Contagem da frequência dos bigramas em palavras de quatro a seis letras do português brasileiro

Bigram frequency for four- through six- letter words of Brazilian Portuguese

Cláudia Nascimento Guaraldo **Justi** [I,*]
Francis Ricardo dos Reis **Justi** [II]

RESUMO

Vários estudos sobre leitura e escrita têm demonstrado a influência de unidades sublexicais no processamento da linguagem. Uma das medidas de unidades sublexicais mais conhecida é a frequência de ocorrência dos bigramas (pares ordenados de letras que co-ocorrem nas palavras de uma língua). Visando preencher uma lacuna nos dados psicolinguísticos relativos ao português brasileiro, o presente estudo disponibiliza aos pesquisadores dados relativos à frequência de ocorrência dos bigramas (*type* e *token*) em palavras de quatro a seis letras do português brasileiro. Os resultados desse estudo demonstraram que a correlação média entre as medidas *type* e *token* foi forte e que, de uma forma geral, há mais redundância ortográfica no final das palavras de quatro a seis letras, uma vez que os bigramas em posições finais foram sempre em menor número e significativamente mais frequentes do que os bigramas nas outras posições.

Palavras-chave: bigramas; redundância ortográfica; análise de corpus linguístico.

ABSTRACT

Many studies about reading and writing have demonstrated the influence of sublexical units on language processing. One of the measures of sublexical units more used in the field of psycholinguistic is the bigram frequency. Aiming to fill in a gap on Brazilian Portuguese's psycholinguistics data, the present study provides for interested researchers data about bigram frequency (type and token) for four- through six- letter words of Brazilian Portuguese. The results of statistical analysis demonstrated that the mean correlation between type and token measures was strong. In addition, there was more orthographic redundancy at the end of the four- through six- letter words, because the bigrams of last position were always in minor number and significantly more frequent than bigram of other positions.

Keywords: bigrams; orthographic redundancy; *corpus* linguistics

[I] Universidade Federal de Pernambuco – UFPE (Recife)
[II] Universidade Federal de Alagoas – UFAL (Maceió)

Como na leitura o acesso lexical é considerado um dos processos mais importantes (Perfetti, 1985), não é surpreendente que muitas pesquisas tenham se dedicado à investigação do efeito de variáveis lexicais como a frequência, a regularidade e a vizinhança ortográfica no reconhecimento visual de palavras (ver Justi & Justi, 2009 e Roazzi, Justi & Justi, 2008 para uma revisão desses achados no português brasileiro). Recentemente, no entanto, vários pesquisadores têm chamado a atenção para a necessidade de serem analisadas também as características sublexicais dos estímulos e seus possíveis efeitos sobre o reconhecimento visual ou produção escrita de palavras (Cassar & Treiman, 1997; Conrad, Carreiras, Tamm & Jacobs, 2009; Deacon, Conrad & Pacton,

2008; Pacton, Perruchet, Fayol & Cleeremans, 2001; Treiman, 1993; Westbury & Buchnan, 2002).

De acordo com Carreiras e Grainger (2004), pode-se pensar nas características sublexicais das palavras como sendo unidades funcionais que intervêm entre o processamento perceptual de baixo nível (como a detecção de traços e bordas) e o acesso à representação mental das palavras na memória de longo prazo (acesso lexical). Nesse sentido, as unidades sublexicais podem ser subdivididas de acordo com suas características em: ortográficas; fonológicas; morfológicas. O foco deste trabalho recai sobre as características ortográficas das unidades sublexicais. Em ortografias alfabéticas como o português, pode-se dizer que a representação abstrata das letras é uma das principais unidades sublexicais do ponto de vista ortográfico, já que muitos modelos de reconhecimento visual de palavras assumem que, de algum modo, ou a representação abstrata das letras é mapeada em uma representação da forma ortográfica da palavra que as contém (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1981; Murray & Forster, 2004; Paap & Johansen, 1994; Whitney, 2001), ou é mapeada diretamente em um *output* fonológico ou semântico (Harm & Seidenberg, 2004; Plaut, Seidenberg, McClelland & Patterson, 1996; Seidenberg & McClelland, 1989). Porém, um problema que surge a partir da representação abstrata das letras é: como codificar a posição das letras nas palavras? De acordo com Carreiras e Grainger (2004) e Grainger e Whitney (2004), uma solução para esse problema seria postular um nível representacional sublexical que codificasse pares ordenados de letras ou bigramas. Por exemplo, a palavra 'toca' seria codificada pelos bigramas 'to', 'oc' e 'ca'. Embora difiram na forma como codificam os bigramas, exemplos recentes dessa tendência são o modelo de Superposição de Bigramas Abertos (Grainger, Granier, Farioli, Van Assche & van Heuven, 2006) e o modelo SERIOL (Whitney, 2001; Whitney & Cornelissen, 2008). No entanto, o modelo de processamento paralelo distribuído de Seidenberg e McClelland (1989) já propunha que esse tipo de unidade sublexical (bigramas) poderia emergir no modelo devido ao seu grau de redundância ortográfica. Nesse sentido, se os bigramas, por um lado, são vistos por alguns pesquisadores como uma forma de codificar a ordem das letras nas palavras, por outro lado, podem ser encarados também como uma medida de redundância ortográfica.

