As estatísticas de vizinhança ortográfica das palavras do português e do inglês são diferentes?[1]

Are orthographic neighborhood statistics of Portuguese and English words different?

Francis Ricardo dos Reis **Justi** *
Cláudia Nascimento Guaraldo **Justi** **

Resumo

Este trabalho avaliou a correlação entre medidas tradicionais de vizinhança ortográfica (N e NF) e o número de vizinhos ortográficos formados pela transposição de duas letras (TLN), bem como se diferenças na distribuição estatística das variáveis de vizinhança ortográfica poderiam explicar as diferenças encontradas nos efeitos dessas variáveis na língua portuguesa e inglesa. Para tanto, foram geradas estatísticas de vizinhança ortográfica para 8465 palavras do português. Não foram observadas grandes diferenças entre as palavras do português e as da língua inglesa no que tange às estatísticas de vizinhança ortográfica, tornando difícil explicar as discordâncias encontradas nos efeitos de N e NF com base na ideia de estruturas de vizinhança ortográfica diferentes entre as línguas. Além disso, a correlação entre as medidas tradicionais de vizinhança ortográfica e TLN foram ínfimas, o que indica que os efeitos de N e NF, sejam quais forem, dificilmente podem ser explicados por sua relação com TLN.

Palavras-chave: vizinhança ortográfica; acesso lexical; reconhecimento visual de palavras; linguística de corpus; comparações entre línguas;

Abstract

This work aimed to evaluate two hypotheses: are traditional orthographic neighborhood measures (N and NF) correlated with transposed letter neighbors (TLN)? Could differences in the neighborhood structure of Portuguese and English words explain the differences observed in N and NF effects across these languages? To shed light on these issues orthographic neighborhood statistics were generated for 8465 Portuguese words. The difference between languages in the relevant variables were minor, making it difficult to explain the differences observed in N and NF effects across these languages appealing to differences in the neighborhood structure of Portuguese and English words. Another point of interest was the correlation between traditional orthographic neighborhood measures and TLN which were very low. Therefore, N and NF effects can hardly be explained by their relationships with TLN.

Key-Words: orthographic neighborhood; lexical access; visual word recognition; corpus linguistics; cross-language comparisons;

[1] Os autores gostariam de agradecer à *Psychonomic Bulletin & Review* por permitir a reprodução dos dados do trabalho de Andrews (1997) nas tabelas 3 e 4 desse estudo.

* Universidade Federal de Alagoas – UFAL / Maceió

Contato: Universidade Federal de Alagoas – Campus A. C. Simões, ICHCA, Curso de Psicologia, Av. Lourival Melo Mota, s/n, Bairro Tabuleiro do Martins, Maceió, AL, Brasil, CEP: 57072-970. E-mail: francisjusti@gmail.com

** Universidade Federal de Pernambuco – UFPE / Recife

Durante a leitura, o acesso lexical é um dos processos cognitivos mais importantes, pois envolve o acesso à representação mental das palavras que reconhecemos em um texto, o que nos permite compreendê-las e pronunciá-las (Perfetti, 1985). Nesse sentido, uma das principais tarefas que o nosso sistema cognitivo desempenha durante o acesso lexical é a seleção da entrada lexical

correta diante das diversas entradas que podem ser ortograficamente similares ao estímulo-alvo. Dentro da tradição de pesquisas sobre o reconhecimento visual de palavras, o efeito que as palavras similares em termos ortográficos desencadeiam no processo de reconhecimento de um estímulo-alvo é conhecido por *efeito de vizinhança ortográfica* e tem sido questão bastante estudada na literatura sobre o tópico (Andrews, 1989, 1992, 1997; Carreiras, Perea & Grainger, 1997; Coltheart, Davelaar, Jonasson & Besner, 1977; Grainger, O'Regan, Jacobs & Segui, 1989; Huntsman & Lima, 1996, 2002; Justi & Pinheiro, 2006, 2008; Perea, Carreiras & Grainger, 2004; Perea & Pollatsek, 1998; Perea & Rosa, 2000; Sears, Hino & Lupker, 1995; Snodgrass & Minzer, 1993; entre outros).

