Para uma caracterização da distinção entre palavras prosódicas e clíticos com base em dados de frequência[*]

Marina Vigário, Sónia Frota & Fernando Martins
Universidade de Lisboa

The frequency of various phonological patterns in prosodic words (PW) and phonological clitics (CL) was systematically inspected in a corpus of European Portuguese containing ca. half a million words. Clear differences were found between the two types of words, in the use of the (stressless) segmental inventory of the language, in the frequency of individual segments, in syllable types, in the (proportion of) allowed word shapes, and in the tokens/type distribution, a.o.. Given the known ability of infants to process statistical patterns in language input, it is proposed that the frequency of phonological patterns should be added to the set of aspects that may play a role in the acquisition of the distinction between PWs and CLs, which in turn can be seen as a precursor of the distinction between lexical and function words.

Key words: Prosodic words, clitics, frequency, function words, lexical words

0. Introdução

São vários os aspectos de natureza gramatical, fonético-fonológicos, morfossintácticos e semânticos, que distinguem as palavras prosódicas (PWs) dos clíticos fonológicos (CLs). Sabe-se, para além disso, que existe uma estreita relação entre os pares PW/palavra lexical e CL/palavra gramatical. Efectivamente, embora não sejam termos coincidentes, frequentemente vão de par. Pela relação muito próxima estabelecida entre PW/palavra lexical e CL/palavra gramatical, uma das áreas em que a diferenciação entre PWs e CLs pode ser explorada é a da aquisição da linguagem. Na realidade, PW e CL são categorias com correlatos fonético-fonológicos aos quais a criança pode ser sensível ainda antes de possuir conhecimento gramatical/semântico implicado nos conceitos de palavra lexical e palavra gramatical.

O presente artigo pretende explorar uma dimensão da distinção sonora entre PWs e CLs que, até onde sabemos, não foi ainda investigada anteriormente e surge na esteira de uma vasta literatura que se tem debruçado, nas três últimas décadas sobre a questão do papel da informação estatística e distribucional presente no contínuo sonoro no processo de aquisição do léxico, da fonologia e da sintaxe. Em particular, no presente trabalho investigamos diferenças entre os dois tipos de palavras no que se refere à frequência de ocorrência de unidades e padrões fonológicos no Português Europeu.

O artigo encontra-se estruturado da seguinte forma. Na secção 1 fazemos uma breve revisão do que se sabe sobre a diferença entre PW e CL e sobre a relação que estas categorias estabelecem com as categorias palavra lexical e palavra gramatical;

[*] Esta investigação foi parcialmente financiada pelo projecto PTDC/LIN/70367/2006.

fazemos também aqui uma breve referência a literatura recente que se tem ocupado do papel da informação estatística e distribucional e da informação fonético-fonológica na aquisição da gramática. Na secção seguinte apresentamos os procedimentos metodológicos. Os resultados da investigação são apresentdos na secção 3 e discutidos na secção 4. Finaliza-se o artigo na secção 5, sintetizando o essencial dos resultados obtidos e levantando-se algumas questões para investigação futura.

1. Enquadramento

Se bem que PW e CL sejam noções fonológicas, estas duas categorias distinguem-se relativamente a diversos outros parâmetros gramaticais.

Do ponto de vista fonológico, no Português, tal como em muitas outras línguas, as PWs são portadoras de acento de palavra, contrariamente aos clíticos fonológicos (note-se que, para além da percepção, também a aplicação ou não de regras sensíveis ao acento evidencia a presença ou a ausência de acento, como a redução vocálica ou a semivocalização de vogais); existem também regras fonológicas que se aplicam no interior ou nos limites de PW (e.g., a queda de vogal alta não-recuada no limite direito de PW não afecta CLs, a não ser que estes incorporem na PW precedente, situação em que se encontram no limite direito de PW); existem restrições fonotácticas sobre o início de PW, que não afectam CLs (veja-se a possibilidade de a lateral palatal iniciar CL mas não PW); e a atribuição de acentos tonais, assim como a possibilidade de focalização estão também circunscritos a PW (Vigário 2003). Também ao nível do tamanho máximo possível PWs e CLs se distinguem: as PWs podem exceder as duas sílabas, enquanto as palavras clíticas são maximamente dissilábicas, tanto quanto se sabe (efectivamente, clíticos dissilábicos estão documentados em várias línguas, mas polissilábicos não – veja-se também sobre este assunto a revisão feita em Vigário 2003: cap. 5).

Outras diferenças conhecidas entre PWs e CLs dizem respeito às características morfossintácticas de ambos os tipos de palavras. Tipicamente, as PWs são formadas por palavras pertencentes às classes morfossintácticas abertas (nomes, adjectivos, verbos, advérbios), com conteúdo semântico referencial (embora também possam integrar palavras com função gramatical, pertencentes a classes fechadas); enquanto os clíticos fonológicos, pelo contrário, são necessariamente palavras gramaticais. Para além disso, PW, mas não CLs, podem ser morfologicamente complexas. Finalmente, do ponto de vista da sua distribuição, as palavras clíticas não podem ocorrer isoladamente, enquanto as palavras que formam PWs podem (e.g., Selkirk 1995, entre muitos outros).

A presente investigação visa fornecer novos dados para a caracterização da distinção entre PWs e CLs, relativos à frequência de ocorrência de unidades e padrões fonológicos nas duas classes de palavras no Português Europeu.

