A frequência que conta na aquisição da fonologia: *types* ou *tokens*?*

Marina Vigário, Sónia Frota & Fernando Martins
Faculdade de Letras da Universidade de Lisboa

Abstract

We examine the frequency of a number of phonological units and patterns in European Portuguese, computed over tokens and over types, in adult speech, and compare it with the frequency and/or order of emergence of those units and patterns in children's early speech. We conclude that, whenever frequency information based on tokens and on types does not converge, it is always the frequency computed over tokens that correlates with the frequency patterns and/or order of emergence of those units/patterns in child speech. This investigation contributes to the understanding of the role of frequency in language acquisition, in addition to providing new frequency data for Portuguese.

Keywords: type frequency, token frequency, acquisition of phonology.
Palavras-chave: frequência em *types*, frequência em *tokens*, aquisição da fonologia.

1. Introdução

A investigação conduzida nos últimos anos sobre aquisição da gramática, e da fonologia muito particularmente, tem mostrado que a frequência de unidades e padrões linguísticos se correlaciona com a ordem de emergência e/ou frequência dessas unidades e padrões nas produções das crianças (e.g. Ingram, 1988; Roark & Demuth, 2000; Demuth & Johnson, 2003; Levelt & Van de Vijver, 2004; os artigos compilados em Demuth, 2006; Freitas, Frota, Martins & Vigário, 2006; Gülzow & Gagarina, 2007, entre muitos outros).

* Este estudo foi desenvolvido no âmbito do projecto *FreP: Padrões de Frequência na Fonologia do Português - Investigação e Aplicações*, PTDC/LIN/70367/2006. Gostaríamos de agradecer aos bolseiros do projecto, Marisa Cruz e Nuno Matos, pela colaboração na aferição atempada dos níveis de fiabilidade da ferramenta FreP apresentados na secção 2, bem como a Mary Beckman, que nos colocou recentemente uma questão que entroncava precisamente no tópico aqui investigado, mais concretamente, quais os dados de frequência que contam para o estabelecimento da restrição de palavra mínima. Agradecemos também a todos os colegas que nos cederam *corpora* para análise de frequência, em particular Fernanda Bacelar e Tiago Freitas.

***Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2010, pp. 749-767**

São várias as questões que se colocam uma vez determinado que a frequência no *input* desempenha um papel importante na emergência de unidades e padrões gramaticais do nível da palavra e abaixo. Uma dessas questões diz respeito ao tipo de dados sobre os quais a criança computa a frequência: sobre o vocabulário de *input* (listagem de palavras únicas – *types*) ou sobre a ocorrência de todas as instâncias das palavras (*tokens*). Até onde sabemos, esta questão não foi ainda tratada sistematicamente na literatura sobre aquisição e desenvolvimento da fonologia. Trabalhos como os de Marchman & Plunkett (1989) sugerem que a frequência computada com *tokens* prediz a aquisição mais precisamente do que a computada com *types*. Contudo, os estudos que têm mostrado correlações entre os valores de frequência no *input* e a emergência e/ou frequência de unidades e padrões na fala da criança são baseados em dados variáveis deste ponto de vista: nuns casos os dados são provenientes de dicionários ou léxicos (e.g. Monin, Loevenbruck & Beckman, 2007); noutros decorrem de *corpora* de fala, tipicamente considerando a frequência em termos de *tokens* (e.g. vários trabalhos sobre a frequência fonológica na fala dirigida à criança, como os citados em Prieto, 2006).

Sabe-se que os valores de frequência para *types* e para *tokens* numa mesma língua podem divergir muito, embora nem todos os aspectos da fonologia variem do mesmo modo (veja-se por exemplo Leung & Law, 2004, para valores de frequência de diversas unidades fonológicas baseados em *tokens* e em *types* no Cantonês de Hong Kong; ou os fornecidos para o Japonês por Ota, 2006). Para o Português, há poucos dados já disponíveis na literatura que mostrem que a frequência de ocorrência de unidades ou padrões fonológicos pode variar em função de se considerarem os dados de frequência de *types* ou de *tokens* (mas veja-se em particular Viana et al., 1996). Um conjunto de dados especialmente interessante para o que nos ocupa aqui diz respeito à ocorrência de palavras sub-mínimas (palavras compostas apenas por uma sílaba aberta). Avaliando se a restrição de palavra mínima (Minimal Word Constraint) está activa no Português Europeu, Vigário (2003) procedeu a uma contagem do número de *types* com esse formato no Portugês Fundamental (Bacelar & Segura, 1987), determinando que este tipo de palavra corresponde a uma percentagem muito diminuta nesse *corpus* (cerca de 0.4%). Contudo, os baixos valores apresentados contrastam com os obtidos por Vigário, Frota & Martins (2006), onde são consideradas as frequências em termos de *token* (atingindo aqui esse formato cerca de 7%).

Casos como os acima mencionados, em que há uma assimetria clara nos valores de frequência se considerados *types* ou *tokens*, permitem-nos determinar qual dos tipos de dados melhor se correlaciona com os da criança. Quando observado o comportamento das crianças a adquirir o Português Europeu, verifica-se que as chamadas palavras sub-mínimas aparecem cedo e mantêm-se com uma frequência significativa mesmo depois de passado o estádio inicial das primeiras produções, tal como sucede em línguas como o Francês (cf. Vigário, Freitas & Frota, 2006; Demuth & Johnson, 2003), mas contrariamente a outras línguas, como o Inglês, onde este tipo de palavras não é admitido. A explicação para esta assimetria no desenvolvimento fonológico tem sido atribuída à

diferente frequência de palavras submínimas no *input*. Contudo, pelo menos no caso do Português Europeu, esta justificação para a emergência precoce do formato de palavra sub-mínima só deverá estar disponível se se considerar que os dados sobre os quais a criança computa a frequência para efeitos de extracção dos formatos mais frequentes na língua são os *tokens* e não os *types*. O facto de nuns trabalhos se mostrar a relevância da frequência com contagens sobre *tokens* e noutros sobre *types* pode significar que os dados relevantes para a computação da frequência podem variar em função do aspecto fonológico em aquisição e/ou da fase de aquisição.

O trabalho que aqui se apresenta fornece um novo contributo para a compreensão da importância relativa da frequência de unidades e padrões fonológicos no *input* computada a partir de *types* e a partir de *tokens*, para a aquisição e desenvolvimento de diferentes aspectos da fonologia. Observando um conjunto de unidades e padrões fonológicos do Português Europeu, procedemos a um levantamento da sua frequência de ocorrência na fala do adulto, obtida através de uma computação (i) em termos de *tokens* e (ii) em termos de *types*. Os resultados obtidos são sistematicamente comparados com os resultados de frequência das mesmas unidades e padrões fonológicos nas produções de crianças numa fase precoce de aquisição. Nos casos em que não há coincidência na frequência extraída com base em *tokens* e em *types*, determina-se qual dos tipos de computação melhor se correlaciona com as produções das crianças.

