Contributos para o estudo da variação na frequência de ocorrência de unidades e padrões fonológicos

Joana Aguiar & Marina Vigário†*
*Universidade do Minho
† Faculdade de Letras da Universidade de Lisboa
Centro de Linguística da Universidade de Lisboa

Abstract

In this paper, we examine the frequency of phonological units and patterns in two oral *corpora* of European Portuguese: the *corpus* TA90PE, containing interviews from different regions of Portugal, and the *corpus* TQT, containing oral samples collected in *Terra Quente Transmontana*. The results contribute to determine the scope and limits of frequency variation, showing that the distribution of some of the phonological units and patterns examined is influenced by geographical and social factors, such as education and age. It is suggested that these later aspects, together with other linguistic data, may contribute to draw linguistic profiles in forensic analyses.

Keywords: Phonological units, phonological patterns, frequency, variation, forensic linguistics.
Palavras-chave: unidades fonológicas, padrões fonológicos, frequência, variação, linguística forense.

1. Introdução

Nas últimas décadas tem crescido o interesse pelos estudos de frequência, nomeadamente no âmbito da fonologia (e.g. Bybee, 2002; Bybee & Hooper, 2001; Pierrehumbert, 2001; Demuth, 2006; Gülzow & Gagarina, 2007, entre muitos outros; e, para o Português, Andrade & Viana, 1994; Vigário & Falé, 1994; Viana *et al.*, 1996; e vários trabalhos recentes de Vigário, Frota, Martins e colegas). Tal interesse justifica-se, designadamente, por ser cada vez mais claro o papel determinante que a frequência de palavras, unidades e padrões gramaticais desempenha em áreas como as do processamento

***Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2010, pp. 95-109**

e uso linguístico, ou a aquisição e desenvolvimento da linguagem: e.g. o efeito da frequência de palavras ou unidades e padrões gramaticais é reiteradamente detectado em tarefas de processamento (e.g. Dell, 1990; Caramazza *et al.*, 2001); sabe-se que a elevada frequência constitui um factor inibidor de processos de regularização (Bybee & Hooper, 2001); tem sido repetidamente mostrado que a frequência relativa de unidades e padrões fonológicos no *input* ao qual as crianças estão expostas tem um forte poder preditivo da ordem de emergência dessas unidades e padrões nas primeiras produções das crianças (e.g. Zamuner *et al.*, 2004; Freitas *et al.*, 2006; Ota, 2006; Prieto, 2006; Vigário *et al.*, 2006a). Dados de frequência parecem ainda distinguir línguas e famílias linguísticas. Por exemplo, a frequência dos diferentes tipos silábicos é muito diferente em línguas românicas, como o Português, o Castelhano ou Francês, e em línguas germânicas, como o Inglês ou o Holandês (cf. Vigário, Frota & Freitas, 2003 e referências aí citadas, e secção 5, abaixo).

Apesar do vasto conjunto de estudos desenvolvidos nesta área, desconhece-se quase por completo até que ponto a frequência de unidades e padrões fonológicos pode variar numa mesma língua. Para além disso, não se sabe se, havendo variação significativa, ela é determinada por factores intra ou extralinguísticos identificáveis.

2. Objectivos

Partindo da comparação de dois *corpora* de fala espontânea provenientes de variedades diferentes, o presente estudo pretende: (i) dar a conhecer dados novos sobre a frequência das classes maiores de segmentos, tipos silábicos, formatos de palavra e padrão acentual, numa variedade do Português; (ii) identificar objectos e padrões fonológicos cuja frequência de ocorrência é relativamente invariante, no sentido em que ela não distingue significativamente os *corpora* analisados; (iii) identificar padrões de variação que se correlacionam com factores externos, como a proveniência geográfica, idade e a escolaridade dos falantes; e (iv) identificar medidas de frequência potencialmente diferenciadoras de variedades e/ou grupos de indivíduos dentro de uma mesma variedade, com eventual aplicação em áreas como a da Linguística Forense.

3. Metodologia

Para o estudo da variação das unidades e padrões fonológicos comparámos dois *corpora* com características muito semelhantes: ambos são constituídos por entrevistas orais a indivíduos com diferentes características sociolinguísticas, transcritas ortograficamente. Um dos *corpora*, designado aqui e noutros trabalhos por TA90PE, é uma amostra do *corpus Português Falado. Documentos Autênticos* e contém dados de fala espontânea de indivíduos oriundos de diversas zonas de Portugal, recolhidos na década de 90 (ver Vigário *et al.*, 2006b), num total de 22994 palavras. O outro *corpus*, o *TQT*,

foi recolhido e analisado em Aguiar (2009) e é composto por 64757 palavras. Estes dados de fala espontânea foram recolhidos dos cinco concelhos da Terra Quente Transmontana: **Alfândega da Fé, Carrazeda de Ansiães, Macedo de Cavaleiros, Mirandela e Vila Flor** (cf. Mapa 1).