A principal tentativa de se definir as palavras que são similares em termos ortográficos consiste em considerar, como vizinhos ortográficos de uma palavra, qualquer palavra que possa ser criada ao se mudar uma letra dessa palavra, enquanto se preservam as posições das demais letras. Desse modo, a palavra ‘alma’ teria cinco vizinhos ortográficos (‘arma’, ‘asma’, ‘alga’, ‘alça’ e ‘alta’). Essa pode ser considerada a definição mais tradicional de uma palavra vizinha (ou similar em sua forma ortográfica) e é atribuída a Coltheart e cols. (1977). O número de palavras que pode ser criado, utilizando esse processo proposto por Coltheart e cols., é chamado de medida N (de *neighborhood*, em inglês) e se refere ao número de vizinhos ortográficos da palavra-alvo. Além dessa medida, nos estudos sobre o efeito de vizinhança ortográfica, também se tem utilizado a medida NF (de *neighborhood frequency*, em inglês) que se refere à frequência de ocorrência dos vizinhos da palavra-alvo, ou seja, se ela possui vizinhos de frequência maior do que a dela ou não (Grainger & cols., 1989). No caso de NF, pode-se dizer que a palavra ‘anão’ tem um vizinho ortográfico de maior frequência que é a palavra ‘ação’. A operacionalização dessas medidas (N e NF) tem o intuito de permitir aos pesquisadores estabelecer se o efeito de vizinhança ortográfica seria facilitador ou inibitório. Assim sendo, se o efeito de vizinhança ortográfica for facilitador, então o reconhecimento de palavras com um grande número de vizinhos ortográficos, ou com pelo menos um vizinho de maior frequência, será realizado com maior acurácia e rapidez; se o efeito for inibitório, o contrário se verificará, ou seja, as palavras nessa condição serão reconhecidas mais lentamente e com menor acurácia.

Devido à importância teórica dos efeitos de vizinhança ortográfica para modelos de acesso lexical, existem atualmente na literatura diferentes hipóteses explicativas para os efeitos de N e de NF. Alguns autores têm proposto que o efeito dessas variáveis pode ser modulado pelo uso de estratégias (Grainger & Jacobs, 1996; Carreiras e cols., 1997; Perea e Rosa, 2000); outros apontam que diferenças no grau de consistência do mapeamento grafema-fonema de uma determinada língua podem influenciar no efeito dessas variáveis (Andrews, 1997; Ziegler & Perry, 1998); alguns levantam a hipótese de que é a distribuição estatística das variáveis N e NF nas palavras de uma determinada língua que pode modular os efeitos dessas variáveis (Siakaluk, Sears & Lupker, 2002); e por fim, alguns questionam a própria definição de vizinhança ortográfica (Grainger, 2008). Considerando-se essas diferentes hipóteses explicativas para as diferenças encontradas nos resultados dos estudos sobre vizinhança ortográfica, o presente trabalho teve como objetivo investigar as duas últimas, e consequentemente, elas serão mais bem detalhadas a seguir.

De acordo com Siakaluk e cols. (2002), uma possível fonte de diferenças entre as línguas nos efeitos de N e NF pode advir de diferenças na estrutura da vizinhança ortográfica dessas línguas, isso é, diferentes distribuições de N e NF nas palavras das línguas. Mais especificamente, Siakaluk e cols. (2002) argumentam que, como mais da metade das palavras da língua inglesa de três a cinco letras têm vizinhos ortográficos de maior frequência de ocorrência, seria bastante contraintuitivo imaginar um mecanismo de processamento da informação que prejudicasse o desempenho na leitura da maioria das palavras. Ora, o que Siakaluk e cols.(op.cit.) parecem estar argumentando é que, de alguma forma, a própria distribuição estatística de N e NF nas palavras da língua inglesa pode levar o nosso sistema cognitivo a tirar proveito dessas variáveis e, no caso do inglês, levar a um efeito facilitador. Dessa forma, uma hipótese diretamente relacionada seria a de que, em uma língua em que a estrutura da vizinhança ortográfica fosse diferente e a maioria das palavras não tivesse vizinhos de maior freqüência de ocorrência, o efeito de N e NF deveria ser diferente do encontrado na língua inglesa (sendo, provavelmente, inibitório). Essa hipótese torna importante que se compare a estrutura da vizinhança ortográfica das palavras de diferentes línguas, para se averiguar se diferenças nessas distribuições também poderiam explicar as variações encontradas nos resultados das pesquisas (Andrews, 1989, 1992; Carreiras & cols., 1997; Justi & Pinheiro, 2006, 2008; Sears & cols., 1995; Siakaluk & cols., 2002).