É a observação de que as palavras escritas de uma língua são formadas apenas por certas combinações possíveis de letras e que, dentro das concatenações permissíveis, alguns padrões ocorrem mais frequentemente do que outros que suscita a ideia de que os bigramas podem ser vistos como medidas de redundância ortográfica (Novick & Sherman, 2004). Por exemplo, em ortografias alfabéticas, a frequência de ocorrência de determinadas combinações de letras varia: o padrão ortográfico 'ck' é mais comum do que o 'lk' na língua inglesa. Há também restrições posicionais sobre a ocorrência de determinados padrões ortográficos: no inglês, palavras podem terminar, mas não começar com 'ck'. Alguns padrões ortográficos envolvem letras duplas (sequência de duas letras idênticas). No inglês, por exemplo, certas letras podem ser duplicadas (como o 'o' em 'b**oo**k' e o 'l' em 'fi**ll**'), ocorrendo, tipicamente, no meio e no fim das palavras, mas não no início (Seidenberg, 1989). A existência desse tipo de regularidade nos padrões ortográficos de uma língua,

principalmente os indexados pela freqüência dos bigramas, tem ensejado diversas pesquisas. Por exemplo, no caso da língua inglesa, estudos de pesquisadores como Treiman (1993; Treiman & Bourassa, 2000; Cassar & Treiman, 1997) têm evidenciado que, a partir do jardim de infância, as crianças já demonstram uma sensibilidade à frequência de ocorrência dos padrões ortográficos de sua língua, seja na leitura (p. ex., ao julgar quais sequências de letras eram mais parecidas com uma palavra real, as crianças consideraram as restrições ortográficas do inglês), seja na escrita (p.ex., como na língua inglesa as palavras não podem começar com 'ck', as crianças raramente cometiam esse tipo de erro ao tentar escrever uma palavra). Ampliando a generalidade desses achados, Pacton e cols. (2001), pesquisando crianças falantes da língua francesa, demonstraram que a sensibilidade aos padrões ortográficos, em especial às letras que podem aparecer duplicadas em palavras do francês, ocorre de forma independente da frequência das letras isoladas que constituem o padrão. Assim sendo, os resultados do estudo de Pacton e cols. (2001) revelaram que as crianças foram sensíveis à frequência com que duas letras iguais ocorriam juntas e não meramente à frequência absoluta de cada letra.

Os resultados dos estudos de Treiman (1993), Cassar e Treiman (1997) e Pacton e cols. (2001) sugerem que, gradualmente, e mesmo após poucos meses de exposição à língua escrita, as crianças desenvolvem uma sensibilidade às restrições ortográficas de sua língua baseadas na frequência de ocorrência das unidades sublexicais presentes nas palavras a que são expostas. Ora, como a frequência de ocorrência dessas unidades sublexicais é normalmente indexada pela frequência de ocorrência dos bigramas, a frequência de ocorrência dos bigramas passa a ser, pelo menos, uma variável a ser controlada nos estudos sobre a leitura e/ou escrita. Por exemplo, no âmbito dos estudos sobre o reconhecimento visual de palavras, pesquisadores, como Rastle, Davis e New (2004) sugerem que alguns dos efeitos encontrados pelos pesquisadores ao investigarem o processamento morfológico podem ser, na realidade, efeitos da frequência de ocorrência dos bigramas que compõem as palavras. De acordo com tais pesquisadores, enquanto as raízes e os afixos têm, tipicamente, bigramas de alta frequência de ocorrência, a frequência dos bigramas que se localizam entre os morfemas é menor. Assim sendo, sequências de letras de baixa probabilidade (bigramas infrequentes) representariam o limite entre o término de um morfema e o início de outro. Uma vez que o ser humano parece ser sensível à frequência de ocorrência desses padrões ortográficos desde cedo, isso explicaria alguns dos efeitos que sugerem uma decomposição morfológica da palavra em tarefas de reconhecimento visual de palavras utilizando o paradigma de *priming* (Longtin, Segui, & Hallé, 2003; Rastle & Davis, 2009).

De uma forma geral, além dos estudos sobre o reconhecimento visual de palavras (p.ex. Conrad, Carreiras, Tamm & Jacobs, 2009) e sobre o desenvolvimento do conhecimento ortográfico (p.ex. Pacton e cols., 2001), estudos realizados em outras áreas do conhecimento têm considerado ser importante investigar o papel dos bigramas, seja na solução de anagramas (p.ex. Mayzner & Tresselt, 1963), seja nos estudos sobre a memorização de palavras novas (p.ex. Dorfman, 1999). Ademais, outro indício da importância dessa variável é que uma pesquisa na base de dados *PsicInfo*, realizada em agosto de

2009 pelos autores desse trabalho, utilizando os termos *bigram* ou *bigrams*, tendo como base apenas trabalhos publicados nos últimos 5 anos, resultou em 49 artigos, o que representa um número bastante razoável de trabalhos considerando-se a especificidade do tema.

Tendo em vista a importância dos bigramas como uma medida de redundância ortográfica ou como unidades sublexicais em modelos de reconhecimento de palavras, é natural que diversos pesquisadores utilizem alguma medida de frequência dos bigramas em seus estudos, seja para investigar diretamente seus efeitos, seja para controlá-los. De acordo com Novick e Sherman (2004), existem diferentes tipos de medidas de frequência de bigramas, algumas sendo sensíveis à posição em que os bigramas aparecem nas palavras, enquanto outras não. As medidas sensíveis à posição em que os bigramas ocorrem, normalmente, levam em consideração também o número de letras das palavras, posto que algumas posições variam de acordo com o número de letras das palavras (p.ex.: contando letra a letra, da esquerda para a direita, a posição 3 é a posição final em palavras de 4 letras, no entanto, em palavras de 6 letras, é uma das posições mediais). Desse modo, normas de bigramas sensíveis à posição especificam quais bigramas são mais frequentes, por posição, em palavras de um determinado número de letras; já normas que não são sensíveis à posição, especificam a frequência de ocorrência de um determinado bigrama, independentemente, do número de letras das palavras e de onde ele ocorre nelas. Além disso, a frequência dos bigramas pode ser computada de duas formas: de acordo com o número de palavras em que o bigrama aparece (*type frequency*); e, pelo número de vezes em que o bigrama aparece de uma forma geral (*token frequency*[1]). Por exemplo, considerando uma contagem sensível à posição e ao número de letras, na frase: "Fábio joga muito bem vôlei, por isso joga na seleção", o bigrama 'jo', na posição inicial de palavras de 4 letras, tem frequência de *type* igual a 1 e frequência de *token* igual a 2. O bigrama 'jo' tem uma frequência de *type* igual a 1 porque, no exemplo dado, aparece apenas na palavra 'joga' e tem uma frequência de *token* igual a dois porque ocorre duas vezes na frase.