São vários os trabalhos que têm pesquisado o papel da informação fonológica ou prosódica na aquisição de aspectos vários da gramática e da segmentação lexical (e.g., Morgan & Demuth 1996, Nespor, Guasti, Christophe 1996, Gout, Christophe & Morgan 2004; veja-se ainda a revisão feita em Gervain & Mehler 2010). É sabido que

emergência de palavras funcionais é mais tardia do que a emergência de palavras lexicais. Guasti (2002: 107) refere que todos os elementos funcionais estão normalmente ausentes nas primeiras produções infantis o que faz com que a fala nesta fase se assemelhe a discurso telegráfico. Independentemente de eventuais razões não-fonológicas para esta diferença, o facto é que, nas primeiras produções, a criança mostra já ter estabelecido uma distinção entre estas duas categorias, ao tratá-las diferentemente (note-se que a omissão de palavras funcionais nesta fase não afecta apenas palavras funcionais clíticas).

Christophe, Millotte, Bernal & Lidz (2008) mostram experimentalmente que as crianças aos dois anos usam o conhecimento que têm de palavras funcionais (como artigos ou pronomes pessoais) na sua língua como pista para a categorização morfossintáctica de palavras lexicais desconhecidas (das categorias nome ou verbo) e para colocarem hipóteses sobre o seu significado possível (designando objectos ou acções). Estes resultados sugerem fortemente que a identificação do que é palavra funcional precede largamente a produção deste tipo de palavras.

Sabe-se que as crianças numa fase precoce são sensíveis à organização prosódica, designadamente ao nível do sintagma fonológico e do sintagma entoacional (e.g., Gout, Christophe & Morgan 2004), e que são sensíveis a informação estatística e distribucional (e.g., Jusczyk, Luce & Charles-Luce 1994; Saffran 2002). Para além disso, desde há muito que se compreendeu que a proeminência ao nível do sintagma fonológico fornece uma pista robusta acerca do parâmetro de direccionalidade sintáctica, uma vez que a proeminência neste nível é sistematicamente inicial em línguas com recursividade à esquerda (como no Japonês ou no Turco) e final em línguas com recursividade à direita (como o Português ou o Francês) (e.g., Nespor & Vogel 1986/2007, entre muitos outros). Este facto deu origem a uma hipótese segundo a qual as crianças em fases muito precoces, pré-verbais, podem fixar o parâmetro sintáctico básico que regula a ordem cabeça-complemento na língua particular a que estão expostas com base naquela pista prosódica (Nespor, Guasti, Christophe 1996; Christophe, Nespor, Guasti & van Ooyen 2003). Note-se que, face aos algoritmos de mapeamento sintaxe-prosódia, a proeminência ao nível do sintagma fonológico identifica catagorias lexicais, estando o material funcional no lado não-recursivo da língua (numa língua como o Japonês, à direita da cabeça e numa língua com o Português, à sua esquerda). Face ao que é conhecido sobre a sensibilidade dos bebés a padrões distribucionais e estatísticos, à proeminência e ao fraseamento prosódico, é plausível colocar a hipótese de que as palavras funcionais podem ser identificadas no contínuo por corresponderem a sílabas muito frequentes que ocorrem tipicamente nas extremidades de constituintes prosódicos (Christophe et al. 2008).

A hipótese da presente investigação é a de que existem outros factores de natureza estatística/distribucional que distinguem PWs e CLs e, dada a correlação estreita entre os pares PW/palavra lexical e CL/palavra funcional, tal informação poderá contribuir para as crianças estabelecerem, mesmo numa fase pré-lexical, uma categorização rudimentar das palavras lexicais e das palavras gramaticais.

2. Metodologia

Com vista à obtenção sistemática de dados de frequência referentes à ocorrência de unidades e padrões fonético-fonológicos capazes de distinguir PWs e CLs, usámos a ferramenta electrónica FreP (Martins, Vigário & Frota, 2009, v. 2.1.09). Esta ferramenta foi concebida para identificar e contar unidades e padrões fonológicos em textos escritos e permite fazer um tratamento diferenciado de PWs e CLs, bem como de *tokens* (listagem de items lexicais incluindo ocorrências múltiplas da mesma palavra) e *types* (listagem de palavras únicas). Sobre o funcionamento geral da ferramenta veja-se Vigário, Frota & Martins (2006). Trabalhos subsequentes desta equipa e colaboradores, detalhes actualizados sobre as funcionalidades da ferramenta e uma demostração do seu funcionamento podem ser encontrados em www.fl.ul.pt/LaboratorioFonetica/FreP.

Os dados de frequência foram computados sobre um corpus de fala com mais de meio milhão de palavras (mais precisamente, 503 948). Trata-se de uma secção da base de dados FrePOP (Frota, Vigário, Martins, & Cruz 2010), presentemente disponível em http://frepop.fl.ul.pt/. Este é um corpus de discurso espontâneo, que contém material proveniente do *Português Falado* anos 90 (CLUL), do corpus CRPC (CLUL), do corpus CORP-ORAL (ILTEC), do corpus C-ORAL-ROM (CLUL), do corpus CORDIAL-SIN (CLUL) e um excerto de um debate televisivo.

Utilizando o FreP, foram extraídos valores de frequência relativos à ocorrência de PW e CL, e sua extensão em número de sílabas, número e diversidade de tipos silábicos, distribuição das classes maiores de segmentos e dos segmentos concretos. Fazendo uso de uma funcionalidade que permite gerar ficheiros .txt que incluem apenas PW e apenas CL, correu-se a ferramenta separadamente sobre os ficheiros assim gerados contendo, de tal modo que os dados de frequência obtidos em cada caso pudessem ser referentes apenas a cada uma das categorias de palavras. Para PWs e para CLs, e sempre com o recurso ao FreP, procedeu-se ainda à criação de um ficheiro contendo apenas as ocorrências únicas de cada palavra (*types*), o que permitiu correr de novo a ferramenta e obter dados de frequência para PWs e CLs referentes a *types*. Os números totais de unidades maiores extraídas encontram-se sintetizados no Quadro 1.