O artigo organiza-se do seguinte modo: na secção 2 descreve-se a metodologia seguida para a selecção e tratamento dos dados; na secção 3 mostramos os resultados obtidos, considerando dados de frequência de formatos de palavra, padrão acentual, tipos silábicos, segmentos e ponto de articulação consonântico; na secção 4 discutimos os resultados e suas implicações para a compreensão da aquisição da fonologia.

2. Metodologia

Para as produções infantis, são aqui considerados os dados disponibilizados na literatura sobre a frequência e/ou ordem de emergência de unidades e padrões fonológicos nos estádios iniciais das produções de crianças a adquirir o Português Europeu: Vigário, Freitas & Frota (2006), para os formatos de palavra; Frota & Vigário (2008) e Correia (2008), para o padrão acentual; Frota *et al.* (2005) e Freitas, Frota, Martins & Vigário (2006), para os tipos silábicos; Freitas (1997) e Jordão (2009), para as classes de segmentos consonânticos na sílaba; e Costa, Freitas, Frota, Martins & Vigário (2007), para o ponto de articulação consonântico. Em todos os casos os dados são computados sobre *tokens*.[1]

Para a extracção de valores de frequência de unidades e padrões na fala adulta, foi constituído um *corpus* de fala expandido em relação ao que tem sido usado na literatura recente sobre a frequência na fonologia, integrando, para além do *corpus TAPE90*

[1] Remetemos para os respectivos trabalhos para uma descrição detalhada dos aspectos metodológicos relativos à recolha e ao *corpus* de fala infantil em análise.

(*Português Falado Anos 90*, CLUL), trabalhado em vários dos estudos recentes citados atrás, também outros dois *corpora* de fala espontânea: parte do *Corpus de Referência do Português Contemporâneo*, disponibilizado pelo Centro de Linguística da Universidade de Lisboa, e parte do *CORP-ORAL*, disponibilizado por Tiago Freitas (ILTEC). Trata-se de uma secção da base de dados *FrePOP – Frequency of Phonological Objects in Portuguese* (Frota, Vigário, Martins & Cruz, em curso), presentemente em construção no Laboratório de Fonética (CLUL/FLUL). O *corpus* totaliza 240.767 *tokens*, 16.702 *types*, 173.355 palavras prosódicas (PW), 72.525 clíticos, 447.331 sílabas e 933.411 segmentos.

A extracção dos valores foi feita com a versão mais actual da ferramenta electrónica *FreP* (Martins, Vigário & Frota, 2009, v.2), que possibilita obter automaticamente, a partir de texto escrito, valores de frequência de unidades e padrões fonológicos desde o nível da palavra, inclusive, até ao traço articultório. Uma avaliação da fiabilidade desta versão da ferramenta feita sobre um *corpus* de 76.000 palavras permitiu verificar níveis de acerto de acima dos 99% para todos os parâmetros (e acima de 99,6%, se excluídas palavras estrangeiras não adaptadas à grafia/fonologia da língua).

Para a separação dos dados em *types* e *tokens* no *corpus* do adulto, usou-se uma funcionalidade nova da ferramenta que permite listar todas as palavras únicas, juntamente com a sua frequência. Sobre as palavras assim listadas correu-se novamente o *FreP*, obtendo-se deste modo os valores de frequência de unidades e padrões fonológicos computados sobre os *types*.

3. Resultados

Nesta secção apresentam-se os resultados das contagens de frequência de ocorrência de um conjunto de unidades e padrões fonológicos, em termos de *tokens* e em termos de *types*, no *corpus* de fala adulta, em confronto com os dados de frequência e/ou ordem de emergência dessas unidades ou padrões na fala da criança (dados disponíveis na literatura computados sobre *tokens*). A nossa análise incide sobre a frequência dos formatos de palavra, do padrão acentual, dos tipos silábicos, das classes de segmentos consonânticos e por fim dos traços de ponto de articulação consonântico.

3.1. Formatos de palavra

Os valores relativos aos formatos de palavra, apresentados no Figura 1, mostram claramente que a contagem em termos de *tokens* está mais próxima dos valores de frequência exibidos nos primeiros estádios de produção infantil, quando comparada com a contagem em termos de *types*. Efectivamente, tal verifica-se não apenas para o formato sub-mínimo, mas também para todos os restantes formatos.

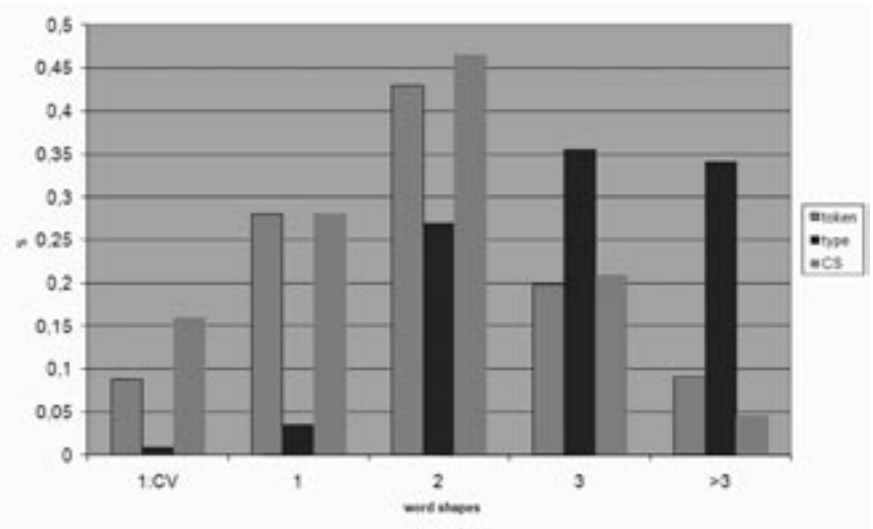

Figura 1: Valores percentuais de frequência dos formatos de palavra prosódica (palavras monossilábicas com sílaba aberta (1:CV), e palavras com 1, 2, 3 e mais de 3 sílabas). Dados de fala adulta computados sobre *tokens* e sobre *types*, e de fala da criança (CS).

Note-se que se confirmam os resultados de frequência do *input* baseados em *tokens* obtidos em Vigário et al. (2006), agora computados num *corpus* 10 vezes maior.