Apesar de serem comunidades do nordeste transmontano marcadamente rurais e demograficamente envelhecidas, há diferenças importantes na caracterização socioeconómica dos cinco concelhos que importa reter: a maioria da população desta região reside e trabalha nas cidades de Mirandela e Macedo, municípios dotados de mais emprego e de infra-estruturas culturais, económicas e industriais privilegiadas[1].

Mapa 1 – Área geográfica da Terra Quente Transmontana[2]

Para a constituição do *corpus* TQT foram entrevistados 100 informantes[3], estratificados de acordo com o concelho (**Alfândega da Fé, Carrazeda de Ansiães, Macedo de Cavaleiros, Mirandela e Vila Flor**), escolaridade (analfabetos/alfabetizados), sexo (masculino/feminino) e idade (intervalos etários de [20-35 anos], [36-50 anos], [51-65 anos] e [>65 anos]). Em nenhum caso se recolheu a fala de indivíduos a trabalhar ou a estudar fora da região de Trás-os-Montes, nem a residir fora dos concelhos em análise. Foram apenas entrevistados falantes que não (e)migraram por um período superior a doze meses nos últimos trinta anos. Todos os falantes com menos de 65 anos são profissionalmente activos.

Para a extracção dos valores de frequência no *Corpus TQT* foi usada a ferramenta electrónica *FreP* (FreP v1.0010 2004-2008, F. Martins, M. Vigário & S. Frota). Esta

[1] Dados do Anuário Estatístico da Região Norte – 2006, disponível em http://www.ine.pt/ (Abril 2008).

[2] Imagem retirada de http://www.terraquentedigital.espigueiro.pt (Junho 2007).

[3] Incluem-se nestas recolhas quinze entrevistas cedidas pela Rádio Ansiães (98.1FM).

ferramenta identifica e conta unidades e padrões fonológicos a partir de textos escritos, de acordo com as convenções ortográficas em vigor. Os valores de frequência relativos ao *corpus* TA90PE, usados neste trabalho para efeito de comparação, são os já publicados e extraídos com a mesma ferramenta.

Neste trabalho observaremos a distribuição dos tipos de segmentos, tamanho de palavra prosódica (PW), acento, e tipos silábicos mais frequentes. Para aferir se as variáveis externas influenciam significativamente a produção das unidades e padrões fonológicos aqui abordados utilizámos o teste do Qui-Quadrado baseado na Tabela de Contingência, com o valor de prova, doravante v.p., de 0,05.

4. Resultados

4.1. Comparação de *corpora*

A comparação dos valores de frequência obtidos para os dois *corpora* mostra que os valores relativos dos diferentes formatos de palavra não se distinguem significativamente nos dois *corpora* (*Corpus* TQT: PW1=28%, PW2=46%, PW≥3=26%; *Corpus* TA90PE: PW1= 29%, PW2=44%, PW≥3=27%).

De igual modo, a distribuição do acento é igual nos dois *corpora*, como vemos na Tabela 1.

Acento	TQT	TA90PE
Final	21,97%	21,56%
Penúltima	76,29%	76,44%
Antepenúltima	1,74%	1,99%

Tabela 1 - Valores relativos para a distribuição do padrão acentual nos *corpora* TQT e TA90PE.

Os dois *corpora* distinguem-se significativamente, no entanto, quanto à percentagem de ocorrência das grandes classes de segmentos Consoante e Vogal, e de V-*Slots*[4], sendo que estes últimos no *Corpus TQT* representam menos de metade dos encontrados no *corpus* TA90PE (cf. Tabela 2).

[4] Sobre a noção de *V-Slot*, ver Mateus & Andrade (2000). Palavras como *segmento*, que podem ser silabificadas como *se.gue.men.to*, ilustram a ocorrência de uma posição deste tipo, que pode ser preenchida por um *schwa*, introduzida entre duas consoantes quando a primeira não é licenciada pela coda e a segunda não é licenciada como segundo elemento de um ataque complexo (ver também Vigário & Falé, 1994).

	Corpora	
Classes de segmento	TQT	TA90PE
Vogal	**43,50%**	**48,00%**
Consoante	**50,70%**	**46,00%**
Glide	5,73%	5,80%
V-Slot	**0,07%**	**0,20%**

Tabela 2 – Valores relativos para as grandes classes de segmentos e *V-Slots* nos *corpora* TQT e TA90PE.