Outra questão que se coloca é a da própria definição de vizinhança ortográfica, uma vez que para autores como Grainger (2008), as definições de N e NF (propostas respectivamente por Coltheart e cols., 1977; e, Grainger e cols., 1989) seriam apenas "boas aproximações" do conceito de similaridade ortográfica. Mais especificamente, para Grainger e Whitney (2004) e Whitney e Lavidor (2005) as definições tradicionais de vizinhança ortográfica (N e NF) pecam por não considerar como vizinhos ortográficos as palavras formadas pela transposição de duas letras (p.ex.: 'acesos' e 'acesso'), conhecidas como *transposed letter neighbors* – TLN, em inglês. Como existe certa evidência de que palavras com vizinhos transpostos são reconhecidas mais lentamente do que palavras sem vizinhos transpostos (Andrews, 1996), há a possibilidade de que os efeitos de N e NF sejam meros subprodutos dos efeitos de TLN. Uma forma de se averiguar essa questão é observar até que ponto N e NF se correlacionam com TLN. Afinal, se a correlação entre essas variáveis e TLN for muito pequena, dificilmente poder-se-ia explicar os efeitos encontrados de N e NF com base em sua correlação com TLN. Colocando de outra forma, se N e NF não se correlacionam com TLN, deve-se admitir que tais variáveis podem ter efeitos genuínos e independentes de TLN e que, muito provavelmente, são esses efeitos que têm sido relatados nos estudos de vizinhança ortográfica.

Tendo-se essas duas hipóteses em mente, o presente trabalho objetivou investigá-las através de comparações entre a distribuição das variáveis de vizinhança ortográfica no português do Brasil e na língua inglesa. A comparação entre essas duas línguas é interessante, tendo em vista que os resultados de estudos sobre o efeito de vizinhança ortográfica realizados com falantes do português brasileiro têm apresentado efeitos inibitórios de NF e N (Justi & Pinheiro, 2006, 2008), o que conflita com estudos realizados com falantes do inglês (Andrews, 1989, 1992; Sears & cols., 1995; Huntsman & Lima, 2002; Siakaluk & cols., 2002).

Dessa forma, os objetivos específicos do presente estudo foram:

a) avaliar a hipótese de Siakaluk e cols. (2002) de que diferenças na distribuição de N e NF poderiam explicar diferenças nos efeitos dessas variáveis;

b) considerando-se a hipótese de Grainger (2008), avaliar, se no caso do português do Brasil, N e NF se correlacionam com TLN, a ponto de se poder questionar se os primeiros teriam efeitos genuínos.

Método

Existem, pelo menos, duas listas de contagem de frequência de ocorrência 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, porque, em um estudo piloto, os índices de frequência de ocorrência das palavras dessa lista apresentaram uma maior correlação com o julgamento subjetivo de frequência (índice de familiaridade – Balota, Piloti & Cortese, 2001) relatado por um grupo de universitários (de todos os estudos citados sobre vizinhança ortográfica nas revisões da literatura efetuadas por Andrews em 1997 e Perea & Rosa em 2000, apenas dois estudos não foram realizados com universitários). Tendo sido selecionada a lista de palavras do corpus NILC (2005), outra questão que se tornou de fundamental importância no presente estudo foi o cômputo de estatísticas quanto à distribuição dos índices de vizinhança ortográfica nas palavras do português do Brasil. A finalidade dessas análises foi permitir a comparação da estrutura da vizinhança ortográfica das palavras do português com a das palavras da língua inglesa e investigar as possíveis correlações entre as diferentes medidas de vizinhança ortográfica no português do Brasil.