	PW	CL
Tokens	353 665	150 283
Types	31 591	54
Sílabas	765 102	160 552
Segmentos	1 641 098	283 354

Quadro 1: Números totais de palavras (*tokens*), de palavras únicas (*types*), de sílabas e de segmentos no corpus analisado.

A observação dos dados visou atingir dois objectivos fundamentais. O objectivo central foi o de identificar as diferenças mais significativas entre PWs e CLs relativamente à frequência de unidades e parâmetros fonético-fonológicos. Para além disso, procurou-se também elementos do mesmo tipo que permitam distinguir CLs de

sílabas iniciais ou finais de PW, uma vez que tais pistas, para além da elevada frequência das sílabas das palavras funcionais (Christophe et al. 2008), poderão ser usadas pelas crianças para a segmentação dos CLs. Um objectivo secundário final foi ainda traçado. Sendo os CLs formas que dependem prosodicamente da PW adjacente, eles formam com estas unidades PW estendidas. Porque o impacto da cliticização no formato de PW (estendida) nunca foi anteriormente avaliado, procedemos à avaliação desse impacto ao nível da frequência dos diferentes padrões acentuais.

Importa precisar, para terminar esta secção, que para a computação da PW estendida usou-se uma nova funcionalidade do FreP que permite determinar o tamanho das PW tendo em conta as palavras que se cliticizam neste domínio. Assume-se que os clíticos no Português Europeu se incorporam ou adjungem a PW, dados os argumentos avançados nesse sentido em Vigário (2003).

3. Resultados

Sabe-se a partir de trabalhos anteriores que a frequência de CLs (*token*) é inferior à das PWs. No corpus analisado em Vigário, Freitas & Frota (2006) e em Vigário, Frota & Martins (2010) as PWs constituem 70.4% e 70.5% do total das palavras do corpus analisado, contra, respectivamente, 29.6% e 29.5% de CLs. Esta proporção parece ser muito robusta, uma vez que, também no nosso corpus, aproximadamente 20 vezes maior do que o avaliado em Vigário et al. (2006) e duas vezes maior do que o corpus usado em Vigário et al. (2010), esta proporção se mantém: as PWs correspondem aqui a 70.2% e os CLs a 29.8%.[1] Contudo, os valores são muitíssimo mais expressivos quando avaliada a proporção de PWs e CL tendo em conta as palavras únicas (*types*): a percentagem de PWs-*types* no corpus é de 99.83%, enquanto a de CLs-*types* é de apenas 0.17%. Isto significa que a proporção *token*/*type* para cada classe é muito diversa. Efectivamente, a média de ocorrência de uma mesma PW é de 11.2, enquanto a de um mesmo CL é de 2783.0. Por outras palavras, as palavras clíticas repetem-se com muitíssimo mais frequência do que as palavras prosódicas.[2]

Como vimos na secção introdutória, o tamanho de palavra distingue de modo evidente PWs e CLs, uma vez que, enquanto as PWs podem ser polissilábicas, os CLs são maximamente dissilábicos. Contudo, há um outro aspecto distintivo referente aos formatos dos dois tipos de palavras. Como pode ver-se na Figura 1, a proporção de palavras monossilábicas e dissilábicas é muito distinta nas duas classes: a larga maioria dos CLs é monosilábica, enquanto o formato mais frequente de PWs é o dissílabo.

[1] Valores um pouco mais altos para palavras sem acento são referenciados em Viana et al. (1996). A diferença identificada pode dever-se a diferentes critérios para classificação das palavras como não-acentuadas. Nos trabalhos mais recentes de Vigário e colegas esta classificação é feita através da ferramenta electrónica FreP, que inclui um algoritmo de identificação de PWs e CLs baseado nos critérios fonético-fonológicos explanados em Vigário (2003).

[2] No mesmo sentido, observando apenas as 20 palavras mais frequentes do corpus, verificamos que 70% são CLs e apenas 30% correspondem a PWs. Este tipo de diferença merece em nosso entender uma avaliação autónoma e mais detalhada.

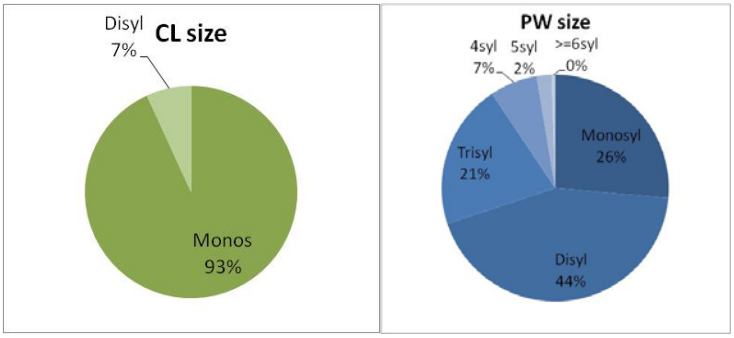


Figura 1: Percentagem de PWs e CLs com diferentes formatos (em número de sílabas).