Comparando *tokens* e *types* na fala adulta, verificamos que há diferenças expressivas em todos os formatos: as palavras com menos de três sílabas são claramente mais frequentes nos *tokens* do que nos *types*, invertendo-se os valores quando consideradas as palavras com 3 ou mais sílabas; 2/3 dos *types* têm 3 ou mais sílabas, mas as palavras mais frequentes (*token*) têm 1 ou 2 sílabas; os monossílabos e palavras submínimas não têm expressão nos *types*.

Observada a frequência dos diversos formatos de palavra na criança, é claro que é a frequência baseada em *tokens* que se correlaciona directamente com os dados das crianças (CS), em todos os formatos.

3.2. Padrões acentuais

As descrições disponíveis sobre a distribuição do acento nas primeiras produções da criança revelam dificuldade inicial na produção do padrão acentual de acordo com o alvo (Correia, 2008; 2009; Frota & Vigário, 2008). Para além disso, o acerto surge mais cedo para o acento final do que para o penúltimo (Frota & Vigário, 2008).

O Figura 2 mostra a frequência das palavras em função da distribuição do acento no adulto.

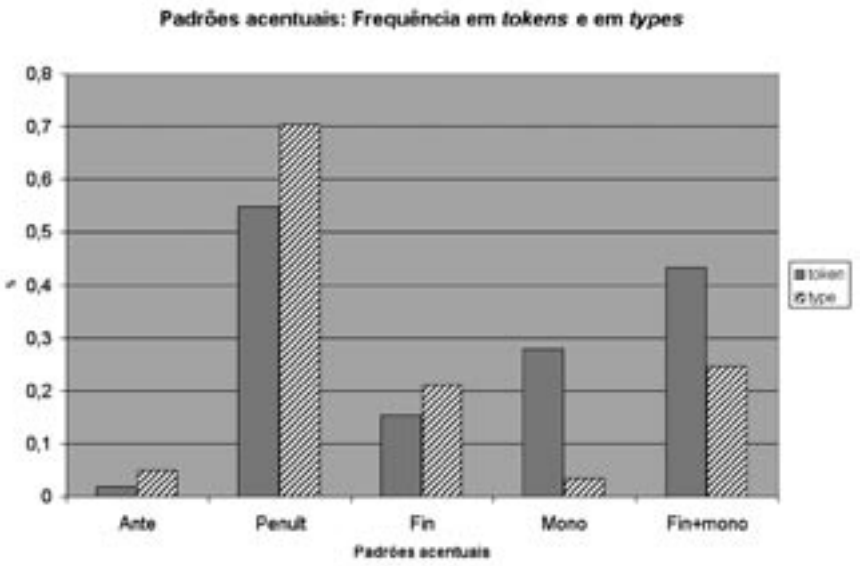

Figura 2: Distribuição do acento na fala adulta, computada sobre *tokens* e sobre *types*: palavras com acento antepenúltimo, penúltimo e final, palavras monossilábicas acentuadas e soma de palavras monossilábicas acentuadas e palavras com acento final.

Destacam-se valores superiores em todos os padrões nos *types*, se excluídas as palavras monossilábicas, que ocorrem muito mais frequentemente na contagem sobre *tokens*. Ignorando as palavras monossilábicas e a contagem em *tokens*, o padrão acentual que predomina é claramente o penúltimo.

Os dados assim considerados não se correlacionam com a dificuldade da criança na aquisição do padrão acentual do Português Europeu, e muito em particular o facto de os erros na colocação do acento deixarem de existir mais cedo nas palavras com acento final do que nas com acento penúltimo.

Podemos explicar esse padrão, porém, admitindo que as palavras monossilábicas também desempenham um papel na computação da distribuição do padrão acentual na língua. Efectivamente, basta que a criança saiba que no *input* o acento é computado a partir da margem direita da palavra (o que pode ser estabelecido muito cedo no processo de aquisição), para os monossílabos engrossarem o número de palavras com acento final. Nesse caso, a proporção de palavras com acento final é muito próxima das com acento penúltimo, se a computação for feita em termos de *tokens* (mas não em termos de *types*).

A proximidade verificada na frequência de acento penúltimo e de acento final no *input* (*token*) permite explicar a dificuldade observada nas primeiras produções da criança na colocação do acento. Porém, ela não justifica que o padrão de acentuação final surja mais cedo correcto, relativamente ao penúltimo.

Se considerarmos, contudo, apenas o formato de palavra mais frequente, tanto no adulto (71% das palavras acentuadas no nosso *corpus*) como na criança (75% nos dados de Vigário *et al*., 2006), isto é as palavras com menos de três sílabas, verificamos que a proporção de palavras com acento final (incluindo monossílabos) é superior à das com acento penúltimo (ver Figura 3). Porém, crucialmente, isso é verdade novamente apenas se considerada a contagem em termos de *tokens*, mas não em termos de *types*.

Figura 3: Distribuição do acento na fala adulta nos formatos de palavras mais frequentes (com 1 e 2 sílabas), computada sobre *tokens* (painel da esquerda) e sobre *types* (painel da direita).

Para além da importância da frequência no *input* computada sobre *tokens* e não sobre *types*, os dados mostram o papel crucial das PWs de 1 sílaba para a preponderância do acento final sobre o acento penúltimo. Como dissemos, a partir do momento em que a criança determina que a atribuição de acento é feita nesta língua com referência à fronteira direita da palavra, os monossílabos podem ser tratados como exibindo acento final (relativamente à margem da palavra relevante). Para além disso, sabe-se que no Português, com a excepção de uma diminuta fatia dos pronomes clíticos verbais, todas as palavras clíticas se adjungem à palavra prosódica seguinte (cf. Vigário, 2003), sendo muito rara a ocorrência de enclíticos na língua (num estudo exploratório, Vigário *et al*., 2005, referem que apenas cerca de 3% dos clíticos incorporam à palavra prosódica precedente; estes valores são confirmados pelos nossos dados, baseados em 245.880 palavras prosódicas e clíticas, onde apenas 3.2% dos clíticos são enclíticos). Isto significa que muitas das palavras monossilábicas podem, de facto, ser processadas como dissilábicas com acento final, no caso de serem precedidas de um proclítico. Note-se que, sendo a ocorrência de clíticos na língua bastante frequente (com 29,5% nos dados de Viana *et al*. 2005 e os mesmos 29,5% no *corpus* 10 vezes maior presentemente em análise - contagem feita sobre *tokens*), é concebível que o número de palavras dissilábicas por via da adjunção de um clítico possa subir em quase um terço. Para além disso, se considerados os tipos silábicos que se destacam em posição tónica relativamente à posição átona (CVGN, CVN, CVG, VN, VG e CVGC – cf. Vigário *et al.* 2006), verificamos que eles totalizam 23.2%

das sílabas tónicas iniciais de palavras com mais de uma sílaba, 33,4% das sílabas tónicas finais de palavras com mais de uma sílaba e 53,9% das sílabas acentuadas que compõem os monossílabos (este último valor desce para 47.2% se excluído o tipo VN, proveniente da palavra *um*, cujo estatuto de palavra prosódica pode ser discutível). Consequentemente, a distribuição das sílabas dos monossílabos acentuados é claramente mais próxima da das sílabas tónicas finais do que da das sílabas tónicas iniciais.

