Frequências no Português Europeu: a ferramenta *FreP**

Marina Vigário[1], *Fernando Martins*[2 3] *e Sónia Frota*[2]
Laboratório de Fonética, Universidade do Minho[1], DLGR,
Universidade de Lisboa[2], ILTEC[3]

0. Introdução

Tem crescido nos últimos anos a consciência da importância dos efeitos de frequência em diversos domínios da linguística, nomeadamente no âmbito da fonologia e da aquisição da linguagem (e.g. Bybee, 2000 e 2001; Bybee e Hooper, 2001; Jurafsky, Bell e Girand, 2002; Moates, Bond e Stockmal, 2002; Pierrehumbert, 2002; entre muitos outros). Por exemplo, é sabido que palavras ou combinações de palavras muito frequentes são mais susceptíveis de redução do que palavras ou combinações de palavras pouco frequentes (e.g. Selkirk, 1984: 7.1.3; e Jurafsky, Bell e Girand, 2002, para o Inglês; Booij, 1995, para o Neerlandês; Vigário, 2003: cap. 7, para o Português Europeu) e que formações morfologicamente irregulares de elevada frequência tendem menos à regularização do que as de baixa frequência (e.g. Bybee e Hopper, 2001). Para além de um indicador do que é não-marcado nas línguas ou numa dada língua (e.g. Vigário e Falé, 1994), a frequência surge também como reveladora da actuação de restrições (*constraints*) nas línguas – Peperkamp (1997), por exemplo, considera que a baixa frequência de palavras monossilábicas no léxico do Italiano é demonstrativa da activação nessa língua da restrição de Palavra Mínima, que limita a duas sílabas ou moras o tamanho mínimo das palavras. No domínio da aquisição da linguagem, são também vários os trabalhos que vêm defendendo a importância dos efeitos de frequência no desenvolvimento linguístico infantil (e.g. Fikkert e Freitas, 1998; Lleó e Demuth, 1999; Beckman e Edwards, 2000; Roark e Demuth, 2000; Demuth e Johnson, 2003; Prieto, 2004) – por exemplo, Fikkert e Freitas (1998) sugerem que a maior rapidez no desenvolvimento de Codas preenchidas revelada por crianças que adquirem o Neerlandês, quando comparadas com as que adquirem o Português, decorre da maior frequência de Codas preenchidas nessa língua na fala adulta relativamente ao Português, e Demuth e Johnson (2003) sustentam que a violação da restrição de Palavra Mínima na fala de crianças que adquirem o Francês, em estádios em que essa restrição se mostra

* Agradecemos a Fernanda Bacelar do Nascimento e Luísa Alice Pereira, do Centro de Linguística da Universidade de Lisboa, a disponibilização graciosa do CD-Rom *Português Falado. Documentos Autênticos*, editado pelo Centro de Linguística da Universidade de Lisboa e Instituto Camões, de onde foram retirados os materiais analisados neste trabalho, bem como as questões e comentários feitos pela audiência do *XX Encontro Nacional da Associação Portuguesa de Linguística*.

***Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2004, pp. 897-908**

inviolável nas produções de crianças que adquirem outras línguas, se explica pela elevada frequência de palavras monossilábicas na fala adulta dirigida a essas crianças.

Entre os *corpora* disponíveis há mais tempo para o Português Europeu (PE) para extracção de informações de frequência encontra-se o Português Fundamental (cf. Bacelar, Marques e Segura da Cruz, 1987). Várias foram as manipulações a que foi sujeito, nomeadamente, para aferir frequências de tipos silábicos (Andrade e Viana, 1994; Vigário e Falé, 1994; Viana *et al.*, 1996), frequências de segmentos fonéticos e distribuição do acento de palavra (Viana *et al.*, 1996) ou frequência de items lexicais particulares (Vigário, 2003: cap.7). Ao contrário do material a partir do qual se pode fazer extracção de informação de frequência, contudo, não existem, até onde nos é dado conhecer, ferramentas que permitam conhecer informação sobre a frequência de unidades fonológicas e que sejam de domínio público.[1]

Interessados na expansão da informação disponível sobre as frequências relativas no Português de um conjunto de unidades linguísticas – não apenas fonológicas mas também morfossintácticas –, propusemo-nos construir uma ferramenta electrónica específica, capaz de fornecer automaticamente essa informação, o *FreP*. O presente artigo descreve o essencial das propriedades e funcionamento desta ferramenta. Para além disso, mostra resultados da sua aplicação a uma amostra do *corpus* do *Português Falado. Documentos Autênticos*, editado em CR-ROM pelo Centro de Linguística da Universidade de Lisboa e Instituto Camões. Esta amostra é constituída pelos dados do Português de Portugal da década de 90 (CD 1).