Quando observada a diversidade de tipos silábicos em ambos os tipos de palavras, verifica-se que, quando comparados com PWs, os CLs apresentam um número significativamente baixo de formatos de sílabas (aqui entendidos como as diferentes combinatórias possíveis, no interior de uma sílaba, de consoantes, vogais e glides, orais e nasais). No nosso corpus, os CL apenas instanciam 8 tipos contra os mais de 60 encontrados em PWs.

Considerando agora os valores de frequência dos tipos silábicos de CL em confronto com as sílabas átonas encontradas nos limites de PW, verificamos que os 5 formatos mais frequentes nos CLs não diferem dos das sílabas átonas de PW: os tipos CV, V, CVC, VC, CVN constituem 96% dos CLs e 87.71% das sílabas átonas de PW. Para além disso, não distinguem também CLs das sílabas iniciais átonas de PW, onde ocorrem os mesmos 5 formatos mais frequentes, pela mesma ordem. Tal não sucede em relação às sílabas átonas em fim de PW, onde apenas 3 dos 5 tipos silábicos mais frequentes são comuns aos 5 mais frequentes nos CLs.

Também considerando a proporção do tipo silábico mais frequente em cada classe, CV, verifica-se grande proximidade nos valores obtidos para este tipo silábico nos CLs e na posição inicial de PW, contrariamente ao que sucede com a posição final de PW ou em qualquer posição de PW (considerando sempre o tipo silábico em posição átona e excluindo, aqui como noutros locais deste artigo, as PW monossilábicas, cujas sílabas são ao mesmo tempo iniciais e finais de PW). Como pode ver-se na Figura 2, no primeiro par de situações os valores de frequência estão pouco acima de 40% (44.1% e 41%, respectivamente) enquanto no segundo estão próximos ou acima dos 60% (66.7% e 59.2%, respectivamente).

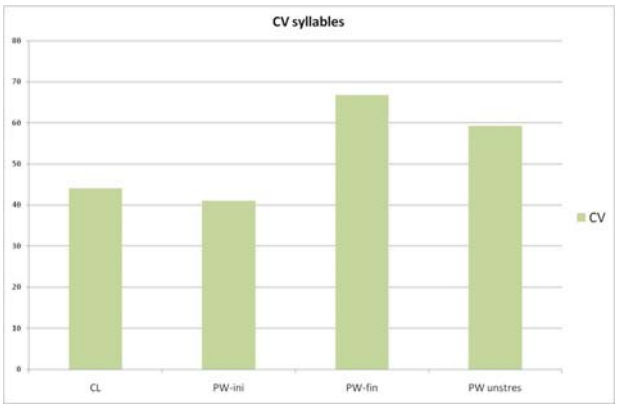


Figura 2: Distribuição do tipo silábico CV átono em posição inicial de CL (*CL*), posição inicial de PW (*PW-ini*), posição final de PW (*PW-fin*) e em qualquer posição de PW (*PW unstres*).

Pelo contrário, observando não apenas o que se passa com o segundo tipo silábico mais frequente nos CLs, o tipo V, mas considerando todos os tipos silábicos iniciados por vogal, verifica-se que os seus valores de frequência nos CL são bastante superiores aos verificados em qualquer outra posição de PW, embora a proximidade seja evidentemente maior, uma vez mais, com o início de PW (34.8% e 44%, respectivamente, para as sílabas iniciais de CLs *versus* 21.1% e 36.2%, respectivamente, para a posição inicial de PW).

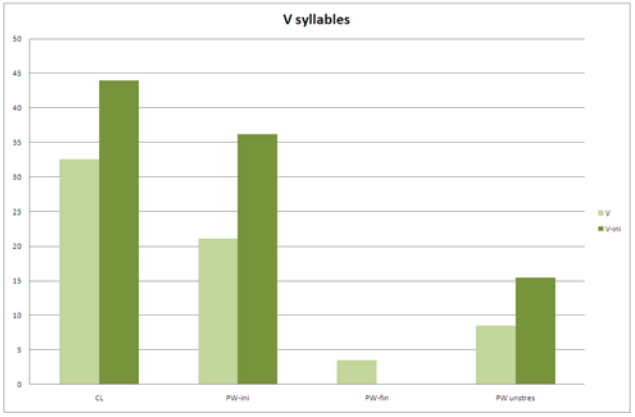


Figura 3: Distribuição dos tipos silábicos V e iniciados por V átonos em posição inicial de CL (*CL*), posição inicial de PW (*PW-ini*), posição final de PW (*PW-fin*) e em qualquer posição de PW (*PW unstres*).[3]

[3] Novamente foram incluídos os valores relativos às sílabas que iniciam CL; incluindo as restantes, a percentagem do tipo silábico V é de 32.5% e a dos tipos iniciados por V é de 41.2%.

Ainda considerando os formatos silábicos, uma inspecção dos três restantes tipos silábicos encontrados em CLs (VG, VGN, VGC) mostra que, embora eles sejam relativamente pouco frequentes nesta classe, eles são exclusivos ou quase exclusivos dos CL: confrontem-se os valores em início de CL apresentados no Quadro 2 com os relativos à distribuição desses tipos silábicos no interior de PW, nunca excedendo os 0.5%.

	CL	Pwini	Pwfin	PW unst
VG	1.18	0.4	0.001	0.14
VGN	2.58	---	0.49	0.19
VGC	0.27	0.004	---	0.001

Quadro 2: Distribuição dos 3 tipos silábicos menos frequentes nos CLs (dos 8 possíveis) em posição inicial de CL (*CL*), posição inicial de PW (*PW-ini*), posição final de PW (*PW-fin*) e em qualquer posição de PW (*PW unstres*), todos em posição átona; valores percentuais para os totais de sílabas iniciais de CL (150275), de sílabas átonas iniciais de PW (141455), de sílabas átonas finais de PW (204006), de sílabas átonas de PW (411430).