Também os valores de ocorrência dos tipos silábicos mais frequentes distinguem estatisticamente os dois *corpora*, concorrendo para tal particularmente as distribuições dissemelhantes dos tipos V, CVG e VN, como se pode observar na Tabela 3.

Tipos Silábicos	Corpus TQT	Corpus TA90PE
CV	46,47%	46,36%
V	**14,94%**	**15,83%**
CVC	10,62%	11,01%
CVN	5,47%	5,37%
CVGN	5,12%	5,62%
CVG	**3,69%**	**2,66%**
VC	3,09%	3,03%
CCV	2,87%	2,18%
VN	**1,85%**	**2,64%**
CVGC	1,38%	1,21%
VG	1,60%	1,51%
Outros	4,51%	4,09%

Tabela 3 - Valores relativos para os tipos silábicos mais frequentes nos *corpora* TQT e TA90PE (Frota *et al.*, 2006, e Vigário *et al.*, 2006b).

4.2. Frequência de unidades e padrões fonológicos no corpus TQT por variáveis externas.

Nesta subsecção analisamos a distribuição das unidades e padrões fonológicos em análise (classes de segmentos, tamanho de PW, tipos silábicos e padrão acentual) no *corpus* TQT, tendo em consideração as variáveis externas concelho, escolaridade e idade, já descritas na secção 3.

4.2.1. Tamanho de palavra prosódica

4.2.1.1. Distribuição por concelho

Apesar de os valores dos diferentes tamanhos de palavra prosódica para cada concelho serem bastante próximos (cf. Tabela 4), salientam-se algumas diferenças significativas (v.p. 0,00).

	Concelhos da TQT				
Tamanho de PW	Alfândega	Carrazeda	Mirandela	Macedo	Vila Flor
PW1	28,10%	29,11%	27,47%	27,73%	27,99%
PW2	46,56%	46,09%	46,30%	45,73%	46,93%
PW≥3	25,35%	24,80%	26,23%	26,54%	25,08%

Tabela 4 – Percentagem de Sílabas por Palavra Prosódica por Concelho.

Nos concelhos de Mirandela e Macedo, os dois centros de desenvolvimento industrial e cultural, regista-se a maior percentagem de palavras prosódicas com três ou mais sílabas (26,23% e 26,54%, respectivamente), em contraste com os valores para Alfândega (25,35%), Carrazeda (24,80%), e Vila Flor (25,08%).

4.2.1.2. Distribuição por escolaridade e idade

A distribuição por escolaridade mostra-nos que nos falantes analfabetos a percentagem de palavras prosódicas com menos de três sílabas é superior, ao contrário dos valores para as palavras prosódicas com três ou mais sílabas (cf. Tabela 5).

	Escolaridade		
Tamanho de PW	Alfabetizado < 65	Alfabetizado > 65	Analfabeto
PW1	28,22%	27,01%	27,52%
PW2	44,30%	47,24%	47,71%
PW≥3	**27,48%**	**25,75%**	**24,77%**

Tabela 5 - Percentagem de Sílabas por Palavra Prosódica por Escolaridade.

Como vemos na Tabela 6, a frequência de palavras prosódicas mais ou menos pesadas varia ao longo da idade do falante. Na faixa etária [51-65] são visíveis alterações na produção de palavras prosódicas de diferentes tamanhos, havendo um aumento de palavras prosódicas com três ou mais sílabas e concomitantemente um decréscimo de palavras prosódicas com uma sílaba. Já no grupo etário seguinte [> 65] há um aumento na produção de palavras prosódicas com uma e duas sílabas e uma diminuição da produção de palavras prosódicas de três ou mais sílabas (27% [20-35]; 27% [36-50]; 29% [51-65]; e 25% [>65]). É também de realçar o aumento progressivo da percentagem de palavras

prosódicas com duas sílabas ao longo das faixas etárias (44% [20-35]; 45% [36-50]; 45% [51-65]; e 48% [>65]).

	Idade			
Tamanho de PW	20-35	36-50	51-65	>65
PW1	29,50%	28,81%	26,35%	27,26%
PW2	43,61%	44,49%	44,80%	47,47%
PW≥ 3	**26,89%**	**26,70%**	**28,85%**	**25,26%**

Tabela 6 - Percentagem de Sílabas por Palavra Prosódica por faixa etária.

4.3. Tipos de Segmentos

4.3.1. Distribuição por concelho e idade

A análise percentual da contagem de grandes classes de segmentos (C, V, G e *V-slots*) por concelho (cf. Tabela 7) revela uma mínima variação nos valores para as *V-Slots* (v.p. 0,00), nomeadamente se compararmos os valores obtidos no concelho de Alfândega (0,04%) com os de Mirandela (0,08%) e Macedo (0,09%).