São aqui apresentados dados de frequência das seguintes unidades gramaticais: palavras monomoraicas – compostas por uma única sílaba terminada em vogal –, palavras monossilábicas, dissilábicas, trissilábicas e com 4 ou mais sílabas; palavras prosódicas *versus* palavras clíticas e seu tamanho em número de sílabas; proporção enclíticos / proclíticos fonológicos.

Note-se que, se bem que alguma desta informação esteja disponível para o *corpus* do Português Fundamental a partir dos trabalhos referidos acima, permitindo assim avaliar eventuais diferenças motivadas pela variável *corpus*, outra constitui informação nova sobre *corpora* do Português Europeu.

Embora ainda em fase de expansão e melhoramento, projecta-se que no momento da publicação deste artigo o *FreP* se encontre *online*, na página da Faculdade de Letras da Universidade de Lisboa (Laboratório de Fonética), com espelhos na página do Instituto de Letras e Ciências Humanas da Universidade do Minho (Laboratório de Fonética) e do Instituto de Linguística Teórica e Computacional. A partir do endereço da FLUL será possível aceder a uma demonstração do programa e seguir um conjunto de procedimentos que permitirão ao visitante da página correr o programa noutros *corpora*, bem como enviar comentários ou solicitações no sentido da melhoria da ferramenta. O programa tornar-se-á, assim, do domínio público, desde que claramente

[1] Integrados em sistemas de síntese de fala existem programas que poderiam permitir colher algum deste tipo de informação (M. Céu Viana, c.p.; J. João Almeida, p.c.). Contudo, segundo pudemos apurar, não são de domínio público e/ou não existem como módulos autónomos dos restantes elementos que compõem esses sistemas.

usado para fins de investigação e não comerciais. Pretende-se, finalmente, que a ferramenta possa crescer e tornar-se progressivamente mais poderosa através da sua eventual adaptação à pesquisa de outros tipos de informação de frequência.

1. Breve descrição da ferramenta *FreP*

O *FreP* é uma ferramenta electrónica primeiramente construída para extrair automaticamente, a partir de textos escritos, informação de frequência de unidades coincidentes ou relacionadas com a sílaba e a palavra no Português.[2]

Quanto às propriedades gerais desta ferramenta, o trabalho em progresso desenvolve-se no sentido de o *FreP* possibilitar: (i) uma utilização amigável, ao apresentar uma estrutura transparente e adoptar um sistema de janelas/comandos baseado no formato Windows; (ii) uma utilização personalizada, por exemplo, ao permitir ao utilizador o acesso a informação sobre a aplicação ou não de regras particulares ou a activação/bloqueio de regras específicas; (iii) a sua adaptabilidade, permitindo introduzir novos módulos para extracção de informação de frequência de outras unidades linguísticas e fazê-lo correr sobre diferentes *corpora*; (iv) a sua portabilidade, sendo compatível com outros sistemas operativos, como o Linux; (v) o seu uso de domínio público, com a disponibilização para a comunidade científica via *download*.

Quanto às sua funções específicas, na fase actual o *FreP* inclui algoritmos que permitem (i) localizar vogais, (ii) dividir sílabas, e (iii) determinar a presença e a localização do acento de palavra. Dependentes dessas três funções básicas, um conjunto de passos possibilitam extrair vários tipos de informação: (i) número de sílabas por palavra; (ii) número de palavras com uma, duas, três, ... N, sílabas; (iii) número de palavras morfossintácticas; (iv) número de palavras prosódicas; (v) número de palavras clíticas; (vi) número de palavras prosódicas com uma, duas, três, ... N, sílabas; (vii) número de palavras clíticas com uma ou duas sílabas; (viii) número de palavras prosódicas monomoraicas (e não-monomoraicas); (ix) número de palavras clíticas monomoraicas (e não-monomoraicas).

Encontra-se em progresso o desenvolvimento dos procedimentos para (i) a determinação da constituição silábica; (ii) a determinação da constituição silábica em função da posição na palavra; (iii) a determinação da constituição silábica em função do acento de palavra; e (iv) a determinação da posição na palavra da sílaba acentuada. Estes procedimentos permitirão extrair informação diversa, designadamente sobre a frequência de sílabas fechadas/abertas, iniciadas ou terminadas por zero, uma ou duas consoantes e/ou com núcleos simples ou complexos, tudo isto em função da posição na palavra ou da localização do acento.