No que diz respeito à distinção entre a frequência de *type* e a frequência de *token*, autores como Novick e Sherman (2004) têm argumentado que a frequência de *token* pode ser enganosa, uma vez que nessa se confundem a frequência do bigrama e a frequência das palavras nas quais ele ocorre (assim, um bigrama pode parecer frequente não porque ocorre em muitas palavras, mas porque ocorre em uma palavra muito frequente). Já no que diz respeito às contagens de bigramas que não são sensíveis à posição, pode-se dizer que estas deixam de capturar informações potencialmente relevantes. Por exemplo, na língua inglesa, o bigrama 'ck', embora possa ser frequente, não ocorre no início das palavras. Sendo assim, as contagens não posicionais de bigramas são cegas a essas possíveis regularidades da língua.

Todos os dados psicolinguísticos apresentados no corpo desse artigo até o momento são relativos à língua inglesa, pois, em nosso conhecimento, ainda não se encontra disponível publicamente nenhuma medida da frequência de ocorrência dos bigramas relativa ao

[1] Optou-se por não traduzir os termos '*type*' e '*token*' por esses já terem se tornados clássicos na área.

português brasileiro. Infelizmente, essa escassez de dados tem dificultado os pesquisadores a investigar e/ou controlar os possíveis efeitos dos bigramas em pesquisas desenvolvidas na língua portuguesa (p.ex.: Justi & Pinheiro, 2006; Justi & Justi, 2007). Tendo essa lacuna nos dados psicolinguísticos da língua portuguesa em mente, além da importância teórica de se estudar os efeitos dos bigramas, o presente estudo teve como objetivo prover os pesquisadores brasileiros de dados sobre a frequência de ocorrência dos bigramas das palavras de 4 a 6 letras do português brasileiro. Para tal, optou-se pelo cômputo das frequências de *type* e de *token* dos bigramas, por posição e número de letras das palavras. Essa opção se justifica porque, conforme ressaltado por Novick e Sherman (2004), as contagens não posicionais podem deixar de capturar regularidades ortográficas relevantes da língua. Além disso, o leitor interessado em uma contagem não posicional pode facilmente obter esses dados com base em uma contagem posicional (basta somar a frequência relatada de um bigrama para todas as posições em que ele ocorre), já o inverso não é verdadeiro.

MÉTODO

Existem, pelo menos, duas listas de contagem da frequência de ocorrência de palavras no português do Brasil, a lista de Pinheiro (1996) e a lista desenvolvida pelo NILC (2005). No presente estudo optou-se pela lista do NILC (2005) por duas razões: 1) ela tem como base uma amostra maior de palavras e isso aumenta a sensibilidade da contagem da frequência de *type* dos bigramas; 2) em estudo anterior Justi e Justi (2008) desenvolveram estatísticas de vizinhança ortográfica para 8465 palavras de 4 a 6 letras da lista do NILC. Sendo assim, parece mais produtivo somar dados psicolinguísticos a esse conjunto de palavras para que os pesquisadores possam ter dados mais completos ao invés de ter dados fragmentários sobre conjuntos de estímulos diferentes. Destarte, a base de dados que serviu de cálculo para a frequência de ocorrência dos bigramas no presente estudo constitui-se das 8465 palavras de 4 a 6 letras retiradas do corpus NILC (2005) que foram analisadas previamente por Justi e Justi (2008). De qualquer forma, para avaliar a generalidade das frequências dos bigramas gerados neste estudo, uma amostra de 2298 palavras de seis letras comuns à terceira e quarta série do trabalho de Pinheiro (1996) foi selecionada, e uma análise da correlação entre a frequência dos bigramas gerados no presente trabalho e a frequência dos bigramas gerados com base na lista de Pinheiro foi efetuada. Foram escolhidas palavras de seis letras da lista de Pinheiro, pois, no conjunto de palavras de 4 a 6 letras dessa lista, as de 6 letras são as que existem em maior número.

Material

Foram utilizadas no presente estudo 8465 palavras de quatro a seis letras do corpus NILC (2005) e 2298 palavras de seis letras comuns à 3ª e 4ª série da lista de Pinheiro (1996).