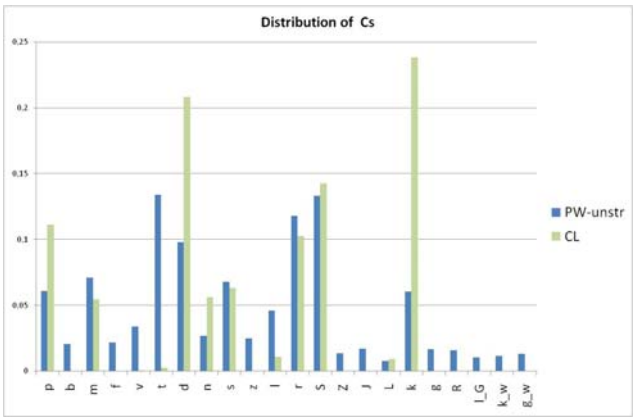
Grandes diferenças entre CLs e PWs são também encontradas quando observados os segmentos. Considerando primeiramente o inventário segmental, verificamos que ele se reduz em perto de 50% no caso dos CLs: apenas 21 dos 40 segmentos fonéticos da língua ocorrem associados a CL (ver Quadro 2).

Inventário segmental em posição átona
p, b, **m**, f, **v, t, d, n, s**, z, **l,** r, **S**, Z, J, **L, k**, g, R, l_G, k_w, g_w, **i,** e, E, **@, a, u, o**, O, i~, e~, **6~,** u~, **o~**, j, **w**, **j~**, w~

Quadro 3: Inventário segmental em posição átona (transcrição em SAMPA). Todos os segmentos ocorrem em PWs; os que não ocorrem em CLs são assinalados a cinza.

Para além disso, a distribuição da frequência dos segmentos admitidos nas duas classes também exibe fortes assimetrias: segmentos como [p], [d], [n], [k], [@], [o~] e [j~] são muito mais frequentes no interior da classe dos CLs e, inversamente, segmentos como [v], [t] e [l] ocorrem significativamente mais no interior da classe das PWs (ver Figura 4).

(a)



(b)

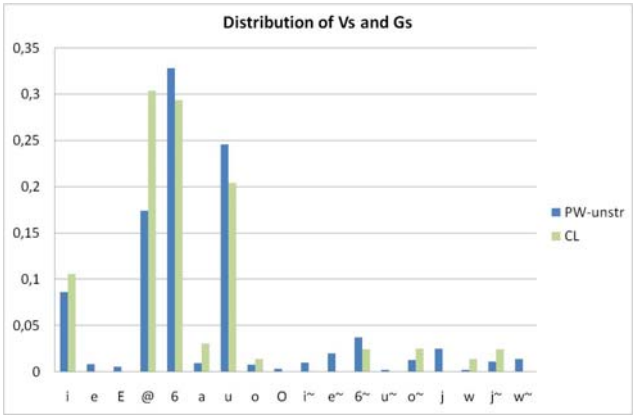


Figura 4: Distribuição dos segmentos consonânticos (a) e vocálicos e semivocálicos (b) em sílaba átona por PWs e CLs.

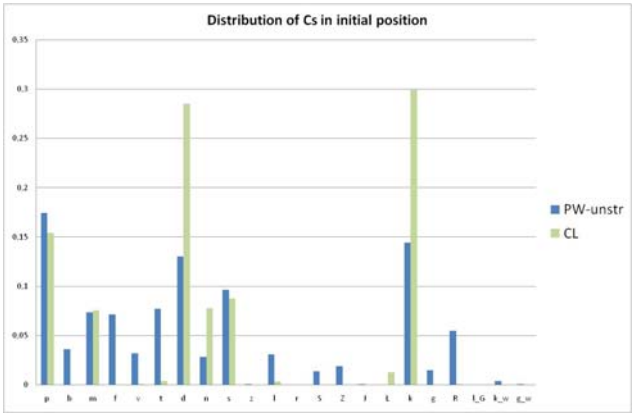
Considerando a distribuição dos segmentos em função da posição, percebe-se que o inventário segmental em início de CL reduz ainda mais um pouco (60%), o que, comparado com a posição inicial de PW é muito mais expressivo. Efectivamente, a redução do inventário segmental em início de PW é apenas de 22% (estão entre os segmentos que não ocorrem em posição inicial de PW [L, l_w, J, r], são raros [z, g_w] e é muito raro [o~]).

Inventário segmental em posição inicial átona
p, b, **m,** f, **v, t, d, n, s,** z, **l,** r, S, Z, J, **L, k,** g, R, l_G, k_w, g_w
i, e, E, **@, 6, a, u, o,** O, i~, e~, 6~, u~, o~, j, w, j~, w~

Quadro 4: Inventário segmental em posição átona inicial (transcrição em SAMPA). Todos os segmentos ocorrem em início de PW; os que não ocorrem em CLs são assinalados a cinza.

Quanto à distribuição dos segmentos em posição inicial, ela é também muito diferente em função das classes (ver Figura 5), o que já podia inferir-se parcialmente pelo que ficou dito acima acerca das sílabas em início de palavra: as consoantes [k], [d], [n] e [L] e as vogais [a], [u] e [6~] são muito mais frequentes em CL (ou exclusivas desta classe, no caso de [L]) do que nas PWs; inversamente, são vários os segmentos que ocorrem só ou mais predominantemente em início de PW ([b, f, v, t, z, l, S, Z, g, R, k_w, g_w, i, e, E, 6, o, O, i~,e~]. Na realidade, são poucos os casos de distribuição equilibrada dos segmentos pelas duas classes (que se restringem aos segmentos [p, m, s,]).