---

[2] Importa sublinhar que a breve descrição que fazemos desta ferramenta na presente secção diz respeito ao momento da elaboração do artigo (e não ao da apresentação da comunicação, altura em que o produto estava em fase mais incipiente), sendo previsível que, na ocasião da publicação, o *FreP* se encontre já numa nova versão.

Apresentamos de seguida um conjunto de critérios seguidos para a determinação das unidades referidas nos parágrafos anteriores. Em cada caso, as decisões foram norteadas pelo conhecimento linguístico de que dispomos no momento presente acerca das unidades em causa e do funcionamento do sistema fonológico / gramatical do Português.

Sobre a detecção de vogais – Considerámos as vogais presentes no nível fonológico, ignorando transformações decorrentes dos processos *opcionais* de queda de vogal ou de semivocalização conducente a ditongos crescentes. Coerentes com este princípio e ao constatarmos que a semivocalização originadora de ditongos crescentes é obrigatória em posição pós-tónica, como em *família*, considerámos aqui que a glide que constitui este tipo de ditongo crescente não seria tratada como uma vogal. As sequências de grafemas <qu> e <gu> foram contadas como representando uma única consoante (labializada, em palavras como *língua*, ou não-labializada, em palavras como *líquido*). No caso das sequências de consoantes violadoras dos princípios de silabificação no Português (cf. Vigário e Falé, 1994; Mateus e Andrade, 2000; Mateus *et al.*, 2003), duas possibilidades foram admitidas: uma considerando a existência de uma posição vocálica entre as consoantes relevantes, de acordo com a proposta de tratamento fonológico destas sequências de Mateus e Andrade (2000), e uma outra em que tal vogal não existe.[3] A versão pública do programa permitirá que o utilizador escolha qualquer das opções.

Sobre a divisão silábica – Adoptámos os critérios de divisão silábica apresentados em Vigário e Falé (1994), Viana *et al.* (1996), Mateus e Andrade (2000). Pelo menos numa primeira fase, e para este efeito particular, no caso de sequências de grafemas que representam sequências de consoantes não respeitadoras do Princípio de Sonoridade, considerámos sempre e apenas os dados resultantes da introdução de uma nova posição vocálica (que pode ser superficialmente vazia ou realizada com um *schwa*, no Português Europeu, ou com um [i], no Português do Brasil). Para além disso, em sequências ortográficas representando VGV, assumiu-se que a glide é ambissilábica (na linha do sugerido em Vigário e Falé 1994) e, consequentemente, esse segmento surge associado tanto à sílaba que domina a vogal precedente como à que domina a vogal seguinte.

Sobre a contagem de sílabas – São duas as formas de contar sílabas possibilitadas pelo *FreP*: uma considerando apenas as posições vocálicas presentes e preenchidas no nível subjacente e outra em que são contabilizadas não apenas essas posições vocálicas mas também as posições vocálicas introduzidas entre sequências de obstruintes que, de outro modo, violariam o Princípio de Sonoridade.

Sobre a detecção da presença do acento de palavra – Os procedimentos relacionados com a identificação da presença de acento no léxico do Português foram

[3] Do ponto de ponto de vista fonológico ambas as informações podem ser interessantes. Do ponto de vista da fiabilidade dos resultados, contudo, a segunda permite resultados maximamente fiáveis, enquanto a primeira resulta na introdução de alguma taxa de erro devido à existência de consoantes mudas, de ocorrência imprevisível, tratadas nesta fase como as primeiras de um sequência de duas consoantes entre as quais é introduzida uma posição vocálica. No total das 22994 palavras ortográficas do *corpus* TA90PE estes casos deram origem 0.291% de erro. Devemos notar que se encontram em fase de implementação alguns procedimentos que permitirão diminuir este tipo de erros.