Procedimentos

Da mesma forma que no estudo de Justi e Justi (2008), foram selecionadas para análise apenas as palavras de 4 a 6 letras do corpus NILC (2005) com frequência de, pelo menos, uma ocorrência por milhão, sendo todas as

palavras estrangeiras e hifenizadas descartadas, resultando em uma base de dados de 8465 palavras. Tais palavras foram organizadas em uma planilha do '*MS Excel*', e comandos para a extração de *strings* foram utilizados para a extração dos bigramas. Os bigramas foram extraídos de acordo com o número de letras das palavras e a posição em que ocorriam nelas. Após esse procedimento, os bigramas gerados foram analisados no programa '*SPSS for Windows*' versão 16.0, sendo geradas as seguintes estatísticas: frequência de ocorrência de *type* e de *token* para bigramas que iniciam em palavras de 4 letras nas posições 1, 2 e 3 (p.ex.: para a palavra 'pato' o bigrama que inicia na posição 1 é 'pa', na 2 é 'at' e na 3 é 'to'); frequência de ocorrência de *type* e de *token* para bigramas que iniciam em palavras de 5 letras nas posições 1, 2, 3 e 4 (p.ex.: para a palavra 'chave' o bigrama que inicia na posição 1 é 'ch', na 2 é 'ha', na 3 é 'av' e na 4 é 've'); frequência de ocorrência de *type* e de *token* para bigramas que iniciam em palavras de 6 letras nas posições 1, 2, 3, 4 e 5 (p.ex.: para a palavra 'casaco' o bigrama que inicia na posição 1 é 'ca', na 2 é 'as', na 3 é 'sa', na 4 é 'ac' e na 5 é 'co').

No que diz respeito às palavras de 6 letras, comuns às 3ª e 4ª séries da lista de Pinheiro (1996), optou-se por gerar bigramas para apenas uma das posições dessas palavras. Para tanto, um sorteio foi realizado e a posição 1 foi selecionada. Esse procedimento se justifica porque gerar bigramas é um processo bastante dispendioso e não há, em princípio, qualquer razão para se imaginar que a correlação entre a frequência dos bigramas das duas listas irá variar sistematicamente de acordo com sua posição, uma vez que se trata de palavras com o mesmo número de letras. O mesmo procedimento empregado para gerar os bigramas da lista do NILC foi utilizado para gerar os bigramas da posição inicial de palavras de seis letras da lista de Pinheiro.

RESULTADOS

Como pode ser observado na Tabela 1, ao se considerar separadamente as palavras de acordo com o número de letras, há mais bigramas diferentes nas posições mediais das palavras do que nas posições iniciais e finais, sendo os bigramas das posições finais sempre em menor número.

	Palavras de 4 letras			**Palavras de 5 letras**			**Palavras de 6 letras**		
Posição	**N_BG**	**F_TY**	**F_TK**	**N_BG**	**F_TY**	**F_TK**	**N_BG**	**F_TY**	**F_TK**
P1	193	5,7	418,6	257	11,5	384,2	277	15,9	299,9
P2	242	4,6	333,8	304	9,8	324,8	371	11,8	223,9
P3	173	6,4	466,9	316	9,4	312,4	336	13,1	247,2
P4	--	--	--	170	17,5	580,8	285	15,4	291,5
P5	--	--	--	--	--	--	166	26,5	500,4
Média	202,7	5,4	398,6	261,7	11,4	377,2	287	15,3	289,4

Nota: P1 = Posição 1; P2 = Posição 2; P3 = Posição 3; P4 = Posição 4 e P5 = Posição 5; N_BG = número de bigramas; F_TY = frequência média de *type*; F_TK = frequência média de *token*.

Tabela 1 – *Número de bigramas em palavras de 4 a 6 letras por posição*

A fim de verificar se haveria alguma diferença estatisticamente significativa entre a frequência dos bigramas das palavras de 4 a 6 letras de acordo com a posição em que eles ocupam nas palavras, duas análises de variância foram efetuadas: uma, tendo como variável dependente a frequência de *type*, e a outra, tendo como variável dependente a frequência de *token*. Para as palavras de quatro letras, a ANOVA que considerou a frequência de *token* não foi estatisticamente significante (p > 0,46). Já a ANOVA que considerou a frequência de *type* foi estatisticamente significante, F(2,605) = 4,341 e *p* = 0,013. Para explorar melhor esse resultado, análises *post hoc*, utilizando a correção de Bonferroni, foram realizadas. Essas análises revelaram que a única diferença estatisticamente significativa ocorreu entre a frequência dos bigramas da posição final e a frequência dos bigramas da posição medial (*p* < 0,05), sendo os bigramas da posição final significativamente mais frequentes. Nenhuma das outras comparações *post hoc* foi estatisticamente significativa (para todas, *p* > 0,17).

No caso das palavras de 5 letras, tanto a ANOVA que considerou a frequência de *type*, F(3,1043) = 9,192, quanto a que considerou a frequência de *token*, F(3,1043) = 5,544, foram estatisticamente significantes (para ambas *p* < 0,01). No entanto, em ambos os casos, análises *post hoc*, utilizando a correção de Bonferroni, revelaram que as únicas diferenças estatisticamente significativas existentes foram entre a frequência dos bigramas da posição final e a frequência dos bigramas das outras posições (para todas, *p* < 0,05), sendo os bigramas da posição final significativamente mais frequentes do que os bigramas das demais posições. Nenhuma das outras comparações *post hoc* foi estatisticamente significativa (para todas, *p* > 0,50).

Por fim, as palavras de 6 letras apresentaram o mesmo padrão das palavras de 5 letras. Tanto a ANOVA que considerou a frequência de *type*, F(4,1430) = 9,276, quanto a que considerou a frequência de *token*, F(4,1430) = 6,831 foram estatisticamente significantes (para ambas *p* < 0,001). Mas, em ambos os casos, análises *post hoc*, utilizando a correção de Bonferroni, revelaram que as únicas diferenças estatisticamente significativas ocorreram entre a frequência dos bigramas da posição final e a frequência dos bigramas das outras posições (para todas, *p* < 0,01), sendo os bigramas da posição final significativamente mais frequentes do que os bigramas das demais posições. Nenhuma das outras comparações *post hoc* foi estatisticamente significativa (para todas, *p* > 0,50).