Material

Foram utilizadas no presente estudo 8465 palavras de quatro a seis letras do corpus NILC (2005).

Procedimentos

Foram selecionadas para a análise, inicialmente, todas as palavras do corpus NILC (2005) que tinham de quatro a seis letras (de acordo com Andrews, 1997, a maioria dos estudos sobre vizinhança ortográfica foi realizada com essa classe de palavras). Então, a frequência de ocorrência bruta relatada para essas palavras pelo corpus NILC foi convertida em um índice que expressa a frequência de ocorrência por milhão de palavras e foram selecionadas apenas as palavras com frequência de, pelo menos, uma ocorrência por milhão. Posteriormente, todas as palavras estrangeiras e hifenizadas foram descartas e a análise final foi desenvolvida com base nas 8465 palavras restantes. Essas palavras foram organizadas em um banco de dados e o programa '*N_Watch*' (Davis, 2005) foi utilizado como auxílio para o cômputo das seguintes estatísticas: FLog – contrapartida logarítmica ($^{\log_{10}(x+1)}$) da frequência por milhão de ocorrências da palavra; N – número de vizinhos ortográficos da palavra; NPV ou *spread* – número de posições da palavra em que, alterando-se uma letra, vizinhos ortográficos são gerados (p.ex. 'missa' tem um NPV igual a dois, pois tem vizinhos formados por mudança na segunda e terceira letras, 'massa' e 'mista', respectivamente); NVmF – número de vizinhos de menor frequência que a palavra; NF – número de vizinhos de maior frequência de ocorrência que a palavra; P_NF – presença (1) ou ausência (0) de, pelo menos, um vizinho de maior frequência; e, P_TLN – presença (1) ou ausência (0) de, pelo menos, um vizinho ortográfico formado pela transposição de duas letras da palavra.

Resultados

Para facilitar a visualização e a comparação com os dados existentes para a língua inglesa, optou-se por apresentar estatísticas descritivas para o total de palavras, e também estatísticas separadas para palavras de quatro, cinco e seis letras (como no trabalho de Andrews, 1997). O mesmo ocorreu para as análises de correlação. A tabela 1 apresenta as estatísticas descritivas para o total de 8465 palavras analisadas.

	Média	Desvio-Padrão
Frequência de ocorrência por milhão de palavras (FpM)	31,02	173,82
Frequência Logarítmica (FLog)	0,85	0,60
Número de letras (L)	5,39	0,71
Número de vizinhos ortográficos (N)	2,76	2,95
Número de posições em que se geram vizinhos (NPV)	1,65	1,28
Número de vizinhos de menor frequência (NVmF)	1,52	2,01
Número de vizinhos de maior frequência (NF)	1,24	1,8
Porcentagem com vizinhos de maior frequência (%NF)	54,11	–
Porcentagem com vizinhos formados pela transposição de duas letras (%TLN)	3,86	–

Tabela 1. *Estatísticas Descritivas de 8465 Palavras de Quatro a Seis Letras Retiradas do Corpus NILC (2005)*