norteados pelos resultados do estudo de Vigário (2003) sobre a palavra prosódica e os clíticos no Português Europeu. Assumimos, assim, que todas as palavras lexicais e as gramaticais com mais de duas sílabas são portadoras de acento de palavra e que um conjunto designado de palavras gramaticais mono e dissilábicas não são portadoras de acento próprio (ver a listagem fornecida em Vigário, 2003: cap. 5). No caso de palavras com mais do que um radical ou com sufixos portadores de acento independente do da base morfológica a que se juntam, a ferramenta permite extrair os dados adequadamente nos seguintes casos: quando os elementos portadores de acento independente são separados por espaço branco ou por hífen; na generalidade dos casos dos advérbios em *–mente*; em muitos casos envolvendo sufixos *z*-avaliativos. Porém, permanece alguma taxa de erro decorrente de um conjunto de situações, entre as quais (i) a existência de compostos (envolvendo a concatenação de radicais ou de palavras – cf. Villalva 1994) que integram mais de uma palavra fonológica mas que não são graficamente separados por espaço branco ou hífen, (ii) a existência de palavras não sufixadas com terminações casualmente coincidentes com as formas dos sufixos *z*-avaliativos (e.g. *razão, cozinha*); (iii) a acentuação das bases morfológicas à quais se juntam sufixos com acento independente, quando monossilábicas e terminadas em sílaba aberta ou fechada por fricativa ou quando a localização do acento da base é de algum modo excepcional (e.g. *somente, orgãozinho*). Corrido o *FreP* em todo o *corpus* do Português de Portugal da década de 90 (CD1 do *Português Falado. Documentos Autênticos*), que passamos a designar por TA90PE, as situações descritas em (ii) e (iii) conduzem à introdução de uma taxa de erro de 0.178%.[4] Encontra-se em estudo a introdução de mecanismos que permitam minimizar estes vários tipos de erro.

Sobre a direcção de cliticização – Relativamente à direcção de cliticização e conduzidos ainda pela investigação detalhada de Vigário (2003: cap. 5), assumimos que todos os clíticos no Português Europeu são proclíticos, com a excepção dos clíticos pronominais pós-verbais e de *de* nas sequências *hei-de, hás-de* e *hão-de*.

Para finalizar esta breve descrição do *FreP*, importa fornecer de modo sistemático dados relativos aos níveis totais de fiabilidade obtidos no momento presente tendo em conta o número total de 22994 palavras ortográficas do *corpus* TA90PE:[5] a fiabilidade na identificação de palavras acentuadas e não acentuadas é de 99,935% e de 99,930% de acerto na localização do acento. A fiabilidade da contagem silábica, por seu turno, é de 99,709%. Quanto à distinção entre enclíticos e proclíticos, ela é exequível com taxa de acerto total, uma vez determinadas as palavras clíticas.

[4] Na avaliação de uma amostra do corpus TA90PE constituída por 4000 palavras não foram encontrados casos do primeiro tipo. Não foram também detectados nessa amostra outros tipos de problema que necessariamente se colocam ao bom desempenho do *FreP*, nomeadamente envolvendo siglas (e.g. *RFM*), abreviaturas (e.g. *etc.*) e dígitos (e.g. *1*).

[5] Ignorou-se para este efeito eventuais erros decorrentes da presença de palavras estrangeiras no *corpus*.

2. Aplicações do *FreP*: a importância da frequência em três áreas da gramática do Português

Apresentamos de seguida um conjunto de resultados da aplicação do *FreP* e suas implicações do ponto de vista da fonologia ou mais genericamente da gramática do Português. Todos os dados referidos foram corrigidos manualmente em relação aos erros descritos na secção anterior.

2.1. Tamanho mínimo de palavra e a restrição de minimalidade

Em línguas tão diversas como o Inglês, o Yidiny, o Árabe, o Japonês, o Lardil, o Estónio, o Bengali, o Baule, o Chamicuro, o Alemão ou o Catalão, a palavra prosódica apresenta um tamanho mínimo definido, sendo pelo menos dissilábica ou bimoraica (veja-se a revisão da literatura em Vigário, 2003: 1.6). Para dar conta deste facto, tem sido proposto que essas línguas são sensíveis a uma restrição sobre o tamanho da palavra, a restrição de Palavra Mínima (*Minimal Word requirement*). Em línguas como o Português (na variedade brasileira ou europeia), a possibilidade de encontrarmos palavras como *pé*, *mi* ou *nu*, levou autores como Bisol (2000) e Vigário (2003) a considerar que tal restrição não se encontra operativa nessa língua. Contudo, Vigário (2003: 159) não deixa de notar que, tendo em conta a lista do *Português Fundamental* (Bacelar, Marques e Segura da Cruz, 1987) que inclui cerca de sete mil formas flexionadas, apenas 138 palavras (lexicais) são monossilábicas, e destas apenas 28 constituídas por sílaba aberta. Tais baixos valores conduzem Booij (2004) a contrapropor que o Português é de facto sensível a restrições de minimalidade, mas que existe um reduzido número de palavras que a violam. Igual discussão decorre em relação ao Italiano e por razões similares (cf. Bafile, 1997 *versus* Thornton, 1996 e Peperkamp, 1997).