Para verificar se a frequência de *type* e a frequência de *token* se relacionam, análises de correlação de Pearson foram desenvolvidas levando-se em consideração o número de letras e a posição dos bigramas nas palavras. Para o cálculo da correlação média, foi utilizado o programa *Meta-Analysis* versão 5.3 de Schwarzer (1989). Foi adotado o seguinte procedimento: primeiro, os valores *r* foram transformados em valores Zr de Fisher; depois o valor Zr médio foi calculado; e, finalmente, foi transformado no valor *r* médio. A correlação média geral entre as frequências de *type* e de *token* foi de 0,74 e a correlação média por número de letras foi de 0,54 para as palavras de 4 letras; de 0,80 para as palavras de 5 letras e de 0,79 para as palavras de 6 letras.

No que diz respeito aos bigramas mais e menos frequentes, a título de ilustração, os ANEXOS 1, 2 e 3 apresentam os dados relativos aos 20

bigramas mais frequentes e aos 20 bigramas menos frequentes, de acordo com a posição, em palavras de quatro, cinco e seis letras, respectivamente. Como a escolha entre a utilização da medida de *type* e a utilização da medida de *token* fica a critério do pesquisador, ambas as medidas estão disponíveis. Como as frequências de *type* e de *token*, porém, não são necessariamente semelhantes, optou-se por levar em consideração o valor da frequência de *type* ao dispor em ordem decrescente os bigramas mais frequentes. Dessa forma, como pode ser observado no ANEXO 1, o bigrama 'ca' apresentou a maior frequência de *type* no início de palavras de 4 letras, mas não apresentou a maior frequência de *token*. No caso dos bigramas menos frequentes, como eles tiveram, via de regra, uma frequência de *type* igual a um, utilizou-se a medida *token* para a disposição dos resultados em ordem decrescente nos anexos. Por exemplo, um dos bigramas na posição inicial de palavras de quatro letras que têm frequência de *type* igual a um é o bigrama 'ég'. Esse bigrama, contudo, tem frequência de *token* igual a quatro e, como foram apresentados nos anexos apenas os vinte bigramas menos frequentes, esse bigrama não aparece no ANEXO 1, pois um bigrama com frequência de *token* menor foi apresentado no lugar dele.

Por fim, objetivando investigar a generalidade da contagem da frequência de ocorrência dos bigramas realizada neste estudo, foi efetuada uma análise da correlação entre a frequência dos bigramas presentes em palavras de seis letras na posição inicial da lista do NILC (2005) e a frequência dos bigramas presentes na posição inicial em palavras de seis letras, comuns à 3ª e 4ª série, da lista de Pinheiro (1996). Os resultados revelaram uma forte correlação entre essas medidas (r = 0,96 para a correlação entre a frequência de *type* e r = 0,85 para a correlação entre a frequência de *token*).

DISCUSSÃO

Os resultados do presente estudo demonstram que há mais redundância ortográfica no final das palavras de 4 a 6 letras do português brasileiro, porquanto os bigramas em posições finais foram, de uma forma geral, sempre em menor número e significativamente mais frequentes do que os bigramas nas outras posições. Uma das possíveis explicações para esse resultado é que os bigramas nas posições finais de palavras podem coincidir com sufixos. Esse provavelmente é o caso do bigrama 'ou' que, embora não tenha aparecido entre os bigramas mais frequentes nas palavras de 4 letras, apareceu entre os dez bigramas mais frequentes no final das palavras de 5 e 6 letras. Isso ocorreria porque a sequência de letras 'ou' no final de palavras do português brasileiro, geralmente, indica uma flexão verbal. Esse argumento é reforçado ao se considerar que a probabilidade de uma palavra ter uma flexão aumenta, muito provavelmente, junto com o seu número de letras. Isso explicaria o fato do bigrama 'ou' ter aparecido entre os mais frequentes nas palavras de 5 e 6 letras, mas não ter aparecido entre os vinte mais frequentes nas palavras de 4 letras. Pode-se dizer que tais resultados corroboram o argumento de Rastle e cols. (2004) de que alguns dos efeitos encontrados pelos pesquisadores ao investigarem o processamento morfológico podem ser, na realidade, efeitos da frequência de ocorrência dos bigramas que compõem as palavras, já que os afixos têm, tipicamente, bigramas de alta frequência de ocorrência. Destarte, torna-se

necessário que pesquisadores interessados em investigar o processamento morfológico controlem a frequência de ocorrência dos bigramas em seus estudos para evitar que os resultados possam ser atribuídos a uma diferença nessa variável.

Outra questão com a qual o presente estudo se deparou refere-se à qual medida seria a mais apropriada da frequência de um bigrama. O presente estudo computou tanto a frequência de *type* quanto a frequência de *token* e encontrou que a correlação entre essas medidas foi moderada para palavras de 4 letras e forte para palavras de 5 e 6 letras. Esse padrão é esperado, já que a frequência de *token* é contaminada pela frequência de ocorrência das palavras e, nesse caso, como a frequência de ocorrência das palavras de 4 letras é maior do que a das palavras de 5 e 6 letras, não é de se surpreender que a correlação entre as medidas *type* e *token* tenha sido menor nesse conjunto de palavras. Nesse sentido, pode-se argumentar, em consonância com Novick e Sherman (2004), que a medida mais segura da frequência de ocorrência de um bigrama é, de fato, a frequência de *type*.