No que diz respeito às principais variáveis de vizinhança ortográfica, pode-se observar na tabela 1 que as palavras do português brasileiro de quatro a seis letras têm, em média, aproximadamente três vizinhos ortográficos (N) e um vizinho ortográfico de maior freqüência de ocorrência (NF), sendo que, geralmente, os vizinhos das palavras são gerados por mudanças em uma ou duas posições das palavras (NPV). Além disso, pelo menos 54% das palavras têm vizinhos ortográficos de maior frequência de ocorrência (%NF), enquanto, aproximadamente, apenas 4% do total das palavras de quatro a seis letras têm vizinhos formados pela transposição de duas letras (%TLN). No caso da língua inglesa, Andrews (1997) apresentou apenas estatísticas específicas pelo número de letras, mas como foi apresentado o número de palavras de quatro, de cinco e de seis letras, nas quais os dados se basearam, tornou-se possível o cálculo da média ponderada para tais estatísticas de vizinhança ortográfica à exceção de %TLN, já que Andrews não apresentou dados quanto à porcentagem de palavras que têm vizinhos ortográficos formados pela transposição de duas letras na língua inglesa. Sendo assim, a média ponderada para N, NF, NPV e %NF na língua inglesa é respectivamente de: 2,81 vizinhos ortográficos; 1,33 vizinhos de maior frequência de ocorrência; 1,53 posições, em média, nas quais os vizinhos podem ser gerados; e, 47,85% das palavras de quatro a seis letras têm vizinhos ortográficos de maior frequência de ocorrência. As análises de correlação para o total das palavras de quatro a seis letras do português do Brasil encontram-se na tabela 2.

	FpM	FLog	L	N	NPV	NVmF	NF
FLog	0,44						
L	-0,10	-0,14					
N	0,08	0,15	-0,49				
NPV	0,06	0,15	-0,33	0,82			
NVmF	0,19	0,42	-0,37	0,80	0,68		
NF	-0,08	-0,23	-0,39	0,74	0,59	0,20	
TLN	0,03	0,03	-0,12	0,10	0,11	0,07	0,08

Nota. FpM = frequência de ocorrência por milhão de palavras; FLog = frequência logarítmica; L = número de letras; N = número de vizinhos ortográficos; NPV = número de posições em que podem ser gerados vizinhos; NVmF = número de vizinhos de menor frequência; NF = número de vizinhos de maior frequência; TLN = presença de um vizinho ortográfico formado pela transposição de duas letras. Todas as correlações foram significantes para $p \leq 0,05$.

Tabela 2. *Correlações Entre Medidas de Vizinhança Ortográfica, Frequência e Número de Letras para 8465 Palavras do Português Brasileiro*

Como pode ser observado na tabela 2, todas as correlações foram significantes. Uma análise cuidadosa, no entanto, indica que isso se deveu mais ao tamanho da amostra do que, propriamente, à força das correlações. As correlações entre N e TLN e entre NF e TLN foram ínfimas: 0,1 para a correlação entre N e TLN e 0,08 para a correlação entre NF e TLN. Dessa maneira, apenas 1% da variação em N pode ser explicada pela variação em TLN, e apenas 0,64 % da variação em NF pode ser explicada pela variação em TLN. O número de vizinhos ortográficos (N) apresentou uma forte correlação positiva com o número de vizinhos ortográficos de maior frequência (NF). O número de posições em que se podem gerar vizinhos (NPV) apresentou uma forte correlação positiva com N e uma correlação positiva moderada com NF. O número de letras apresentou uma correlação negativa moderada com N e uma correlação negativa, um pouco menor, com NF. Já a frequência logarítmica apresentou uma correlação positiva moderada com o número de vizinhos de menor frequência (NVmF) e uma pequena correlação negativa com o número de vizinhos de maior frequência (NF).

A visualização comparativa dos dados estatísticos da estrutura da vizinhança ortográfica das palavras do português do Brasil e das palavras da língua inglesa está organizada na tabela 3. Esta tabela provê as estatísticas descritivas da estrutura da vizinhança ortográfica das duas línguas por número de letras, cujos dados da língua inglesa foram extraídos do trabalho de Andrews (1997). Além disso, como no trabalho do autor citado não foram relatadas estatísticas quanto aos vizinhos que podem ser formados pela transposição de duas letras (TLN), nem quanto à frequência logarítmica, e como as estatísticas dessas variáveis já foram descritas nas análises gerais (ver tabela 1 e tabela 2), optou-se por não incluir estatísticas para essas variáveis na tabela 3.