Os dados referidos em Vigário (2003) não têm em conta a frequência relativa das palavras listadas, tendo cada item da lista sido contabilizado apenas uma vez, independentemente do número de ocorrências. Contudo, uma observação que tenha em conta a frequência relativa de todas as palavras monossilábicas e monomoraicas no universo de todas as palavras prosódicas de um mesmo *corpus* não conduz necessariamente a resultados idênticos.

De modo a contribuir para a discussão, corremos o *FreP* sobre o *corpus* TA90PE, comparável na sua natureza com o do *Português Fundamental*, e extraímos o número de ocorrências de palavras prosódicas monossilábicas, monossilábicas com sílaba aberta (monomoraicas), e com duas, três ou quatro ou mais sílabas. Os resultados são apresentados no Quadro 1.[6]

[6] A contagem de sílabas teve em conta a introdução da posição vocálica entre consoantes cuja sequência violaria de outro modo o Princípio de Sonoridade.

PWs monossilábicas		PWs com mais de uma sílaba		
31,46		68,54%		
com sílaba fechada	com sílaba aberta	com 2 sílabas	com 3 sílabas	com 4 ou mais sílabas
11,66	19,80	42,55	18,35	7,64

Quadro 1: Distribuição de palavras prosódicas (PWs) em função do número e/ou constituição silábica no *corpus* TA90PE (valores percentuais relativos ao total de 17.162 PWs).

Os resultados mostram que a proporção das palavras monomoraicas / monossilábicas em relação aos restantes formatos de palavra é muito maior do que a revelada nas observações referidas acima. O efectivo uso de formas monomoraicas / monossilábicas aponta, assim, para que a palavra prosódica no Português (Europeu) não seja de facto sensível a restições de tamanho mínimo.

2.2. Tamanho de palavras, frequência relativa e aquisição do Português

Referimos na secção introdutória deste artigo que uma das áreas em que mais se tem explorado dados de frequência é a da aquisição da linguagem. Surge exactamente nesta linha a investigação de Vigário, Freitas e Frota (2005).

A literatura da especialidade tem avançado com a hipótese de o processo de aquisição estar condicionado nos primeiros estádios por restrições de Palavra Mínima e de Palavra Máxima, dado que as produções que ocorrem nesses estádios iniciais em línguas como o Inglês, o Neerlandês, Espanhol, Japonês ou Hebreu formam minimamente e maximamente um pé binário (ver revisão da literatura em Demuth e Johnson, 2003). Contudo, os dados da aquisição do Português Europeu apresentados em Vigário, Freitas e Frota (2005) não validam esta hipótese. Nesse trabalho mostra-se que o tamanho das palavras e sua frequência relativa na fala adulta podem explicar o facto. Apresentamos em seguida uma breve sinopse desse estudo nos aspectos que aqui nos interessam.

Já com o recurso ao *FreP* e tendo em conta dados da fala adulta decorrentes do presente trabalho, o estudo avalia a importância do tamanho das palavras na fala adulta, e sua frequência relativa, nos padrões de palavras exibidos nos primeiros estádios de desenvolvimento linguístico por crianças a adquirir o Português Europeu. Correndo o *FreP* sobre um segundo *corpus*, de fala dirigida à criança (CDS – *Child-Directed Speech*), e detectando um conjunto de diferenças nos resultados obtidos a partir de cada *corpus* de fala adulta, o trabalho identifica ainda o tipo de dados de fala adulta que melhor se correlaciona com as produções da criança.

Os materiais em que se baseia o estudo são constituídos por um total de 23.207 palavras de uma base de dados de CDS, 21.184 palavras de fala adulta retiradas de TA90PE, e 4.300 palavras de duas crianças em estádios iniciais de desenvolvimento linguístico. Depois de corrido o *FreP*, os resultados foram integralmente corrigidos manualmente e são os que resumimos de seguida.

Na fala da criança há uma clara predominância de formas dissilábicas, tal com sucede no CDS (46.6%) e na fala adulta não-dirigida a crianças, que passamos a designar por ADS (43.6%). Quanto às formas monossilábicas, elas ocorrem com bastante frequência na fala da criança (CS), atingindo os 28.6%. A sua incidência é, contudo, mais baixa do que a exibida no CDS (43.9%), mas muito próxima dos valores do ADS (29,5%). Relativamente às palavras com mais de duas sílabas, elas também surgem com valores muito expressivos no CS, valores que se aproximam muito dos do ADS (26.9), mas afastam do CDS (que não chega a ter 10% de formas com esta constituição).