Para finalizar, tendo a lista do NILC (2005) sido desenvolvida com base em textos jornalísticos, voltados primordialmente para o público adulto, uma questão torna-se pertinente: são os dados gerados pelo presente trabalho válidos para pesquisas a serem realizadas com crianças? Diferentemente da lista do NILC (2005), a lista de Pinheiro (1996) foi desenvolvida tendo como base livros infantis utilizados por crianças da pré-escola a 4ª série. Em suporte à generalidade dos resultados do presente estudo, foi observada uma forte correlação entre a frequência dos bigramas das palavras da lista do NILC e a frequência dos bigramas das palavras da lista de Pinheiro, o que atesta a generalidade da contagem de frequência de ocorrência dos bigramas efetuada nesse estudo.

CONSIDERAÇÕES FINAIS

Uma das características da mente humana é a sensibilidade à frequência de ocorrência dos estímulos a que é exposta. Nesse sentido, a frequência de ocorrência dos bigramas tem sido considerada uma importante variável sublexical nas pesquisas psicolinguísticas da atualidade, sendo uma das principais variáveis em estudos sobre a aquisição de padrões ortográficos (Cassar & Treiman, 1997; Pacton & cols., 2001; Treiman, 1993; Treiman & Bourassa, 2000) ou uma das variáveis a serem controladas em estudos sobre o processamento morfológico (Longtin & cols., 2003; Rastle & cols., 2004; Rastle & Davis, 2009). O presente artigo foi desenvolvido com o intuito de prover os pesquisadores brasileiros de dados psicolinguísticos referentes à frequência de ocorrência dos bigramas em palavras de 4 a 6 letras do português brasileiro, suprindo, dessa forma, uma importante lacuna nos dados psicolinguísticos dessa língua. Assim sendo, os pesquisadores brasileiros passam a contar com uma relevante fonte de dados psicolinguísticos que pode permitir uma investigação mais sofisticada do papel da redundância ortográfica na leitura e na escrita, bem como um melhor controle experimental em estudos em que outras variáveis são o foco da pesquisa. O banco de dados completo pode ser obtido gratuitamente mediante contato eletrônico com os autores deste artigo.

REFERÊNCIAS

Carreiras, M., & Grainger, J. (2004). Sublexical representations and the 'front end' of visual word recognition. *Language and Cognitive Processes*, *19*, 321-331.

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: children's knowledge of double letters in words. *Journal of Educational Psychology*, 89, 631-644.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.

Conrad, M., Carreiras, M., Tamm, S., & Jacobs, A. (2009). Syllables and bigrams: orthographic redundancy and syllabic units affect visual word recognition at different processing levels. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 461-479.

Deacon, S., Conrad, N., & Pacton, S. (2008). A statistical learning perspective on children's learning about graphotactic and morphological regularities in spelling. *Canadian Psychology*, 49, 118-124.

Dorfman, J. (1999). Unitization of sublexical components in implicit memory for novel words. *Psychological Science*, 10, 387-392.

Grainger, J., & Jacobs, A. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review*, 103, 518-565.

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences*, 8, 58-59.

Grainger, J., Granier, J., Farioli, F., Van Assche, E., & van Heuven, W. (2006). Letter position information and printed word perception: the relative-position priming constraint. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 865-884.

Harm, M., & Seidenberg, M. (2004). Computing the meaning of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, 111, 662-720.

Justi, C., & Justi, F. (2009). Os efeitos de lexicalidade, frequência e regularidade na leitura de crianças falantes do português brasileiro. *Psicologia: Reflexão e Crítica*, 22, 163-172.

Justi, F., & Pinheiro, A. (2006). O efeito de vizinhança ortográfica no português do Brasil: acesso lexical ou processamento estratégico? *Interamerican Journal of Psychology*, 40, 275-288.

Justi, F., & Justi, C. (2007). O efeito da freqüência de ocorrência dos bigramas e da sílaba inicial no reconhecimento visual de palavras no português do Brasil. *Anais do V Congresso Internacional da Associação Brasileira de Lingüística* (pp. 649-650). Belo Horizonte: FALE/UFMG.

Justi, F., & Justi, C. (2008). As estatísticas de vizinhança ortográfica das palavras do português e do inglês são diferentes? *Psicologia em Pesquisa*, 2, 61-73.

Longtin, C., Segui, J., & Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, 18, 313-334.

Mayzner, M., & Tresselt, M. (1963). Anagram solution times: a function of word length and letter position variables. *Journal of Psychology*, 55, 469-475.

McClelland, J., & Rumelhart, D. (1981). An Interactive activation model of

context effects in letter perception: part 1. An account of basic findings. *Psychological Review*, 88, 375-407.

NILC (2005). *Corpus NILC / São Carlos v.7.1*. Retirado em 30/08/2005, do site http://www.nilc.icms.usp.br.

Murray, W., & Forster, K. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111, 721–756.

Novick, L., & Sherman, S. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods, Instruments, & Computers*, 36, 397-401.

Paap, K., & Johansen, L. (1994). The case of the vanishing frequency effect: a retest of the verification model. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1129-1157.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of lab: the case of orthographic regularities. *Journal of Experimental Psychology: General*, 130, 401-426.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Pinheiro, A. (1996). *Contagem de freqüência de ocorrência e análise psicolingüística de palavras expostas a crianças na faixa pré-escolar e séries iniciais do 1º grau*. São Paulo: Associação Brasileira de Dislexia.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, 103, 56-115.

Rastle, K., Davis, M., & New, B. (2004). The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11, 1090-1098.

Rastle, K., & Davis, M. (2009). Morphological decomposition based on the analysis of orthography. In R. Frost, J. Grainger & M. Carreiras (Orgs.), *Advances in morphological processing* (pp. 942-971). Hove: Psychology Press.

Roazzi, A., Justi, C., & Justi, F. (2008). Da tinta à mente: uma discussão sobre os modelos computacionais de reconhecimento visual de palavras. In M. Maluf & S. Guimarães (Orgs.), *Desenvolvimento da linguagem oral e escrita* (pp. 95-121). Curitiba: Editora UFPR.