(a)



(b)

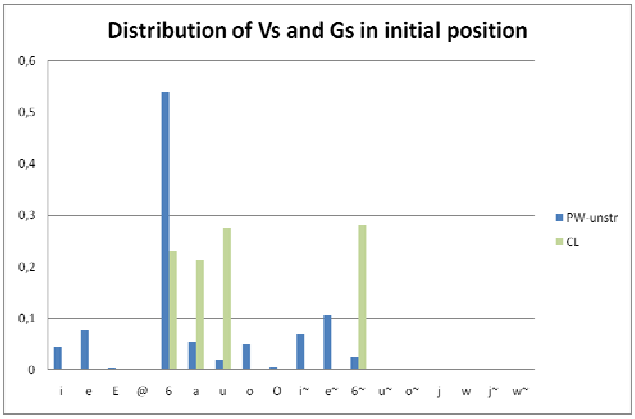


Figura 5: Distribuição dos segmentos consonânticos (a) e vocálicos e semivocálicos (b) em sílaba átona inicial por PWs e CLs.

Consideremos agora o inventário segmental em posição final de palavra. Em ambas as classes, o inventário de segmentos que podem terminar sílaba é muito reduzido, o que é previsível, uma vez que estamos a falar, no caso de consoantes, de uma posição silábica que licencia menos segmentos, no Português como noutras línguas. Apesar disso, há segmentos que só ocorrem, ou ocorrem fundamentalmente, no final de cada uma das classes: apenas 2 consoantes podem aparecer no final dos CLs ([S, r]); [l_G] não ocorre em CLs, contrariamente a [o~] e [w], que são relativamente frequentes nesta classe (cf. *com* e *ao*); para além de [a], não existem CLs com vogais abertas; para além das consoantes típicas de final de sílaba, [S, r, l_G], há consoantes excepcionais que podem ocorrer em final de PW (como [n]), e vogais excepcionais em posição átona, como [E, O]; [w] é muito raro no final de PW e [o~] não ocorre em posição átona final nesta classe.

A partir do que foi dito acima e da observação dos formatos concretos das sílabas é possível estabelecer um conjunto de generalizações ou fortes tendências para a ocorrência de certos segmentos ou sílabas em certas posições da palavra, e que podem funcionar como pistas para a distinção/segmentação de CLs e PWs. Entre elas destacam-se as seguintes (considerando sempre posições átonas):

(i) sílabas iniciada por [i] coincidem quase exclusivamente com início de palavra (PW ou CL) (99.8%) e maioritariamente com clítico (77.8%);
(ii) sílabas iniciadas por [u] coincidem quase sempre com início de palavra (95.2%) e quase sempre com início de CL (91.5%);
(iii) sílabas iniciadas por [aw] são praticamente exclusivas do início de palavra (99.9%) e são quase sempre também iniciais de CL (87.8%);
(iv) sílabas iniciadas por [k@] pertencem quase sempre a CLs (95.7%); e em 85% de sílaba iniciada por [k@] estão em início de CL;
(v) as palavras iniciadas pela sílaba [d@] são na sua larga maioria CLs - 75% destas sílabas iniciais pertencem a CLs, contra apenas 25% pertencentes a PW;
(vi) as palavras iniciadas por [6~j~] pertencem sempre à classe dos CL;
(vii) as palavras terminadas em [a], [o] e [o~] são sempre da classe CL;
(viii) as palavras terminadas em [w], [i] e [r] (sempre sílabas átonas) são praticamente exclusivamente da categoria CL (99.8%, 99.6% e 96.8%, respectivamente);
(ix) as palavras iniciadas por [b, f, z, Z, g, k_w, g_w, e, E, O, i, e~, u~] pertencem sempre à categoria PW;
(x) as palavras terminadas em [l_G, n,] são sempre da categoria PW.

Para finalizar esta análise da frequência das unidades e padrões fonético-fonológicos envolvendo PWs e CLs, gostaríamos de proceder à avaliação do impacto da cliticização sobre os padrões acentuais de PW.

Como a Figura 6 revela, a adição de clíticos reduz o número de palavras monossilábicas e faz aumentar o número de palavras agudas. Isto deve-se à presença de proclíticos ligados a PWs monossilábicas. Apesar de pouco mais de 50% das palavras apresentarem o padrão acentual grave, a fatia complementar divide-se entre palavras

agudas e palavras monossilábicas, não emergindo um único padrão acentual como sendo esmagadoramente dominante na língua.

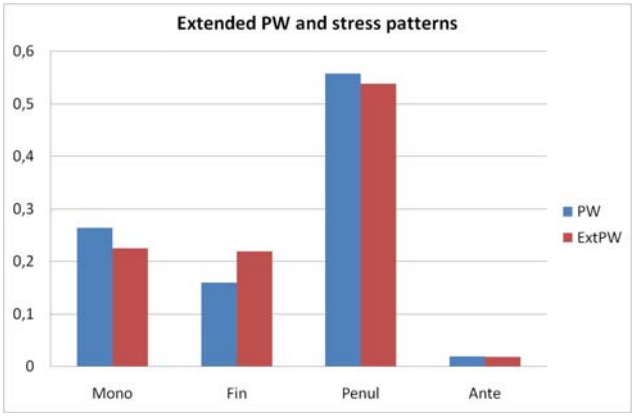


Figura 6: Distribuição dos padrões acentuais considerando PW e PW estendida (i.e., incluindo CLs incorporados ou adjuntos a PW) – proporção de palavras monossilábicas (*Mono*), oxítonas (*Fin*), paróxitonas (*Penul*) e proparóxitonas (*Ante*).