Isto significa que nos primeiros estádios de produções de crianças a adquirir o Português há uma relativa elevada frequência de palavras com menos de duas sílabas e de palavras com mais de duas sílabas (violadoras das restrições de Palavra Mínima e de Palavra Máxima). Para além disso, os dados da criança correlacionam-se estreitamente com os da fala adulta (ADS). Os dados de frequência que caracterizam o Português na fala adulta podem portanto estar na origem da extracção dos padrões exibidos pela criança e sua frequência relativa. Outros trabalhos recentes mostraram que a aquisição de línguas como o Francês ou o Catalão revela efeitos de frequência similares (cf. Demuth e Johnson, 2003; Prieto, 2004). Este é pois um exemplo de como a frequência pode desempenhar um papel importante, neste caso, no âmbito do processo de aquisição e desenvolvimento da linguagem.

2.3. A importância da frequência na direcção de cliticização dos pronomes verbais átonos

Mostrámos nas secções anteriores como a frequência parece estabelecer relações intrincadas com a fonologia e, mais genericamente, com a gramática. Vigário (2003) sustenta que no Português a grande generalidade das palavras clíticas é fonologicamente dependente da palavra prosódica seguinte, sendo que apenas os pronomes átonos pós-verbais, e a preposição *de* em sequências com *hei-de*, *hás-de* e *hão-de*, se cliticizam à palavra prosódica precedente. Desde Frota (1994), por seu turno, que se notou a emergência de uma tendência, progressivamente mais generalizada, para colocação enclítica dos pronomes verbais átonos, mesmo em contextos típicos de próclise (veja-se também Duarte, Matos e Faria, 1995; e Duarte e Matos, 2000). Esta situação é interessante porque a colocação pós-verbal desses pronomes conduz necessariamente à sua encliticização fonológica. Dado que em Vigário (2003) não são consideradas as frequências de uso dos diversos clíticos, contudo, desconhece-se a efectiva proporção de enclíticos e proclíticos fonológicos. De modo a contribuirmos para a discussão sobre até que ponto a frequência se correlaciona e/ou condiciona a gramática também neste caso, apresentamos dados sobre a frequência de ocorrência de ambas as direcções de cliticização.

Uma contagem das palavras clíticas num extracto do TA90PE (correspondente a um total de 4827 palavras) revela que elas constituem 29,48% do total de palavras extraídas. Destas, apenas 3% coincidem com formas enclíticas. A percentagem de palavras proclíticas é pois larguíssima, atingindo 97% do total de clíticos fonológicos.

Os dados mostram que a colocação pós-verbal dos pronomes átonos conduz a uma direcção de cliticização fonológica muito pouco frequente na língua. Isso não parece impedir, contudo, que a tendência para essa colocação seja cada vez mais evidente. Tal situação revela, portanto, que, ao contrário do que verificámos nas secções anteriores, pelo menos em alguns casos, o uso (de estruturas / unidades / ordens de palavras decorrentes de exigências da gramática, por hipótese) pode não ser sensível a efeitos de frequência.

3. Aplicações futuras

O *FreP* poderá vir a ser aplicado a outros corpora, permitindo, por exemplo, estudar eventuais semelhanças e diferenças de frequência entre variedades do Português, entre tipos de texto, entre estilos, ou entre o discurso de grupos definidos por variáveis como idade, sexo, ou profissão. O estudo do primeiro aspecto acima referido encontra-se já em curso, com a aplicação da ferramenta ao corpus do Português falado no Brasil (CD 4 do *Português Falado. Documentos Autênticos*). Na Fig.1 são apresentados os resultados comparativos da composição de palavras prosódicas no Português Europeu e Português Brasileiro. São notáveis as semelhanças entre as duas variedades a este respeito, que podemos relacionar com as semelhanças existentes quanto à sílaba fonológica (Mateus e Andrade 2000, Frota e Vigário 2001).