3.3. Tipos silábicos

Observemos agora os resultados relativos à frequência dos diversos tipos silábicos. Como pode ser visto no Figura 4, *tokens* e *types* exibem diferenças na frequência de ocorrência dos tipos silábicos mais frequentes: a ordem de frequência obtida para *tokens* é CV>**V**>CVC>CVN>CVGN>CCV; enquanto para *types* é CV>CVC>CVN>**V**> CCV>CVGN. Sabe-se que V emerge logo nas primeiras produções infantis, pois a ordem de emergência dos tipos silábicos é CV,V>(C)VN>(C)VG> (C)VC (de acordo com Frota *et al.*, 2005, e Freitas *et al.,* 2006). Assim, os dados de frequência que melhor se correlacionam com a ordem de aquisição são os baseados nos *tokens*. Efectivamente, nos *tokens* V é o segundo tipo mais frequente, enquanto nos *types* este tipo silábico é apenas o quarto mais frequente.

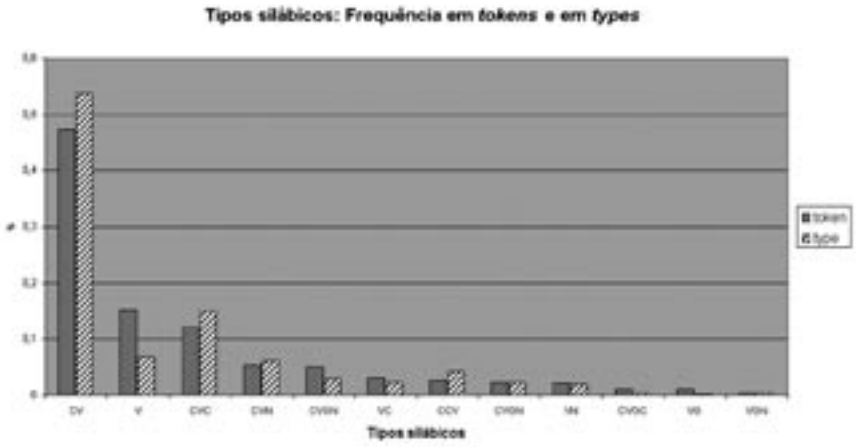

Figura 4: Distribuição dos tipos silábicos mais frequentes (contagem sobre *tokens* e sobre *types*).

A frequência no *input*, quer computada sobre *tokens* quer sobre *types*, não explica, contudo, por que razão os tipos (C)VN e (C)VG surgem antes de (C)VC e o tipo V aparece a par do tipo CV, quando no *input* este último é muitíssimo mais frequente.

Frota *et al*. (2005) e Freitas *et al*. (2006) explicam a ordem de emergência de V e (C)VN e (C)VG pela interacção entre frequência e posições proeminentes (posições

acentuadas e junto aos limites da palavra – computação feita sobre *tokens*). As Figuras seguintes permitem comparar os dados obtidos contando a frequência de cada tipo em posição acentuada (Figura 5) e em função da posição na palavra (Figuras 6 e 7), em *tokens* e em *types*.

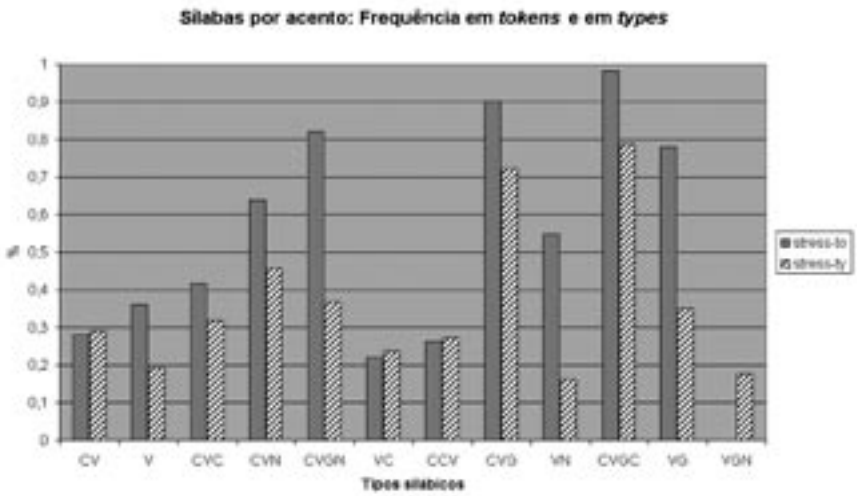

Figura 5: Distribuição dos tipos silábicos mais frequentes em posição acentuada (contagem sobre *tokens* e sobre *types*).

Da observação do Figura 5 ressalta o facto de os tipos (C)VN e (C)VG serem claramente mais frequentes em posição proeminente do que o tipo (C)VC, sendo esta diferença mais expressiva se considerados *tokens*, e não *types*. Para além disso, quando considerada a posição acentuada, o tipo silábico V chega a ser mais frequente do que o tipo CV, isto se, uma vez mais, considerada a contagem sobre *tokens*.

A distribuição da frequência dos tipos silábicos em função da posição na palavra, mostrada nos Figuras 6 e 7, revela que V ocorre quase exclusivamente em posição inicial de palavra ou em monossílabos, se considerada a contagem sobre *tokens* (91% na contagem sobre *tokens* vs. 65% apenas na contagem sobre *types*), enquanto a frequência de ocorrência do tipo CV se distribui pelas várias posições possíveis, incluindo a interna. E a mesma tendência, embora com valores um pouco mais baixos, é observada nos tipos (C)VN e (C)VG . Uma vez mais, a contagem sobre *tokens* produz resultados mais expressivos do que a sobre *types* (considerando os resultados dos diversos tipos com e sem consoante inicial, respectivamente, 65 a 89% dos tipos (C)VN ocorrem em posição inicial e em monossílabos nos *tokens* vs. 47 a 88% nos *types* e 64 a 99% dos tipos (C)VG ocorrem nessas posições nos *tokens* vs. apenas 32 a 85% nos *types*). Se (C)VG (N/C) ocorrem predominantemente em monossílabos acentuados (em quase todos os formatos acima de 50%, chegando a 90% na contagem sobre *tokens*), já (C)VC surge nesta posição apenas em 22 a 30% dos casos (*tokens*). Note-se que, uma vez mais, é na contagem sobre

tokens que os resultados melhor distinguem o tipo (C)VC do (C)VG(N/C), uma vez que, sobre *types*, o primeiro tipo ocorre em monossílabos 0.4 a 0.3% e o segundo entre 0.3 e 17% dos *types*).