4. Discussão

Os resultados aqui apresentados mostram que vários aspectos relativos à frequência e distribuição de unidades e padrões fonético-fonológicos distinguem PWs e CLs. PWs distinguem-se muito evidentemente dos CLs pela maior diversidade lexical, maior diversidade de formatos ou tamanhos de palavra, maior diversidade de tipos silábicos, maior diversidade de segmentos, sendo também bastante mais frequentes do que os CL (ocorrendo mais de 2 vezes mais do que estes no corpus). Para além disso, as duas classes distinguem-se também em relação às frequências de muitas das unidades fonético-fonológicas: CLs monossilábicos predominam claramente sobre os dissilábicos, verificando-se o inverso com as PWs; 3 dos 8 tipos silábicos que ocorrem em CLs são residuais em PWs; vários segmentos que podem ocorrer em sílabas átonas de PWs e CLs apresentam proporções muito diversas nas duas classes ([p], [d], [n], [k], [@], [o~] e [j~] são muito mais frequentes nos CLs e [v], [t] e [l] são muito mais frequentes em PWs).

Se bem que alguns dados pareçam mostrar que existe maior proximidade entre os valores de frequência observados nos CLs e na posição inicial de PW, especialmente no que se refere aos valores de frequência dos tipos silábicos, a verdade é que são muitos os casos em que os CLs apresentam comportamento diferenciado quer da posição inicial, quer da posição final de PW: os tipos silábicos começados por V são claramente mais frequentes nos CL, três do 8 tipos silábicos encontrados na classe das palavras clíticas são residuais tanto em posição inicial como em posição final de PW; os segmentos [k, d, n, L, a, u, 6~] são muito mais frequentes em CLs, ou ocorrem só nesta classe; [b, f, v, t, z, l, S, Z, g, R, k_w, g_w, i, e, E, 6, o, O, i~,e~] ocorrem só ou

predominam na classe das PWs, sendo poucos os casos de distribuição equilibrada dos segmentos pelas duas classes (correspondendo apenas a 10% da totalidade dos segmentos que podem oculpar a posição incial de palavra); há segmentos e/ou sílabas concretas que predominam muito claramente em certas posições, coincidentes com os limites direito ou esquerdo de PW ou de CL (e.g., sílabas iniciadas por [i, u, aw, k@] pertencem quase sempre a CLs; palavras iniciadas por [d@, 6~j~] são quase sempre ou sempre CL; sílabas átonas finais terminadas por [a, o, o~, w, i, r] pertencem exclusivamente ou quase exclusivamente a CLs; palavras iniciadas por [b, f, z, Z, g, k_w, g_w, e, E, O, i, e~, u~], e palavras terminadas por [l_G, n] são sempre PW). Todos estes aspectos nos sugerem não apenas que PWs e CLs diferem de um modo muito sistemático relativamente a parâmetros de frequência, mas também que os dados de frequência podem ser usados para a segmentação de sílabas pertencentes a CLs. Esta é efectivamente uma questão relevante no quadro da aquisição da linguagem, uma vez que a criança apenas pode ser sensível às diferenças estatísticas entre PWs e CLs se ela puder distinguir uma sílaba pertencente a um CL de uma sílaba átona de PW. A nossa proposta aqui é a de que a diferença estatística aqui descrita pode contribuir para a segmentação, para além da categorização dos dois tipos de palavras em classes distintas.

Não deixa, porém, de ser um facto que enclíticos e proclíticos cliticizam a PW, acrescentando-a. Não é pois de descurar a possibilidade de a criança poder tratar, pelo menos para certos efeitos, a PW com os respectivos clíticos como uma única PW (estendida). Este tópico é relevante no quadro das discussões acerca do padrão mais frequente do acento de palavra no Português e sua relação com o comportamento infantil em relação à produção do acento, que revela dificuldades na aquisição do acento que não seriam esperadas num quadro em que o padrão acentual grave seja de facto esmagador (veja-se a este respeito, em particular, Correia 2009 e Vigário et al. 2010). Vigário et al. (2010) colocam a hipótese de as palavras monossilábicas contarem como agudas, fazendo aproximar muito a frequência das palavras com acento grave e com acento agudo. Uma outra linha de pesquisa, com efeitos similares, é considerar que os proclíticos adicionados a PWs monossilábicas podem torná-las di ou polissilábicas. Analisando o impacto da cliticização na distribuição dos padrões acentuais, verificamos que ele existe (diminuem os monossílabos e aumentam as PWs agudas), mas que, contudo, ele não é muito expressivo. De acordo com estes resultados, não deverá ser a cliticização que conduz a uma maior aproximação da frequência de palavras graves e agudas. Isto é compatível com o que vimos anteriormente sobre a diferenciação entre PWs e CLs: existe uma multiplicidade de evidências de natureza estatística para a diferencição entre PWs e CLs, e para a segmentação de CLs e PWs. Isso significa que o bebé dispõe precocemente de evidência estatítica para essa separação. Fazendo-a, é possível que, apesar de CLs e PWs serem integrados num mesmo constituinte fonológico num certo nível, eles sejam considerados distintos para efeitos da identificação do domínio de atribuição do acento. Os dados aqui discutidos sugerem, assim, que a hipótese alternativa de que as palavras monossilábicas contem para a contabilidade da distribuição do acento como palavras com acento final é presentemente a mais capaz de explicar as dificuldades iniciais manifestadas pelas crianças produção do acento.