	Concelhos					
Segmentos	Alfândega	Carrazeda	Mirandela	Macedo	Vila Flor	Total
Consoantes	43,38%	43,38%	43,78%	43,53%	43,43%	43,50%
Vogais	50,72%	50,57%	50,68%	50,81%	50,69%	50,70%
Glides	5,89%	5,99%	5,46%	5,57%	5,80%	5,74%
V-Slots	0,04%	0,06%	0,08%	0,09%	0,07%	0,07%
Total	100%	100%	100%	100%	100%	100,00%

Tabela 7 - Percentagem de ocorrência dos tipos de segmentos por concelho.

No que diz respeito à influência da variável idade, verifica-se que a ocorrência de *V-Slots* é mais frequente nos falantes com menos de 65 anos: 0,09% [20-35]; 0,11% [26-50]; 0,10% [51-65]; 0,04% [> 65 alfabetizados]; 0,01 [> 65 analfabetos].

4.3.2. Distribuição por escolaridade

A distribuição das grandes classes de segmentos por falantes analfabetos e alfabetizados mostra-nos, novamente, oscilações significativas na percentagem de *V-Slots*: nos falantes alfabetizados com menos de 65 anos é superior, ocorrendo em 0,10% do total de classes de segmentos, principalmente em *advogado (a)* e nas formas do verbo *optar*;

nos falantes alfabetizados com mais de 65 anos o valor de *V-Slots* diminui para 0,04; e nas produções dos falantes analfabetos as *V-Slots* constituem apenas 0,01% do total de segmentos. Estes valores mostram que os falantes analfabetos quase nunca produziram palavras com combinações consonânticas que desencadeiam *V-Slots*. Interessantemente, nas poucas ocorrências de palavras deste tipo, nestes falantes nunca se verificou a inserção de vogal em núcleo vazio (e.g. *era o adbogado das troboadas*), ao contrário do verificado em outros falantes com mais de 65 anos (e.g. *os co**gun**omes dos reis*).

4.4. Padrão Acentual

4.4.1. Distribuição por concelho

Como podemos verificar na Tabela 8, a ocorrência de proparoxítonas é superior nos falantes provenientes das cidades de Mirandela (1,36%) e Macedo (1,72%), tendência já verificada nos valores para as palavras prosódicas com três ou mais sílabas.

	Concelho				
Acento	Alfândega	Carrazeda	Mirandela	Macedo	Vila Flor
Monossílabo	28,00%	29,39%	27,23%	27,46%	27,10%
Final	14,63%	15,86%	16,43%	14,97%	16,68%
Penúltima	56,05%	53,74%	54,98%	55,85%	55,15%
Antepenúltima	1,32%	1,01%	**1,36%**	**1,72%**	1,07%

Tabela 8 – Percentagem de ocorrência dos diferentes padrões acentuais por concelho.

4.4.2. Distribuição por escolaridade e idade

Em relação à escolaridade (cf. Tabela 9) e idade (cf. Tabela 10), verifica-se a mesma tendência já observada relativamente à frequência de palavra prosódica com três ou mais sílabas: os falantes analfabetos apresentam valores mais baixos para a acentuação na antepenúltima sílaba do que os calculados para os falantes alfabetizados (v.p. 0,00), o mesmo acontecendo com a acentuação final.

	Acento			
Escolaridade	Monossílabos	Final	Penúltima	Antepenúltima
Alfabetizado < 65	28,22%	15,96%	54,37%	1,45%
Alfabetizado > 65	26,99%	16,08%	55,79%	1,13%
Analfabeto	27,51%	14,61%	56,88%	1,00%

Tabela 9 – Percentagem de ocorrência dos diferentes padrões acentuais por escolaridade.

A distribuição do acento por idades revela, também, que no intervalo etário [20-35] há um aumento de monossílabos e uma clara diminuição de palavras acentuadas na penúltima sílaba. Finalmente é de realçar o aumento de palavras proparoxítonas nos falantes entre os 51 e os 65 anos (v.p.0,005).

	Acento			
Idades	Monossílabos	Final	Penúltima	Antepenúltima
20-35	**29,50%**	16,25%	**52,85%**	**1,40%**
36-50	28,81%	15,69%	54,08%	**1,41%**
51-65	26,35%	15,93%	**56,18%**	**1,54%**
>65	27,25%	15,35%	**56,33%**	**1,06%**

Tabela 10 - Distribuição da ocorrência do padrão acentual por intervalo etário.

4.5. Tipos Silábicos mais frequentes

4.5.1. Distribuição por concelho

À semelhança do verificado para os segmentos, também a distribuição dos tipos silábicos mais frequentes por concelho da TQT não revela oscilações significativas.