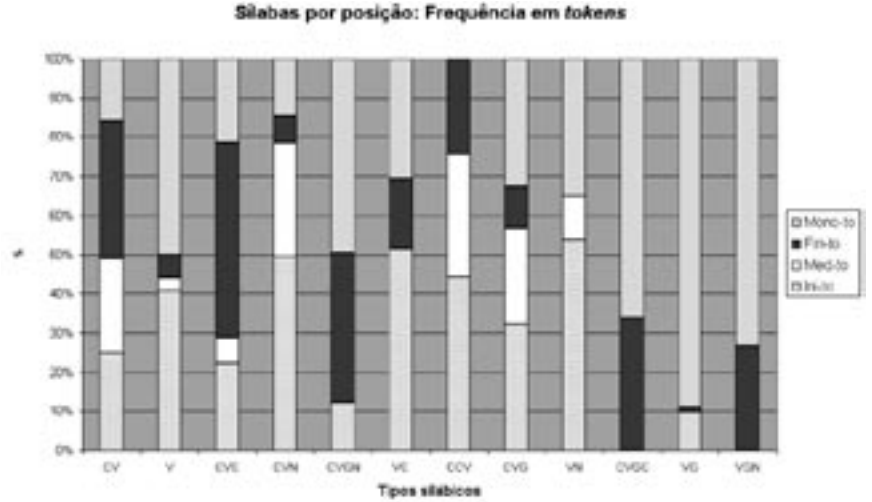

Figura 6: Distribuição dos tipos silábicos mais frequentes em função da posição na palavra (contagem sobre *tokens*).

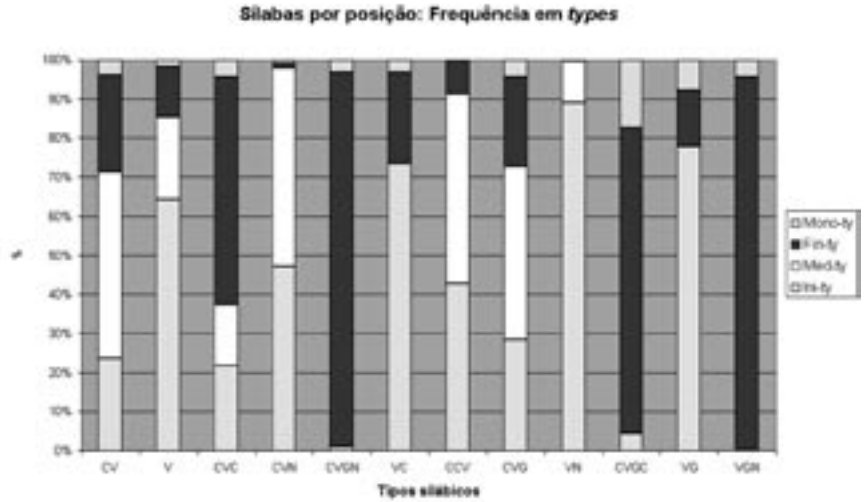

Figura 7: Distribuição dos tipos silábicos mais frequentes em função da posição na palavra (contagem sobre *types*).

Estes resultados confirmam e ampliam os resultados de Frota *et al.* (2005) e Freitas *et al.* (2006), já que a frequência dos tipos silábicos computada sobre *tokens* em posições prosodicamente proeminentes (acentuada e inicial) se correlaciona com o padrão de emergência dos tipos silábicos na criança.

3.4. Classes de segmentos consonânticos

Analisemos agora a distribuição das frequências das grandes classes de segmentos consonânticos: oclusivas, fricativas e líquidas (cf. Figura 8).

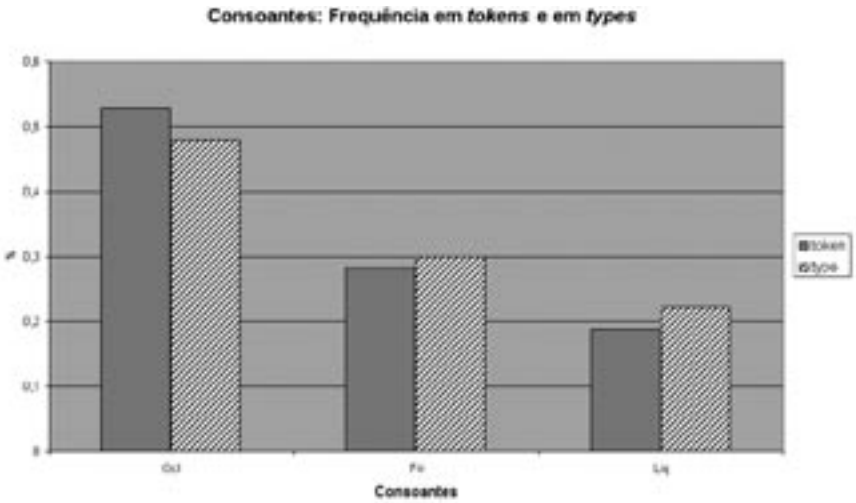

Figura 8: Distribuição da frequência das grandes classes de segmentos consonânticos: oclusivas, fricativas e líquidas (contagem sobre *tokens* e *types*).

Neste caso, as frequências computadas sobre *tokens* e *types* apresentam a mesma distribuição. Na linha do que tem vido a ser demonstrado, verificamos que também aqui a ordem de emergência das três classes predita pela frequência no *input* (quer computada sobre *tokens* quer sobre *types*) é a verificada na fala da criança, uma vez que a emergência das oclusivas precede a das fricativas e das líquidas (em ataque não-ramificado) (cf. Freitas, 1997).