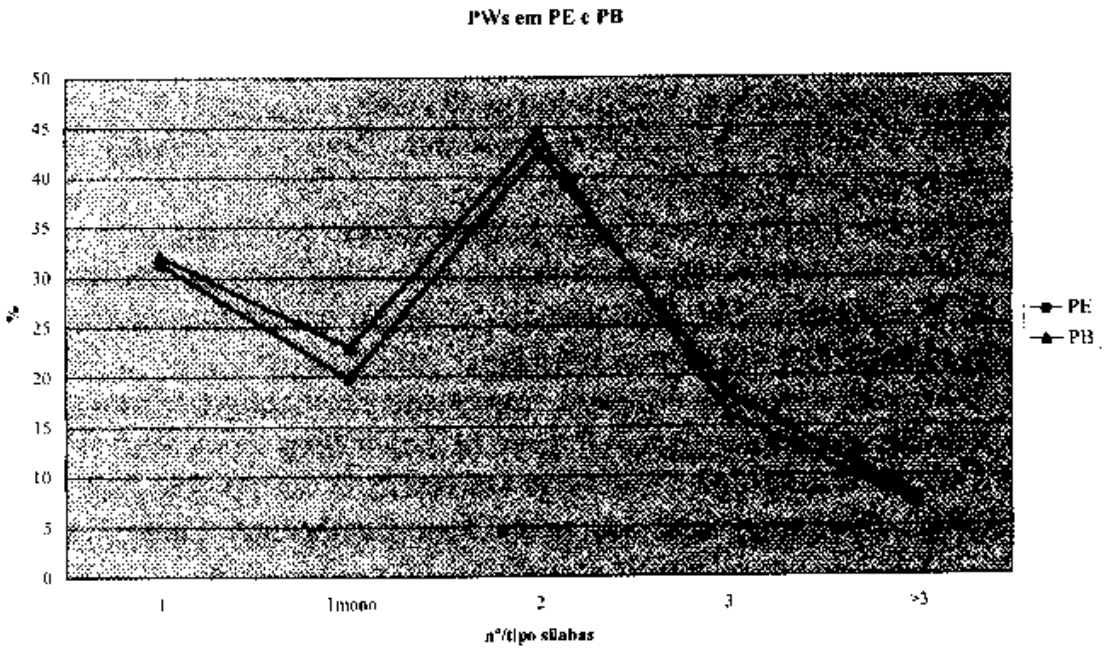

Figura 1. Composição das palavras prosódicas em PE e PB: PWs com 1 sílaba, com 1 sílaba aberta (monomoraicas), com 2, 3 e mais do que 3 sílabas. Dados de 17.162 (PE) e 16.222 (PB) palavras prosódicas.

Consta igualmente do plano de desenvolvimento desta ferramenta, estender o leque de unidades (morfo)fonológicas tratadas pelo FreP, sempre que tal seja útil como teste a hipóteses de análise da gramática do Português ou possibilite a obtenção de novos dados com consequências para a análise linguística.

4. Conclusão

Neste artigo apresentamos a ferramenta de detecção automática de frequências de unidades (morfo)fonológicas a partir de texto escrito, o *FreP*. Ilustramos a sua aplicação e as implicações dos resultados que dela decorrem em três áreas da fonologia: o papel da restrição de Palavra Mínima, o impacto do tamanho da palavra prosódica e sua frequência na aquisição e desenvolvimento da linguagem e a relevância dos dados de frequência para a colocação dos pronomes verbais átonos. Outras potenciais aplicações são enunciadas como estudos em curso ou caminhos a percorrer em investigação futura.

Referências

Andrade, E. & M.C. Viana (1994) Sinérese, diérese e estrutura silábica. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, pp. 31-42.

Bacelar, M.F., M.L. Marques & M.L. Segura da Cruz (1987) *Português Fundamental. Vol. II. Métodos e Documentos, Tomo 2: Inquérito de Disponibilidade*. Lisboa: INIC/CLUL.

Bafile, L. (1997) Parole grammaticali e struttura prosodica: dati dell'italiano e del napoletano. *Lingua e Stile* a. XXXII, 3, pp. 433-469.

Bisol, L. (2000) O Clítico e o seu Status Prosódico. *Revista de Estudos de Linguagem UFMG*. Belo Horizonte, 9(1), pp. 5-30.

Booij, G. (1995) *The Phonology of Dutch*. Oxford: Clarendon Press.

Booij, G. (2004) The morphology-phonology interface in European Portuguese. Review article of M. Vigário, *The Prosodic Word in European Portuguese*. *Journal of Portuguese* Linguistics 3(1), pp. 175-182.

Bybee, J. (2000) Lexicalization of sound change and alternating environments. In M.B. Broe e J.B. Pierrehumbert (eds.), *Papers in Laboratory Phonology V. Acquisition and the Lexicon*. Cambridge: Cambridge University Press, pp. 250-268.

Bybee, J. (2001) *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. & P. Hopper (2001) (eds.), *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Beckman, M. & J. Edwards (2000) Lexical frequency effects on young children's imitative productions. In M.B. Broe e J.B. Pierrehumbert (eds.), *Papers in Laboratory Phonology. Acquisition and the Lexicon*. Cambridge: Cambridge University Press, pp. 250-268.

Demuth, K. & M. Johnson (2003) Truncation to subminimal words in Early French. *Canadian Journal of Linguistics* 48, pp. 211-241.