Schwarzer, R. (1989). *Meta-Analysis v.5.3*. Retirado em 30/09/2007, do site http://userpage.fu-berlin.de/~health/meta53.exe.

Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.

Seidenberg, M. (1989). Reading complex words. In G. Carlson & M. Tanenhaus (Orgs.), *Linguistic structure in language processing* (pp. 53-105). Kluwer Academic Publishers.

Treiman, R. (1993). *Beginning to spell*: a study of first-grade children. New York: Oxford University Press.

Treiman, R., & Bourassa, D. (2000). The Development of Spelling Skill. *Topics in Language Disorders, 20,* 1-18.

Westbury, C., & Buchnan, L. (2002). The probability of the least likely non-length- controlled bigram affects lexical decision reaction times. *Brain and Language*, 81, 66–78.

Whitney, C. How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review*, 8, 221-243, 2001.

Whitney, C., & Cornelissen, P. (2008). SERIOL reading. *Language and Cognitive Process*, 23, 143-164.

Endereço para correspondência:
R. Engenheiro Mário de Gusmão, 1295/104 - Ponta Verde - Maceió / AL – Brasil. CEP: 57035-000.
Tel.: (82)3235-1823.
E-mail: claudia.ngjusti@gmail.com

Recebido em Setembro de 2009
Aceito em Outubro de 2009

* Nota. Bolsista de doutorado do CNPq

Anexo 1

Frequências de type e de token para os vinte bigramas mais e menos frequentes de palavras de 4 letras por posição

POSIÇÃO 1			POSIÇÃO 2			POSIÇÃO 3		
BG	TY	TK	BG	TY	TK	BG	TY	TK
mais frequentes								
ca	36	2710	at	25	737	ra	38	12391
pa	23	12267	ot	24	506	to	36	1559
vi	21	1148	al	23	513	ta	35	1454
ma	20	4909	ai	22	5252	la	33	2892
to	20	988	el	20	4878	ca	31	542
co	19	4582	er	19	1273	ro	28	553
me	19	902	ol	19	401	ia	26	512
fi	18	562	ur	19	367	ma	26	946
pi	18	185	ar	18	11414	as	24	2624
so	18	245	ir	17	198	lo	23	2510
ba	17	327	in	16	127	co	22	214
bo	17	408	on	16	610	ga	22	468
fa	17	959	ac	15	113	os	22	1758
pe	17	4745	ed	15	596	da	21	2397
ra	17	164	ei	15	1145	go	21	1041
sa	17	677	et	15	374	de	20	2603
se	17	3119	oc	15	810	do	19	2427
te	17	943	od	14	2421	te	18	1111
re	16	903	ag	13	250	is	17	6577
do	15	1405	ic	13	507	na	17	651
menos frequentes								
éd	1	1	ál	1	1	ús	1	1
ím	1	1	ça	1	1	aí	1	1
aí	1	1	ço	1	1	bé	1	1
ae	1	1	íd	1	1	có	1	1
ai	1	1	íe	1	1	cô	1	1
bó	1	1	ôr	1	1	ci	1	1
cá	1	1	aú	1	1	dô	1	1
cz	1	1	ba	1	1	du	1	1
eg	1	1	cé	1	1	en	1	1
eu	1	1	ca	1	1	gá	1	1
fô	1	1	dr	1	1	gã	1	1
og	1	1	ft	1	1	gi	1	1
ox	1	1	gr	1	1	iô	1	1
pí	1	1	iç	1	1	lê	1	1
ps	1	1	iz	1	1	lô	1	1
sã	1	1	lc	1	1	lu	1	1
só	1	1	mã	1	1	pé	1	1
ui	1	1	mé	1	1	rê	1	1
vã	1	1	mi	1	1	tó	1	1
vé	1	1	oí	1	1	zó	1	1

Nota: BG = Bigrama; TY= Type; TK = Token

Anexo 2

Frequências de type e de token para os vinte bigramas mais e menos frequentes de palavras de 5 letras por posição em que iniciam na palavra

POSIÇÃO 1			POSIÇÃO 2			POSIÇÃO 3			POSIÇÃO 4		
BG	TY	TK	BG	TY	TK	BG	TY	TK	BG	TY	TK
mais frequentes											
ca	91	2207	or	99	4478	st	63	3094	as	202	4043
co	63	1676	ar	90	3008	ra	59	2290	os	172	5531
po	59	2884	er	90	3579	nt	57	2540	ar	117	3025
pa	58	1989	en	73	3390	ta	55	1283	ta	116	4190
ma	57	2045	al	70	1544	nd	45	3863	ia	97	3049
vi	55	1167	ra	65	1184	to	44	1324	to	92	4616
re	54	845	ur	64	985	la	43	1094	am	87	1729
ba	52	1302	an	57	2015	re	43	542	ra	86	2277
se	52	2299	in	55	3808	rr	43	878	es	82	2971
te	52	2879	ol	50	1894	it	41	2649	ou	81	1397
mo	50	969	as	46	1162	ri	41	1888	ão	78	3059
sa	50	1504	ri	46	739	ro	41	662	em	64	3393
pe	49	1876	es	45	5041	ca	37	1343	te	63	3701
su	47	373	re	45	1134	ia	35	268	do	62	2254
fa	46	2193	on	41	1478	te	35	2415	da	61	2749
to	45	1951	ot	41	947	rt	34	2586	ca	55	1765
de	43	2468	is	40	2773	nh	33	2244	ro	54	2002
tr	43	419	el	39	1532	ga	32	944	sa	50	1891
ve	40	1251	ir	39	435	ss	32	3892	al	44	2957
fo	38	3663	ei	37	1255	de	30	1394	io	43	856
menos frequentes											
pí	2	2	rê	2	2	íc	1	1	dê	2	2
rí	2	2	rs	2	2	úo	1	1	du	2	2
ró	2	2	so	2	2	ae	1	1	há	2	2
tí	2	2	ze	2	2	bá	1	1	ió	2	2
tc	2	2	ál	1	1	cã	1	1	lé	2	2
xa	2	2	áq	1	1	dã	1	1	mi	2	2
óc	1	1	çu	1	1	fá	1	1	pu	2	2
úr	1	1	ín	1	1	fê	1	1	ts	2	2
ax	1	1	ío	1	1	fo	1	1	tu	2	2
bó	1	1	ôr	1	1	iô	1	1	uê	2	2
ec	1	1	bs	1	1	jã	1	1	ur	2	2
fô	1	1	dr	1	1	jé	1	1	xá	2	2
ia	1	1	dv	1	1	jô	1	1	ós	1	1
lé	1	1	gn	1	1	lá	1	1	ôo	1	1
ló	1	1	lú	1	1	lb	1	1	aé	1	1
nó	1	1	nz	1	1	oj	1	1	bó	1	1
of	1	1	ré	1	1	oo	1	1	cô	1	1
ox	1	1	sc	1	1	pó	1	1	ix	1	1
tô	1	1	uç	1	1	pu	1	1	li	1	1
zí	1	1	vn	1	1	rô	1	1	té	1	1