	Tipos Silábicos						
Concelhos	CV	V	CVC	CVN	CVG	CVGN	Outros
Alfândega	45,80%	15,35%	10,73%	5,50%	4,01%	5,50%	13,12%
Carrazeda	45,43%	15,47%	10,70%	5,30%	3,65%	5,44%	14,00%
Mirandela	47,22%	14,29%	11,01%	6,09%	3,62%	4,62%	13,15%
Macedo	46,66%	14,84%	10,19%	5,40%	3,30%	5,05%	14,56%
Vila Flor	47,26%	14,76%	10,49%	5,05%	3,86%	4,98%	13,59%

Tabela 11 - Distribuição dos tipos silábicos mais frequentes por concelho.

4.5.2. Distribuição por escolaridade

A análise dos tipos silábicos mais frequentes mostra-nos que, contrariamente ao observado em função do concelho, quando considerado nível de escolaridade dos sujeitos há diferenças significativas na percentagem de ocorrência de alguns tipos silábicos (v.p. 0,00). Assim, no Tabela 12 podemos ver que os valores para os tipos CV e V são mais altos nos falantes analfabetos; e os tipos CVC são mais altos para os alfabetizados.

	Tipos Silábicos						
Escolaridade	CV	V	CVC	CVN	CVG	CVGN	Outros
Alfabetizado < 65	**45,65%**	**14,65%**	**11,49%**	5,63%	3,46%	5,30%	13,82%
Alfabetizado > 65	**47,52%**	**14,74%**	**9,69%**	5,13%	4,16%	4,56%	14,19%
Analfabeto	**47,89%**	**16,00%**	**8,96%**	5,32%	3,90%	5,15%	12,78%

Tabela 12 - Distribuição dos tipos silábicos mais frequentes por nível de escolaridade.

As diferenças percentuais para cada intervalo etário são, uma vez mais, significativas (v.p. 0,00). Veja-se que a percentagem de ocorrência do tipo CV tem tendência para aumentar ao longo da idade (45% [20-35]; 45% [36-50]; 47% [51-65] e 48% [>65]), ao contrário do tipo CVC, cuja realização diminui ao longo dos intervalos etários (12% [20-35]; 12% [36-50]; 11% [51-65] e 9% [>65]).

	Tipos Silábicos						
Idades	CV	V	CVC	CVN	CVG	CVGN	Outros
20-35	**45,26%**	14,74%	**11,85%**	5,55%	3,44%	5,47%	13,69%
36-50	**44,96%**	14,89%	**11,59%**	5,55%	3,41%	5,57%	14,03%
51-65	**46,74%**	14,33%	**11,02%**	5,80%	3,53%	4,85%	13,74%
>65	**47,71%**	15,37%	**9,33%**	5,22%	4,03%	4,85%	13,49%

Tabela 13 - Distribuição dos tipos silábicos mais frequentes por intervalo etário.

5. Discussão

A comparação dos *corpora* TQT e TA90PE revela que a distribuição percentual de palavras com diferentes tamanhos e padrões acentuais é muito semelhante nos dois *corpora.*

O facto de o formato de palavra variar pouco nos dois *corpora* é tanto mais interessante quanto se sabe que o formato das palavras varia grandemente nas línguas – por exemplo, a frequência de monossílabos em fala dirigida a crianças no Inglês é de 80%, no Catalão é de 35%, no Espanhol de 26% (ver referências em Prieto, 2006), e de 45% no Português (cf. Vigário *et al.*, 2006a); e a frequência de palavras prosódicas com mais de 2 sílabas é de 5% no Inglês, 15% no Catalão e próximo de 30% no Espanhol e no Português (cf. Roak & Demuth, 2000; Prieto, 2006; Vigário *et al.*, 2006a).

Para além disso, mesmo nos parâmetros em que há variação entre os dois *corpora*, ela não deixa de ser muito diminuta. Efectivamente, os valores relativos ao *corpus* TQT nunca se distanciam muito dos anteriormente descritos para o PE, agrupando-se em ambos os casos, frequentemente, com os das línguas românicas e distinguindo-se claramente dos das línguas germânicas. Veja-se, por exemplo, a frequência dos diferentes tipos silábicos apresentados na Tabela 14, onde esse agrupamento é evidente. O mesmo

é também observado relativamente às palavras com mais de duas sílabas, cuja frequência varia num ponto percentual nos dois *corpora*, mas que ocorrem numa proporção muito semelhante a línguas como o Espanhol, e muito distinta da encontrada em línguas como o Inglês, ou mesmo o Catalão (em cuja história operou um processo de queda de vogal final responsável pela perda, em larga escala, da sílaba final de palavras polissilábicas, com impacto, por isso, no número de ocorrências de palavras maiores) (ver Tabela 15).