Duarte, I. & G. Matos (2000) Romance Clitics and the Minimalist Program. In J. Costa (ed.), *Portuguese Syntax. New Comparative Studies*. Oxford: Oxford University Press, pp. 116-142.

Duarte, I., G. Matos & I. Faria (1995) Specificity of European Portuguese clitics in Romance. In I. Faria e M.J. Freitas (eds.), *Studies on the Acquisition of Portuguese*. Lisboa: APL/Colibri, pp. 129-154.

Fikkert, P. & M.J. Freitas (1998) Acquisition of syllable structure constraints: Evidence from Dutch and Portuguese. In *Proceedings of the GALA'97 Conference on Language Acquisition*. Edinburgh: University of Edinburgh, pp. 217-222.

Frota, S. (1994) Is Focus a phonological category in Portuguese? In P. Ackema e M. Schoorlemmer (eds.), *Proceedings of ConSole 1*. The Hague: Holland Academic Graphics, pp. 69-86.

Frota, Sónia & Marina Vigário (2001) On the correlates of rhythmic distinctions: the European Portuguese/Brazilian Portuguese case. *Probus* 13(2), pp. 247-275.

Jurafsky, D., A. Bell & C. Girand (2002) The Role of Lemma in Form Variation. In N. Warner e C. Gussenhoven (eds.), *Papers in Laboratory Phonolgy VII*. Cambridge: Cambridge University Press, pp. 3-34.

Lleó, C., & K. Demuth (1999) Prosodic constraints on the emergence of grammatical morphemes: Crosslinguistic evidence from Germanic and Romance languages. In A. Greenhill, H. Littlefield e C. Tano (eds.), *Proceedings of the 23rd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, pp. 407-418.

Mateus, M.H. & E. Andrade (2000) *The Phonology of Portuguese*. Oxford: Oxford University Press.

Mateus, M.H., I. Faria, I. Duarte, A.M. Brito & S. Frota, G. Matos, F. Oliveira, M. Vigário, A. Villava (2003) *Gramática da Língua Portuguesa*. (5ª edição revista e aumentada). Lisboa: Caminho.

Moates, D.R., Z.S. Bond & V. Stockmal (2002) Phoneme frequency in spoken word reconstruction. In C. Gussenhoven e N. Warner (eds.), *Laboratory Phonology 7*. Berlin/New York: Mouton de Gruyter, pp. 141-169.

Peperkamp, S. (1997) *Prosodic Words*. HIL Dissertations 34. The Hague: Holland Academic Graphics.

Pierrehumbert, J.B. (2002) Word-specific phonetics. In C. Gussenhoven e N. Warner (eds.), *Laboratory Phonology 7*. Berlin/New York: Mouton de Gruyter, pp. 101-139.

Português Falado. Documentos Autênticos (2001). CD-Rom produzido e editado pelo Centro de Linguística da Universidade de Lisboa e Instituto Camões.

Prieto, P. (2004) Early prosodic word acquisition in Catalan. Comunicação apresentada no *Second Lisbon Meeting on Language Acquisition – with special reference to Romance Languages*, Lisboa, Junho.

Roark, B., & K. Demuth (2000) Prosodic constraints and the learner's environment: A *corpus* study. In S. K. Howell, S. A. Fish e T. Keith-Lucas (eds.), *Proceedings of the*

24th Annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press, pp. 597-608.

Selkirk, E. (1984) *Phonology and Syntax. The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Thornton, A. (1996) On Some Phenomena of Prosodic Morphology in Italian: Accorciamenti, Hipocoristics and Prosodic Delemitation. *Probus* 8, pp. 81-112.

Viana, M.C., I.M. Trancoso, F.M. Silva, G. Marques, E., d'Andrade & L.C. Oliveira (1996) Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu. In *Actas do Congresso Internacional sobre o Português*, I. Duarte e I. Leiria (orgs.), vol. III. Lisboa: Colibri/APL, pp. 481-517.

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

Vigário, M. & I. Falé (1994) A Sílaba no Português Fundamental: uma descrição e algumas considerações de ordem teórica. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, pp. 465-477.

Vigário, M., M.J. Freitas & S. Frota (2005) Grammar and frequency effects in the acquisition of the Prosodic Word in European Portuguese. Submetido a *Language and Speech* (*Special Issue on the Acquisition of the Prosodic Word*, editado por Katherine Demuth).

Villava, A. (1994) *Estruturas Morfológicas. Unidades e Hierarquias nas Palavras do Português*. Dissertação de doutoramento, Universidade de Lisboa.

marina.vigário@mail.telepac.pt; fmartins@fl.ul.pt; sonia.frota@mail.telepac.pt