5. Conclusão

A investigação aqui conduzida pôs em evidência uma multiplicidade de elementos de ordem estatística que permitem distinguir PWs e CLs e diferenciar sílabas átonas nas extremidades de PW de sílabas átonas pertencentes a CL (veja-se o resumo feito na secção precedente). Estes resultados são compatíveis com a hipótese de que informação estatística de frequência pode contribuir para que os bebés, numa fase muito precoce do processo de aquisição da linguagem, ainda pré-verbal, consigam segmentar CLs e PWs e assim estabelecer estas duas grandes categorias de palavras. Que os bebés são sensíveis a informação estatística no input tem sido recorrentemente demonstrado através de procedimentos experimentais diversificados (e.g., Jusczyk, Luce & Charles-Luce 1994; Saffran 2002).

A elevada frequência de ocorrência de certos formatos silábicos nos limites de constituintes prosódicos foi já proposta como podendo estar na base da identificação de palavras funcionais em momentos iniciais da aquisição da linguagem (Christophe et al. 2008). O que aqui propomos é a inclusão de muitos outros elementos relacionados com a frequência e distribuição de unidades e padrões fonético-fonológicos, fortalecendo assim a ideia original de que a criança pode chegar à segmentação de palavras funcionais a partir de informação presente no sinal sonoro. Efectivamente, os dados parecem mostrar que o bebé dispõe de informação muito diversificada no sinal, convergente no sentido da identificação de palavras gramaticais.

Para além da diversidade de aspectos de natureza distribucional, a nossa abordagem é distinta da de Christophe et al. (2008) por se concentrar nas diferenças entre PWs e CLs e não entre palavras lexicais e palavras funcionais. A nossa hipótese é a de que a criança usa a abundância de evidências distribucionais e fonético-fonológicas para diferenciar estas duas classes fonológicas, e que, dada a relação próxima entre os pares PW/palavra lexical e CL/palavra gramatical, essa categorização fonológica constitui a base para uma categorização inicial rudimentar das palavras como palavras lexicais e palavras gramaticais. Entre as questões que deixaremos aqui em aberto está a de efectivamente determinar se e quando os bebés são sensíveis aos vários aspectos estatísticos que distinguem PWs e CLs, que aqui descrevemos. Este tópico é deixado para investigação experimental futura.

Referências

Christophe, A., S. Millotte, S. Bernal, J. Lidz (2008) Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech* 51(1&2): 61-75.

Christophe, A., Guasti, M. T., Nespor, M., & van Ooyen, B. (2003). Prosodic structure and syntactic acquisition: the case of the head-complement parameter. *Developmental Science 6*, 213-222.

Correia, S. (2009) *The Acquisition of Primary Word Stress in European Portuguese*. Unpublished PhD thesis. University of Lisbon

Frota, S., M. Vigário, F. Martins & M. Cruz. (2010). FrePOP. Versão 1.0. Laboratório de Fonética (CLUL), Faculdade de Letras da Universidade de Lisboa. ISBN 978-989-95713-2-7. Disponível *online* em http://frepop.fl.ul.pt/.

Gervain, J. & J. Mehler (2010) Speech Perception and Lanugage Acquisition in the First Year of Life. *Annual Review of Psychology* 61: 191-218.

Gout, A., A. Christophe & J. Morgan (2004) Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language* 51: 548-567.

Guasti , Maria Teresa (2002) *Language acquisition. The growth of grammar.* MIT press.

Jusczyk, P., P. Luce, J. Charles-Luce (1994) Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33: 630-645.

Martins, F., M. Vigário & S. Frota (2009) *FreP – Frequency in Portuguese* v.2. *Software* de contagem de frequência (registo IGAC nº 209/2010) (versão 2.1.0.9).

Morgan, J. & K. Demuth (eds.) (1996) Signal to Syntax. Bootstrapping From Speech to Grammar in Early Acquisition. Mahwah, NJ: Lawrence Erlbaum Associates, 1-22.

Morgan, Shi & Allopena (1996) Perceptual Bases of Rudimentary Grammatical Categories: Toward a Broader Conceptualization of Bootstrapping. In J. Morgan & K. Demuth (eds.), 263-283.

Nespor, Marina & Irene Vogel (1986/2007) *Prosodic Phonology*. 2ª edição (2007). Berlin: Mouton de Gruyter.

Nespor, M., M. T. Guasti & A. Christophe (1996) Selecting Word Order: The Rhythmic Activation Principle. In U. Kleinhenz (ed.) *Interfaces in Phonology*. Berlin: Akademie Verlag, 1-26.

Saffran, J. R. (2002) Constraints on statistical language learning. *Journal of Memory and Language* 47: 172-196.

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

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): 175-203 (Special Issue on Crosslinguistic Perspectives on the Development of Prosodic Words, guest-edited by Katherine Demuth).

Vigário, M., F. Martins & S. Frota (2006). 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.

Vigário, M., S. Frota & F. Martins (2010) A frequência que conta na aquisição da fonologia: *types* ou *tokens*. In Ana Maria Brito, Fátima Silva, João Veloso e Alexandra Fiéis (orgs.) *XXV Encontro Nacional da Associação Portuguesa de Linguística. Textos seleccionados*. Porto: Associação Portuguesa de Linguística, 749- 767.

www.fl.ul.pt/LaboratorioFonetica/FreP.