Tipo Silábico mais frequente				
	TQT	TA90PE	Português/Espanhol/Francês	Inglês/Holandês
CV(N)	52%	52%	> 50%	30 – 36 %

Tabela 14 – Distribuição percentual do tipo silábico CV (N) em línguas românicas e germânicas (cf. Vigário, Frota & Freitas, 2003 e referências aí citadas).

PW > 2 sílabas				
TQT	TA90PE	Português/Espanhol	Inglês	Catalão
26%	27%	~30%	5%	15%

Tabela 15 - Distribuição percentual de palavras prosódicas com mais de duas sílabas nos *corpora* em análise, no Português e Espanhol, Inglês e Catalão.

A invariância na frequência de ocorrência de palavras com diferentes tamanhos e padrões acentuais poderá indicar que estes parâmetros são independentes das especificidades de cada variedade (mesmos que variáveis externas possam associar-se a variações significativas a estes níveis dentro de cada dialecto, tópico a que regressamos abaixo). Antes, ela pode decorrer de propriedades invariantes da gramática da língua.

Há pelo menos três aspectos da gramática do Português que podem ter impacto nos parâmetros em causa: (i) a restrição de palavra mínima, que opera em muitas língua e que é considerada responsável pela impossibilidade de ocorrência de palavras mais pequenas do que duas sílabas ou moras, não está activa na língua, permitindo formatos de palavras impedidos noutras línguas (veja-se Vigário, 2003: 5.2 e Vigário *et al.*, 2006b a propósito desta questão); (ii) por outro lado, a estrutura silábica do Português é bastante simples, o que não permite a quantidade de contrastes que se pode obter com formatos de palavras muito pequenos em línguas como o Inglês, cuja estrutura silábica é bastante mais complexa; ou seja, numa língua como o Português, os contrastes lexicais dependem mais da ocorrência de mais sílabas, do que numa língua como o Inglês, onde a complexidade silábica permite muito mais contrastes em palavras com uma ou poucas sílabas; (iii) as regras de atribuição do acento são tais que ele incide regularmente sobre a última ou a penúltima sílaba das palavras; embora haja excepções definidas de modo idiossincrático, as palavras marcadas lexicalmente de tal modo que desencadeiam padrões excepcionais parecem ser relativamente raras.

Existem também dados de frequência que distinguem significativamente as duas variedades. Os nossos resultados mostraram variação significativa na ocorrência das grandes classes de segmentos, sendo que C são claramente mais frequentes no *corpus* TQT, ao contrário de V e de *V-Slot* que ocorrem em muito menor percentagem. Estes resultados são tanto mais interessantes quanto o facto de, com a excepção da frequência de ocorrência de *V-Slots*, as restantes duas classes de segmentos se comportarem uniformemente no interior desta variedade do Português.

Foi também observada variação significativa ao nível da frequência de ocorrência dos tipos silábicos mais frequentes, em particular pela menor ocorrência dos tipos V e VN no *corpus* TQT e a maior frequência do tipo CVG. O facto de os falantes analfabetos e mais velhos mostrarem valores mais próximos dos descritos no *corpus* TA90PE pode sugerir que estas diferenças decorrem de inovações da variedade da TQT, embora não tenhamos dados suficientes para explorar esta hipótese.

A análise da distribuição das unidades e padrões fonológicos aqui apresentada revela, ainda, que o comportamento dos falantes não é uniforme, mas condicionado por factores externos, como a idade e a exposição a processos de escolarização. Entre as variáveis externas mais consistentemente associadas a variações de frequência significativas no interior do *corpus TQT* está nível de escolarização: os indivíduos analfabetos quase não produzem V-*Slots* (0,01% contra 0,04% nos falantes alfabetizados com a mesma idade e 0,10% nos alfabetizados com menos de 65 anos); apresentam uma percentagem de ocorrência de acentuação antepenúltima significativamente mais baixa do que os alfabetizados (1,00% nos falantes analfabetos, 1,13 nos falantes alfabetizados com mais de 65 anos, e 1,45% nos falantes alfabetizados com menos de 65 anos); produzem mais palavras prosódicas com duas sílabas e menos com mais de três sílabas; e produzem mais tipos silábicos CV e V do que os alfabetizados.

No caso do aumento das sílabas de tipo V nos falantes analfabetos, ele pode ter sido impulsionado pela frequência de a-protéticos, fenómeno mais frequente nestes falantes (e.g. *depois alebantaram*; *alebantabamo-nos às seis da manhã*). O ligeiro aumento do tipo CV pode ainda ficar a dever-se à inserção de [ɨ] em final de sílaba (e.g. ***fize-te*** *algum mal*; *andaba só com obelhas e cabras era* ***difícile;*** *e atão* ***dize-me*** *o meu filho mai nobo*.