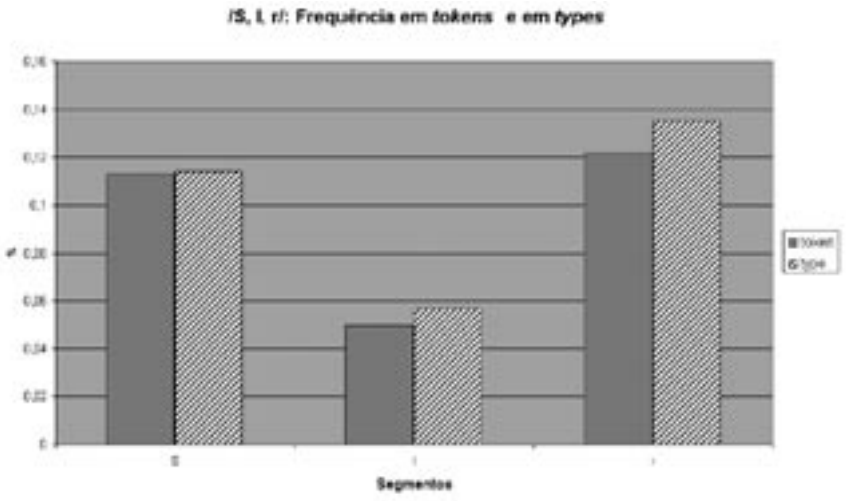

Figura 9. Distribuição da frequência da fricativa, líquida lateral e líquida vibrante (contagem sobre *tokens* e *types*).

Consideremos agora em particular as três consoantes que podem aparecer em coda (Figura 9)[2]. A consoante mais frequente é a líquida vibrante (que é, aliás, a consoante mais frequente no PE), seguida da fricativa surda (a segunda mais frequente na língua) e por fim da lateral, sem diferenças entre a contagem em *tokens* ou *types*. Esta distribuição geral da frequência das três consoantes não se correlaciona nem com a sua ordem de emergência em ataque, nem em coda: tanto em ataque como em coda, a fricativa precede as duas líquidas (Freitas, 1997; Jordão, 2009).

Se atendermos à distribuição relativa das três consoantes na sílaba (Figuras 10 e 11), a líquida lateral ocorre predominantemente em ataque, ao contrário da vibrante que se distribui pelas posições de ataque e coda e da fricativa que surge em coda em mais de 90% dos casos (sem diferenças entre *tokens* e *types*). Assim, a distribuição na sílaba também não se correlaciona com a ordem de emergência em ataque, mas prediz a ordem de emergência em coda (fricativa > líquidas).

[2] De modo a podermos comparar a ocorrência da fricativa em coda com a sua frequência geral, consideramos aqui apenas a ocorrência da pré-palatal surda, que é claramente mais frequente do que a sonora. Os valores de cada segmento em função da posição na sílaba são os seguintes: *onset* S=2983; Z=5519; *coda* S=46764; Z=1208.

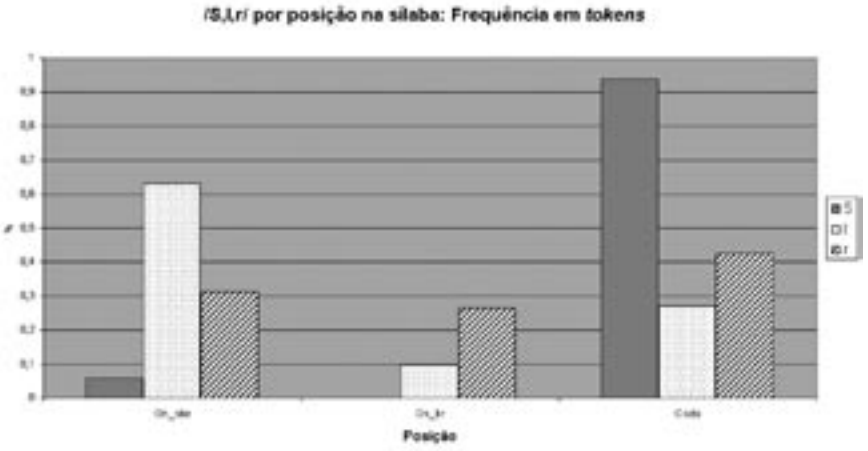

Figura 10. Distribuição da fricativa, líquida lateral e líquida vibrante por posição na sílaba: contagem sobre *tokens*.

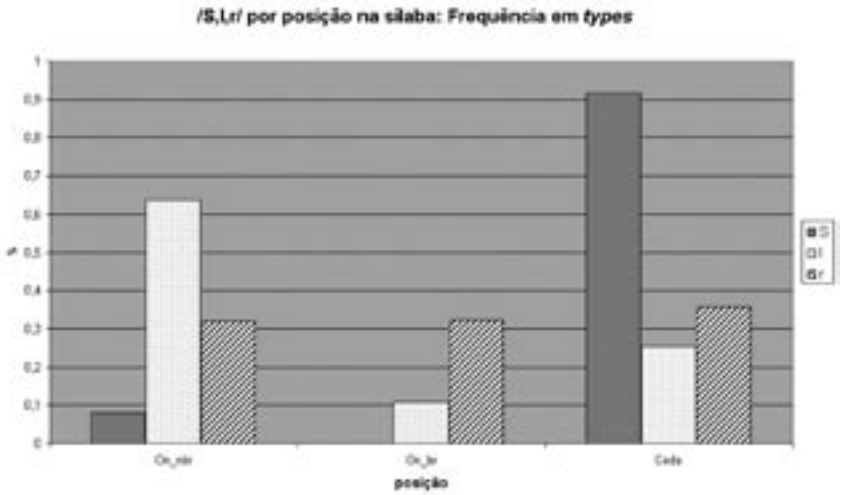

Figura 11. Distribuição da fricativa, líquida lateral e líquida vibrante por posição na sílaba: contagem sobre *types*.

Estes resultados apontam para uma interacção entre estrutura (ataque vs. coda), tipo de segmento (fricativa vs. líquidas) e frequência, a explorar em trabalhos futuros.

3.5. Traços de ponto de articulação consonântico

Finalmente, consideremos os dados de frequência de ponto de articulação consonântico (L=Labial; C=Coronal; D=Dorsal) em ataque silábico no *input*, contados sobre *tokens* e sobre *types* (Figura 12). Tal como sucedeu no caso das grandes classes de segmentos, as frequências obtidas sobre *tokens* e sobre *types* apresentam a mesma distribuição: C > L > D (sendo L e D muito próximos).