Nota: BG = Bigrama; TY= Type; TK = Token

Anexo 3

Frequências de type e de token para os vinte bigramas mais e menos frequentes de palavras de 6 letras por posição em que iniciam na palavra

POSIÇÃO 1			POSIÇÃO 2			POSIÇÃO 3			POSIÇÃO 4			POSIÇÃO 5		
BG	TY	TK	BG	TY	TK	BG	TY	TK	BG	TY	TK	BG	TY	TK
mais frequentes														
ca	149	1741	ar	127	2560	st	100	1307	ad	217	3131	as	371	6229
re	131	1512	or	117	2608	nt	98	3615	ta	148	2332	os	361	7233
co	123	3168	en	116	2647	ti	76	1522	to	120	2192	ar	249	3584
de	96	3214	Ra	110	4043	ra	73	778	id	107	1587	do	190	4742
ma	90	1004	an	106	1717	ri	60	1181	ra	103	1327	am	189	1571
tr	82	701	er	103	1586	nd	59	689	ia	92	1352	da	171	1631
pa	78	2289	al	96	652	ta	58	1652	ic	79	1440	ou	166	2238
ba	71	779	es	79	1406	it	57	1179	te	75	869	es	158	2459
pe	70	1236	Re	75	1378	rr	55	730	da	64	669	ão	131	2380
es	66	2592	on	74	2169	ca	51	560	av	61	811	ia	117	1393
pr	65	1656	ol	71	814	di	50	1131	ro	61	1634	ra	116	3190
sa	65	381	as	70	893	rt	50	1312	do	60	571	ta	115	1069
vi	65	660	Ri	69	775	ss	49	1688	in	60	3413	em	96	1083
po	63	2000	ur	64	508	an	48	3991	er	59	1396	to	87	2027
ch	62	1073	In	62	742	en	48	1716	it	59	903	ca	81	1025
se	59	2076	om	60	1194	rd	45	468	ca	57	622	is	70	2055
mo	58	1198	RO	55	820	li	44	163	sa	57	1387	va	70	838
al	55	1377	Is	47	569	vi	44	723	re	53	863	na	69	3930
ve	55	885	ac	46	188	ci	43	535	ai	50	922	la	65	912
su	53	295	ec	46	520	er	42	957	ar	50	454	co	61	972
menos frequentes														
um	2	2	áu	1	1	áx	1	1	úd	1	1	ía	1	1
vô	2	2	ço	1	1	ér	1	1	úr	1	1	ôo	1	1
zâ	2	2	çu	1	1	ét	1	1	aí	1	1	bu	1	1
zu	2	2	és	1	1	ên	1	1	af	1	1	gi	1	1
ág	1	1	Ia	1	1	ío	1	1	cê	1	1	hó	1	1
áu	1	1	Ío	1	1	ót	1	1	dô	1	1	hu	1	1
êm	1	1	Íz	1	1	úb	1	1	iu	1	1	jó	1	1
óx	1	1	ôq	1	1	ax	1	1	ju	1	1	je	1	1
ôh	1	1	úf	1	1	bó	1	1	lç	1	1	lã	1	1
dó	1	1	úr	1	1	cõ	1	1	ld	1	1	lé	1	1
gã	1	1	ae	1	1	eu	1	1	lg	1	1	li	1	1
gn	1	1	ao	1	1	fõ	1	1	lm	1	1	mã	1	1
hí	1	1	ej	1	1	iz	1	1	nô	1	1	nê	1	1
ji	1	1	Ft	1	1	jõ	1	1	nu	1	1	ni	1	1
mâ	1	1	ga	1	1	mõ	1	1	pê	1	1	pá	1	1
né	1	1	hm	1	1	mf	1	1	rç	1	1	pó	1	1
nô	1	1	Iê	1	1	né	1	1	ró	1	1	rô	1	1
og	1	1	nj	1	1	nl	1	1	rô	1	1	rs	1	1
pê	1	1	oj	1	1	oj	1	1	sp	1	1	tá	1	1
sê	1	1	ox	1	1	ps	1	1	ug	1	1	ur	1	1

Nota: BG = Bigrama; TY= Type; TK = Token