Destaca-se ainda a diminuição do tipo silábico CVC e um aumento do tipo CV nos falantes com mais de 65. Para estes resultados poderão contribuir as paragoges de [l] e [ɾ] em sílaba final fechada por consoante, formando uma nova sílaba, como em: *e se fosse a comer eram só* ***deze***; *nem* ***sequera*** *sabe*; *ou de* ***qualquera*** *maneira.*

Os dados mostram que os desvios mais importantes encontrados nas produções dos sujeitos analfabetos e dos mais idosos representam em geral um incremento das tendências de frequência de uso observadas na língua (e.g. menor uso das unidades e padrões mais raros, como *V-Slots* e palavras proparoxítonas, e maior uso das unidades e padrões mais frequentes, como os tipos silábicos CV e V e as palavras dissilábicas). A hipótese que levantamos é a de que aspectos como a eventual menor dimensão do léxico da população não-escolarizada e da mais idosa e a ocorrência de regras fonológicas *não-standard* podem estar relacionados com os factos observados.

Foram também encontradas diferenças resultantes da proveniência dos sujeitos, dentro da região da TQT: nos concelhos de Mirandela e de Macedo, centros de maior desenvolvimento industrial e cultural, a percentagem de ocorrência de *V-Slots* e a frequência de palavras maiores é superior ao verificado para os restantes concelhos, tal como a ocorrência de palavras proparoxítonas, numa inversão da tendência geral encontrada na língua. Uma vez mais, colocamos a hipótese de o contexto sociocultural favorecer a diversidade lexical, a qual se poderá correlacionar com o aumento relativo da frequência das unidades e padrões menos frequentes na língua. A relevância da dimensão do léxico como um dos factores explicativos de alguns dos resultados encontrados merece, contudo, um estudo específico, ficando a hipótese aqui levantada a aguardar validação.

6. Conclusões

Pretendeu-se com esta investigação contribuir para o conhecimento da amplitude e limites da variação na frequência de ocorrência de um conjunto de unidades e padrões fonológicos, bem como das razões subjacentes à presença / ausência de variação encontrada. Trata-se de um domínio largamente desconhecido, tanto no quadro do estudo do Português, como das restantes línguas.

Verificou-se que a frequência de uso de palavras com diferentes tamanhos e padrões acentuais não distingue as variedades estudadas. Pelo contrário, a frequência de *V-Slots* e de C e V, bem como de diferentes tipos silábicos, em particular V, VN e CVG, distingue-se nas duas variedades. Apesar disso, notou-se que, mesmo quando se observa variação significativa entre *corpora*, ela não deixa de ser relativamente diminuta, quando comparada com a que pode ser detectada entre línguas.

Dentro do *corpus* da Terra Quente Trasmontana, sobretudo o factor escolarização apareceu frequentemente associado a diferenças significativas, caracterizando-se as produções dos indivíduos analfabetos por um incremento das tendências de frequência observadas na língua (menos *V-Slots*, menos palavras proparoxítonas, mais tipos silábicos CV e V e mais palavras dissilábicas).

Os dados aqui apresentados e sua discussão trazem um contributo para várias áreas do conhecimento: (i) com os novos dados sobre a frequência de unidades e padrões fonológicos na TQT e sua comparação com um *corpus* contendo dados de diversas regiões de Portugal, deu-se aqui um contributo para os estudos sobre variação dialectal no Português; efectivamente, trata-se do primeiro estudo comparando dialectos do ponto de vista da frequência de ocorrência de unidades e padrões fonológicos; (ii) ao determinarmos zonas em que a variedade da TQT se afasta do já descrito para o Português, estabelecemos novos dados de referência para este dialecto, o que pode ser relevante, designadamente, no âmbito dos estudos sobre a aquisição da fonologia; (iii) finalmente, ao determinarmos zonas da fonologia em que a variedade da TQT, no seu todo ou apenas em certos grupos de falantes, se afasta do até aqui descrito para o Português, contribuímos para os estudos visando traçar perfis linguístico, os quais, juntamente com outros tipos de

dados linguísticos (e.g. estruturas morfossintácticas, léxico, regras fonológicas), poderão ter aplicação forense. Considerando este último aspecto em particular, julgamos poderem ter ficado lançadas as fundações para um novo domínio de investigação em larga escala, o da caracterização de indivíduos e grupos de indivíduos através de índices de frequência de uso de unidades e padrões linguísticos. Tal caracterização pode ter vantagem e/ou ser complementar a outras, como a feita com base acústica: ela pode incidir não apenas sobre dados orais (transcritos), mas também escritos; não exige a existência de suspeitos para estabelecer comparações; pode ser usada por investigadores criminais sem elevados níveis de especialização. Para além da ferramenta FreP, encontra-se presentemente em construção no Laboratório de Fonética da Faculdade de Letras de Lisboa uma base de dados frequência, a *FrePOP*, baseada num corpus actualmente com mais de três milhões e meio de palavras, que julgamos poder vir a ser, também, um contributo relevante para esta linha de investigação.