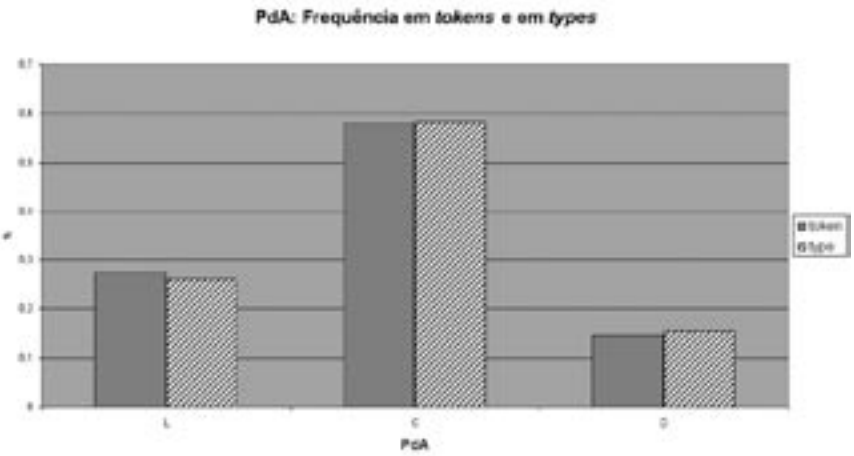

Figura 12: Distribuição da frequência dos pontos de articulação consonânticos: L=Labial; C=Coronal; D=Dorsal (contagem sobre *tokens* e *types*).

A frequência no *input* não se correlaciona directamente com a ordem de emergência dos pontos de articulação. Efectivamente, de acordo com Costa *et al.* (2007), no discurso da criança, no estádio em que a posição inicial é autonomizada, são possíveis nessa posição tanto segmentos *Labiais* como *Dorsais* (em contraste com o que sucede em línguas como o Holandês, onde *Dorsal* não é permitido em posição inicial, facto que tem sido atribuído à distribuição no *input* nesta língua). Nos dados gerais do *input* (Figura 12), *Dorsal* é o traço menos frequente pelo que a sua frequência geral não parece poder explicar a sua emergência precoce em posição inicial no PE.

Observemos agora os dados do *input* em função da posição na palavra e do acento (Figuras 13 e 14).

Observam-se diferentes distribuições para L e D vs. C segundo a posição, bem como diferenças entre contagens sobre *tokens* e sobre *types*. L e D destacam-se em posição inicial na frequência em *tokens* (ocorrem maioritariamente nesta posição, contrariamente a C) e D destaca-se nos monossílabos, quando considerados apenas os *tokens* (Figura 13). Há também diferentes distribuições para L e D, contrapondo-se a C, quando tida em conta a presença/ausência de acento e a posição na palavra, bem como grandes diferenças entre contagens sobre *tokens* e sobre *types*. Efectivamente, L e D destacam-se claramente

em posição acentuada inicial, mas apenas nas contagens sobre *tokens*. Em suma, L e D ocorrem em posições proeminentes – início de palavra e sílaba acentuada em início de palavra –, ao contrário de C.

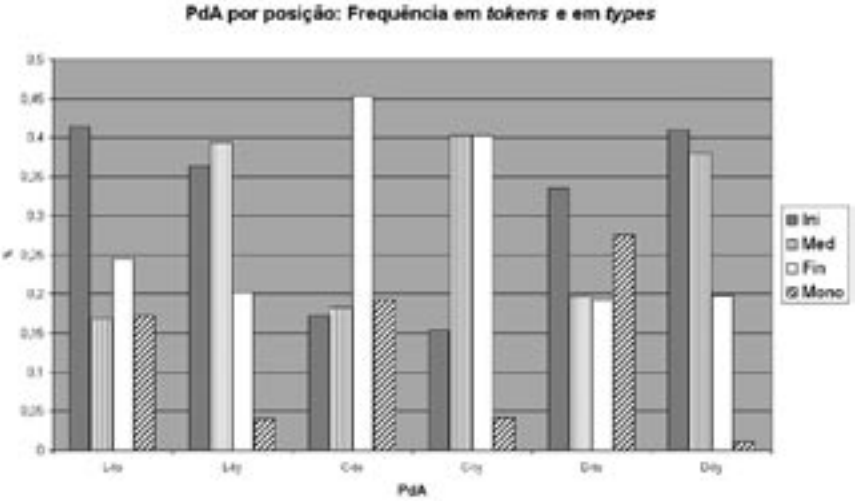

Figura 13: Distribuição da frequência dos pontos de articulação consonânticos (L, C, D) por posição (Inicial, Medial, Final e Monossílabos): contagem sobre *tokens* e *types*.

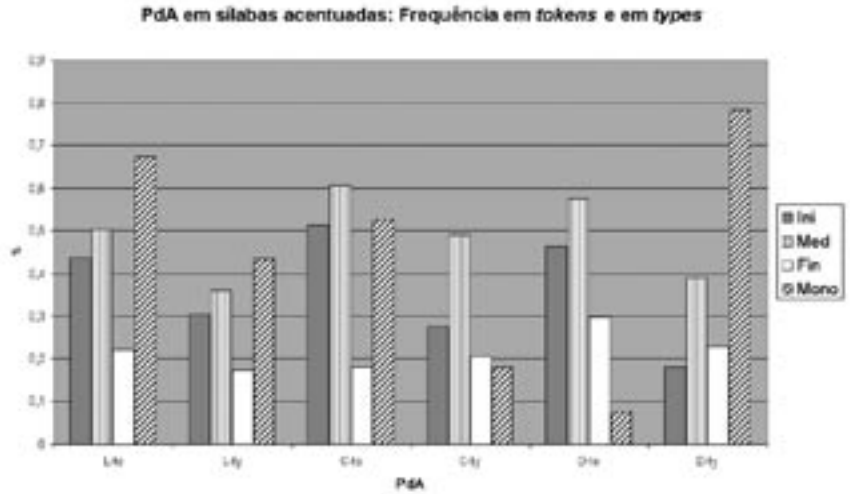

Figura 14: Distribuição da frequência dos pontos de articulação consonânticos (L, C, D) em sílabas acentuadas, por posição: contagem sobre *tokens* e *types*.

Face aos resultados obtidos, a distribuição no *input* (*tokens* apenas) prediz a disponibilização de L e D em posição inicial (reforçando e ampliando os resultados parcelares do *input* reportados em Costa *et al.* 2007).

4. Discussão e conclusões

No presente estudo investigámos o tipo de dados de frequência que é relevante para a aquisição de diferentes aspectos da fonologia, designadamente a frequência considerada em termos de *types* ou em termos de *tokens*. Verificou-se que a frequência no *input* constitui um bom preditor dos dados do discurso da criança para os formatos de palavra (emergência e frequência na produção), o padrão acentual (emergência e evolução), os tipos silábicos (emergência), as grandes classes de segmentos (emergência), os segmentos em coda (emergência) e a aquisição do ponto de articulação consonântico. Sempre que *types* e *tokens* diferem, fazendo predições diferentes, o preditor é a frequência considerada em termos de *tokens*. Isto mesmo foi demonstrado para os formatos de palavra, o acento, os tipos silábicos e o ponto de articulação (para classes de segmentos e segmentos em coda, *types* e *tokens* coincidem). Estes resultados apontam fortemente para que seja a frequência computada sobre *tokens* a relevante para a aquisição da fonologia.