	Português	Inglês
Palavras de quatro letras	1106 palavras	1895 palavras
	Média (DP)	Média (DP)
Frequência de ocorrência por milhão de palavras (FpM)	73,0 (415,0)	100,3 (549,8)
Número de vizinhos ortográficos (N)	6,0 (4,4)	7,2 (4,9)
Número de posições em que se geram vizinhos (NPV)	2,4 (1,2)	2,5 (1,1)
Número de vizinhos de menor frequência (NVmF)	3,2 (3,2)	3,4 (3,8)
Número de vizinhos de maior frequência (NF)	2,8 (2,9)	3,5 (3,4)
Porcentagem com vizinhos de maior frequência (%NF)	74,5	80,3
Palavras de cinco letras	2968 palavras	2895 palavras
	Média (DP)	Média (DP)
Frequência de ocorrência por milhão de palavras (FpM)	33,3 (107,5)	34,2 (153,3)
Número de vizinhos ortográficos (N)	3,3 (2,8)	2,4 (2,3)
Número de posições em que se geram vizinhos (NPV)	2,0 (1,3)	1,5 (1,2)
Número de vizinhos de menor frequência (NVmF)	1,8 (2,0)	1,1 (1,6)
Número de vizinhos de maior frequência (NF)	1,5 (1,8)	1,1 (1,5)
Porcentagem com vizinhos de maior frequência (%NF)	62,1	52,0
Palavras de seis letras	4391 palavras	4166 palavras
	Média (DP)	Média (DP)
Frequência de ocorrência por milhão de palavras (FpM)	18,9 (80,6)	16,5 (62,6)
Número de vizinhos ortográficos (N)	1,6 (1,6)	1,1 (1,6)
Número de posições em que se geram vizinhos (NPV)	1,2 (1,1)	0,8 (0,9)
Número de vizinhos de menor frequência (NVmF)	0,9 (1,2)	0,5 (1,1)
Número de vizinhos de maior frequência (NF)	0,7 (1,0)	0,5 (1,0)
Porcentagem com vizinhos de maior frequência (%NF)	43,6	30,2

Nota. Os dados da língua inglesa foram retirados do trabalho de Andrews (1997).

Tabela 3. *Estatísticas Descritivas de Medidas de Vizinhança Ortográfica e de Frequência de Ocorrência para Palavras do Português do Brasil e do Inglês*

Pode-se observar na tabela 3 que as estatísticas descritivas de vizinhança ortográfica para as duas línguas são razoavelmente similares. Contudo, devido ao tamanho excessivamente grande das amostras ao se calcular testes *T* para amostras independentes com base nos dados fornecidos, praticamente todas as médias diferiram significativamente. As únicas médias que não diferiram significativamente foram as de frequência de ocorrência (devido aos grandes desvios-padrão) e as do número de vizinhos de menor frequência (NVmF) para palavras de quatro letras. Em situações como essas (em que mesmo pequenas diferenças são estatisticamente significativas

devido ao tamanho da amostra), a magnitude das diferenças acaba sendo mais informativa do que os valores de *p* (Cohen, 1994). Destarte, optou-se por calcular o *d* de Cohen (1988) como uma estimativa da magnitude da diferença entre as médias. Com base nos valores de *d*, todas as diferenças entre as médias ficaram abaixo do valor de 0,5, critério proposto por Cohen (1988) para que uma diferença seja considerada razoável (média). As maiores diferenças entre as línguas ocorreram para as palavras de cinco e de seis letras e foram entre o número de posições em que se geram vizinhos (NPV). Em ambos os casos o valor de *d* foi de 0,4. Outras diferenças que merecem menção ocorreram entre o número de vizinhos de menor freqüência (NVmF) e o número de vizinhos ortográficos (N), também no caso das palavras de cinco letras e de seis letras. Para palavras de cinco letras, o valor de *d* para a diferença em NVmF entre as línguas foi de 0,39 e para a diferença em N foi de 0,35. No caso das palavras de seis letras, o valor de *d* para a diferença em NVmF entre as línguas foi de 0,35 e para a diferença em N foi de 0,31. A diferença entre o número de vizinhos de maior frequência de ocorrência (NF) entre as línguas manteve-se razoavelmente constante, independentemente do número de letras (valores *d* de 0,22, 0,24 e 0,20 para palavras de quatro, cinco e seis letras, respectivamente). No que diz respeito às palavras de quatro letras, o valor de *d* para a diferença entre as línguas no número de vizinhos ortográficos (N) foi de 0,26 que é um valor considerado pequeno, enquanto as demais diferenças foram ínfimas ($d < 0,1$). Por fim, Andrews (1997) não relatou o desvio-padrão para a porcentagem de palavras com vizinhos de maior frequência de ocorrência (%NF) em seu estudo, fato que levou à omissão de tal dado na tabela 3. Utilizando, porém, o desvio-padrão calculado com base nos dados do presente estudo, os valores de *d* para as diferenças entre as línguas em %NF foi pequeno (0,13 para palavras de quatro; 0,21 para palavras de cinco e 0,27 para palavras de seis letras).

