonológica, parece lícito notar a este respeito que, em termos relativos, estes *corpora* manifestam tendências muito semelhantes às referidas em Barbosa (1965), nomeadamente no que diz respeito à predominância de sílabas abertas.

É interessante notar, por outro lado, que em termos gerais, também a distribuição relativa das classes de consoantes, é bastante semelhante à apresentada pelo mesmo autor[6] e ainda à referida por Delgado Martins (1975) embora, neste último caso, a ocorrência de consoantes nasais seja proporcionalmente muito menor.

(8)

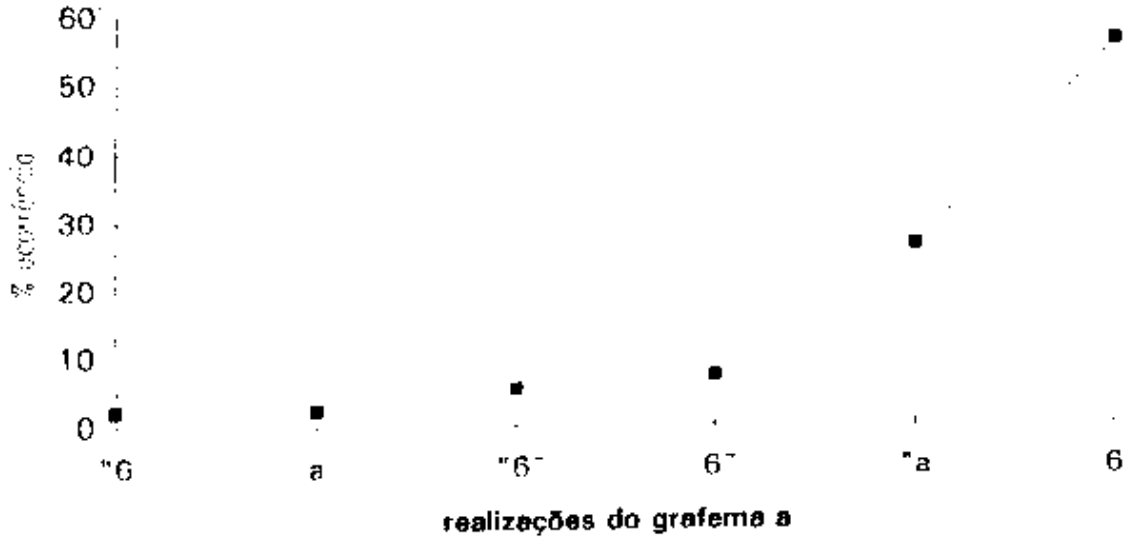


Como qualquer campo ou combinação de campos pode ser pesquisado, é possível analisar também as diferentes realizações fonéticas dos grafemas, como no exemplo em (9), onde se descrevem as possíveis realizações do grafema *a* em PF_FON, independentemente do contexto.

---

6 As percentagens referidas para Barbosa(1965) correspondem ao conjunto de ocorrências de cada classe de consoantes em posição inicial, medial e final de sílaba, relativamente ao número total de consoantes. Note-se a este respeito, no entanto, que na figura em (8) não estão contempladas as semivogais em posição final de sílaba, que este autor trata como consoantes.

(9)



É possível, no entanto, fazer apelo aos contextos adjacentes. Esta possibilidade, muito útil nas fases de desenvolvimento e teste de conjuntos de regras de transcrição fonética, tem vindo a ser utilizada, também, na construção de díades e tríades (difones e trifones) para o português. Em (10) são apresentados os 10 trifones mais frequentes, em posição inicial, medial e final de palavra. Utilizou-se o alfabeto SAM_PA e considerou-se que a transição em fronteira de palavra faz parte integrante do trifone.

(10)

posição inicial		posição medial		posição final	
ocorrências	trifones	ocorrências	trifones	ocorrências	trifones
24203	#n"6~	23082	n"6~w~	36063	"6~w~#
14904	#pu	13132	pur	34603	6S#
12819	#p6	11593	p6r	31628	uS#
10028	#m6	10248	um6	23320	r6#
9346	#um	9657	6r6	19351	tu#
8838	#6k	8545	m6S	19255	du#
8549	#pr	7451	m"e~t	14924	jS#
7326	#Sl	7268	muS	13236	d6#
7215	#k"o	7025	"ajS	13104	"er#
7039	#m"u~	6924	m"u~j~	12779	6~j~#

Estabelecer a lista completa dos trifones que ocorrem nos dois corpora e determinar a sua frequência relativa de ocorrência é um passo fundamental para se poder constituir um corpus fonética e fonotácticamente equilibrado, de dimensões

relativamente reduzidas. Desde que as condições de gravação e os critérios de controlo de qualidade se encontrem claramente definidos, este conjunto de materiais sonoros, depois de recolhido e tratado, virá a constituir o núcleo de uma base de dados de fala que poderá ser progressivamente alargada.

Apesar de os recursos humanos e materiais disponíveis na área do processamento de fala serem poucos, verifica-se que há vários grupos ou laboratórios nacionais (ou pelo menos mais do que um) a estudar os mesmos fenómenos e a trabalhar no desenvolvimento de produtos equivalentes. Seria muito vantajoso se fosse possível reunir alguns esforços a nível nacional, instituindo um conjunto de normas comuns para a elaboração de corpora de fala fonetica e fonotacticamente equilibrados. Só assim se poderá tornar possível uma comparação dos métodos de análise e dos resultados dos estudos realizados nos diferentes laboratórios. A principal vantagem será, contudo, a de facilitar a tarefa de construção de uma importante base de dados de fala para o Português.

Bibliografia

Andrade, E. e M. Céu Viana (1985) - "Corso 1 - Um conversor de texto ortográfico em código fonético para o Português". *RGFF*, 5, CLUL-INIC.

Barbosa, Jorge Morais (1965) - *Etudes de Phonologie Portugaise*. Junta de Investigações Cientificas do Ultramar, Lisboa. (2ª Edição - Universidade de Évora - Divisão de Línguas e Literatura, Évora, 1983).

Delgado Martins, M. Raquel (1975) - "Vogais e Consoantes do Português: Estatística de ocorrência, duração e intensidade". *Boletim de Filologia*, 24: 1-11.

Nascimento, Fernanda, Lúcia Marques e Luísa Segura (1987) - *Português Fundamental: Métodos e Documentos*. INIC-CLUL, Lisboa.

Oliveira, Luís C., M. Céu Viana e Isabel M. Trancoso (1991) - "DIXI - Portuguese Text-to-Speech System". *Proceedings of the 2nd European Conference on Speech Communication and Technology*, vol. 3: 1239-1242.

Viana, M. Céu, Ernesto d'Andrade, Luís C. Oliveira e Isabel M. Trancoso (1991) - "Ler_PE: Um utensílio para o estudo da ortografia do Português". *Actas do VII Encontro da Associação Portuguesa de Linguística*, Lisboa, 1992, pp. 474-489.