Referências

Aguiar, Joana (2009) *Unidades e processos fonológicos no falar da região da Terra Quente: contributos para a Linguística Forense.* Dissertação de mestrado, Universidade do Minho.

Andrade, Ernesto & Maria do Céu Viana (1994) Ainda sobre o Acento e o Ritmo em Português. *Actas do IV Encontro Nacional da Associação Portuguesa de Linguística.* Lisboa: APL/Colibri, pp. 3-15.

Bybee, Joan (2002) Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14, pp. 261-290.

Bybee, Joan & Paul Hopper (eds) (2001) *Frequency and the emergency of linguistic structure.* Amesterdam/Philadelphia: John Benjamins Publishing Company.

Caramazza, Alfonso, Albert Costa, Michele Miozzo & Yanchao Bi (2001) The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27, pp.1430-1450.

Dell, Gary (1990) Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* 5, pp. 313-349.

Demuth, Katherine (ed.) (2006) Special Issue on Crosslinguistic Perspectives on the Development of Prosodic Words. *Language and Speech* 49 (2).

Freitas, Maria João, Sónia Frota, Marina Vigário & Fernando Martins (2006) Efeitos prosódicos e efeitos de frequência no desenvolvimento silábico em Português Europeu. In *XX Encontro Nacional da Associação Portuguesa de Linguística. Textos Seleccionados.* Lisboa: APL/Colibri, pp. 397-412

Frota, Sónia, Maria João Freitas & Marina Vigário (2006) Early prosodic words in European Portuguese. Paper given at the 5th Workshop on Phonological Development, Potsdam, March.

Gülzow, Insa & Natalia Gagarina (eds) (2007) *Frequency effects in language acquisition: Defining the limits of frequency as an explanatory concept.* Berlin: Walter de Gruyter.

Mateus, Maria Helena & Ernesto d'Andrade (2000) *The Phonology of Portuguese.* Cambridge: Cambridge University Press.

Ota, Mits (2006) Input frequency and word truncation in child Japanese: Structural and lexical effects. *Language and Speech* 49(2): (Special issue on the Crosslinguistic Perspectives on the Development of Prosodic Words).

Pierrehumbert, Janet (2001) Exemplar dynamics: Word frequency, lenition, and contrast. Joan Bybee & Paul Hopper (eds.) *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins, pp. 137-157.

Prieto, Pilar (2006) The Relevance of Metrical Information in Early Prosodic Word Acquisition: A Comparison of Catalan and Spanish. *Language and Speech* 49 (2): 233-261(Special issue on the Crosslinguistic Perspectives on the Development of Prosodic Words).

Viana, Maria do Céu, Isabel Trancoso, Fernando Silva, Gonçalo Marques, Ernesto d'Andrade e Luís Caldas de Oliveira (1996) Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu. Inês Duarte & Isabel Leiria (orgs.) *Actas do Congresso Internacional sobre o Português*. Volume III. Lisboa: Colibri/APL, pp.481-517.

Vigário, Marina. (2003) *The Prosodic Word in European Portuguese*. Berlin/New York: Mouton de Gruyter.

Vigário, Marina & Isabel Falé (1994) A sílaba no português fundamental: uma descrição e algumas considerações de ordem teórica. *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, pp. 465-478.

Vigário, Marina, Sónia Frota & Maria João Freitas (2003) From Signal to Grammar: Rhythm and the Acquisition of Syllable Structure. A. Beachley (ed.) *Proceedings of the 27th Boston University Conference on Language Development.* Somerville: Cascadilla Press, pp. 809-821.

Vigário, Marina, Maria João Freitas & Sónia Frota (2006a) Grammar and frequency effects in the acquisition of the Prosodic Word in European Portuguese. *Language and Speech* 49(2):175-203 (Special issue on the Crosslinguistic Perspectives on the Development of Prosodic Words).

Vigário, Marina, Fernando Martins & Sónia Frota (2006b). A ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português. In *XXI Encontro da Associação Portuguesa de Linguística. Textos Seleccionados.* Porto: APL/Colibri, pp. 675-687.

Zamuner, Ania, Louann Gerken & Michael Hammond (2004) Phonotactic probabilities in young children's production of coda consonants. *Journal of Child Language* 31, pp. 515-536.