Verificou-se igualmente que, para certos aspectos da fonologia, os dados de frequência relevantes são computados em *tokens* e tendo em conta a estrutura da língua, isto é as posições prosodicamente proeminentes. É esse o caso dos tipos silábicos e do ponto de articulação consonântico. Julgamos ser assim porque as posições prosodicamente proeminentes se salientam na percepção e favorecem a apreensão dos elementos (unidades ou padrões) que nelas ocorrem com frequência.

O papel da frequência em *tokens* tem várias implicações. Acentua a importância da distribuição de unidades e padrões *efectivamente presentes* no *input*, que varia de língua para língua, e portanto o efeito da especificidade da língua desde os momentos iniciais do processo de aquisição. Acentua também a relevância do *uso* da língua e do estudo dos possíveis *diferentes inputs* numa mesma língua a que diferentes crianças possam estar expostas, sugerindo que estes possam ser despoletadores de diferentes percursos na aquisição, com as consequentes implicações metodológicas que esta importância de um *input* variável traz para os estudos nesta área.

Finalmente, a relevância da frequência computada sobre *tokens* vai no sentido de propostas recentes sobre a representação do conhecimento fonológico através de *exemplar-based models* e sobre o processo de aquisição da linguagem a partir do uso ou *usage-based accounts of language learning* (e.g. Pierrehumbert 2001; Bybee 2001; Bybee & McClelland, 2005), na sua versão moderada dada a interacção verificada entre frequência e estrutura da língua. Em trabalho futuro, estas interacções serão exploradas, designadamente no domínio da estrutura silábica e da emergência dos segmentos. Encontra-se igualmente em preparação uma análise sistemática da frequência em *types* e

tokens para as várias unidades e padrões da fonologia do PE, em *corpora* diversificados (com recurso à *FrePOP* – Frota, Vigário, Martins & Cruz em curso), que permitirá estabelecer um conjunto de predições sobre os padrões a esperar na aquisição da língua, com base no efeito de frequência.

Referências

Bacelar do Nascimento, Fernanda, M. Lúcia Marques & M. Luísa Segura (1987) *Português Fundamental: Métodos e Documentos Inquérito de Frequência.* Instituto Nacional de Investigação Científica - Centro de Linguística da Universidade de Lisboa. Lisboa.

Bybee, Joan (2001) *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. & J. McClelland (2005) Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22, pp. 381-410.

Corpus de Referência do Português Contemporâneo, Centro de Linguística da Universidade de Lisboa.

CORP-ORAL, Instituto de Linguística Teórica e Computacional.

Correia, S. (2007) *Acoustic correlates of stress in early disyllabic productions of 2 Portuguese children.* Comunicação apresentada no GALA 2007, UAB, Barcelona.

Correia, Susana (2009) *The Acquisition of Primary Word Stress in European Portuguese*. Dissertação de Doutoramento, Universidade de Lisboa.

Costa, T., M. J. Freitas, S. Frota, F. Martins & M. Vigário (2007) Sobre o PA na periferia esquerda da palavra. In M. Lobo & M. A. Coutinho (Orgs.) *Textos Seleccionados do XXI Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: APL, pp. 315-328.

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

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

Freitas, Maria João (1997) *A aquisição da estrutura silábica do Português Europeu.* Dissertação de Doutoramento, Universidade de Lisboa.

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

Frota, Sónia, Maria João Freitas, Marina Vigário & Fernando Martins (2005) Prosody and frequency effects on the development of syllable structure in European Portuguese. Comunicação apresentada no *Symposium Exploring the Effects of Prosody, Mor-*

phology, Frequency and Representation on the Development of Syllable Structure in Romance Languages, Xth International Congress for the Study of Child Language, Berlim, Julho.

Frota, S. & M. Vigário (2008) The intonation of one-word and first two-word utterances in European Portuguese. *XIth International Congress for the Study of Child Language (Symposium 'Acquisition of intonation: interfaces with word stress and grammar. Cross-linguistic evidence')* Edimburgo, Julho.

Frota, S., M. Vigário, F. Martins, M. Cruz (em curso) *FrePOP – Frequency of Phonological Objects in Portuguese*. Laboratório de Fonética da Faculdade de Letras de Lisboa.

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

Ingram, David. 1988. The Acquisition of Word-Initial [v]. *Language and Speech*. 31(1), pp. 77-85.

Jordão, Raquel (2009) *A Estrutura Prosódica e a Emergência de Segmentos em Coda no PE. Um Estudo de Caso*. Dissertação de Mestrado, Universidade de Lisboa.

Leung, Man-Tak & Sam-po Law (2004) Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, & Computers* 36 (3), pp. 500–505.

Levelt, C. & R. van de Vijver (2004) Syllable Types in Cross-Linguistic- and Developmental Grammars. In R. Kager, J. Pater & W. Zonneveld (eds.) *Fixing Priorities: Constraints in Phonological Acquisition*. Cmabridge: Cambridge University Press, pp. 204-218.

Plunkett, K. & Marchman, V. (1989) Pattern association in a back-propagation network: implications for child language acquisition. Technical Report 89-02. Center for Research in Language. University of California, San Diego, La Jolla, CA.

Martins, F., M. Vigário & S. Frota (2009) *FreP – Frequency in Portuguese* v.2. *Software* de contagem de frequência (registo IGAC nº 209/2010).

Monnin, J., Lœvenbruck, H., & Beckman, M. E. (2007) The influence of frequency on word-initial obstruent acquisition in Hexagonal French. *Proceedings of the XVIth International Congress of Phonetic Sciences*, 6-10 August 2007, Saarbruecken, 1569-1572.

Ota, Mits (2006) Input Frequency and Word Truncation in Child Japanese: Structural and Lexical Effects. *Language and Speech*, 49(2), pp. 262-295 (Special Issue on *Crosslinguistic Perspectives on the Development of Prosodic Words*, guest edited by Katherine Demuth).

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

Português Falado. Documentos Autênticos (CD1, Anos 90). CD-ROM, Centro de Linguística da Universidade de Lisboa e Instituto Camões.

Prieto, Pilar (2006) The Relevance of Metrical Information in Early Prosodic Word Acquisition: A Comparison of Catalan and Spanish. *Language & Speech* 49 (2), pp. 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 & Luís Caldas de Oliveira (1996) Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu. In 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, Fernando Martins & Sónia Frota (2005) Frequência no Português Europeu: A ferramenta Frep. In Duarte & I. Leiria (orgs.) *Actas do XX Encontro da Associação Portuguesa de Linguística.* Lisboa: APL/Colibri, pp. 897-908.

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

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