A tabela 4 apresenta as correlações entre as diferentes medidas de vizinhança ortográfica e frequência de ocorrência no português do Brasil e no inglês, de acordo com o número de letras. Os dados da língua inglesa encontram-se entre parênteses e foram retirados do trabalho de Andrews (1997) que não relatou quais correlações foram significantes ou não. De qualquer forma, devido ao tamanho das amostras, é esperado que todas as correlações relevantes tenham sido significantes, pois que, no caso das palavras do presente estudo, só não foram significantes as correlações menores que 0,042, no caso das palavras de quatro letras, e, as menores que 0,03, no caso das palavras de seis letras.

		FpM	FLog	N	NPV	NVmF
FLog	4 letras	0,45 (0,45)				
	5 letras	*0,65 (0,52)*				
	6 letras	**0,54 (0,62)**				
N	4 letras	0,04 (0,03)	0,07 (0,24)			
	5 letras	*0,04 (0,05)*	*0,11 (0,20)*			
	6 letras	**0,00 (0,01)**	**0,09 (0,08)**			
NPV	4 letras	0,04 (0,03)	0,10 (0,21)	0,80 (0,72)		
	5 letras	*0,05 (0,03)*	*0,12 (0,16)*	*0,85 (0,82)*		
	6 letras	**0,01 (0,01)**	**0,09 (0,09)**	**0,92 (0,87)**		
NVmF	4 letras	0,18 (0,19)	0,48 (0,70)	0,74 (0,67)	0,62 (0,49)	
	5 letras	*0,22 (0,18)*	*0,44 (0,55)*	*0,77 (0,69)*	*0,66 (0,54)*	

	6 letras	**0,10 (0,14)**	**0,32 (0,37)**	**0,79 (0,70)**	**0,74 (0,60)**	
	4 letras	-0,14 (-0,15)	-0,42 (-0,36)	0,70 (0,65)	0,54 (0,45)	0,04 (-0,11)
NF	*5 letras*	*-0,18 (-0,11)*	*-0,32 (-0,21)*	*0,69 (0,70)*	*0,58 (0,58)*	*0,07 (0,01)*
	6 letras	**-0,11 (-0,10)**	**-0,24 (-0,18)**	**0,67 (0,73)**	**0,60 (0,64)**	**0,08 (0,09)**

Nota. FpM = frequência de ocorrência por milhão de palavras; FLog = frequência logarítmica; N = número de vizinhos ortográficos; NPV = número de posições em que podem ser gerados vizinhos; NVmF = número de vizinhos de menor frequência; NF = número de vizinhos de maior frequência. Os dados referentes à língua inglesa encontram-se entre parênteses e foram extraídos do trabalho de Andrews (1997).

Tabela 4. *Correlações entre Medidas de Vizinhança Ortográfica e Frequência, para Palavras de 4 letras, 5 letras (Itálico) e 6 Letras (Negrito) do Português Brasileiro e do Inglês (Entre Parênteses)*

De uma forma geral, as correlações entre as variáveis de vizinhança ortográfica no português do Brasil e no inglês foram muito similares, apresentando quase sempre a mesma magnitude e o mesmo padrão. Para ambas as línguas, as maiores correlações ocorreram entre o número de vizinhos ortográficos (N) e o número de posições onde se geram vizinhos (NPV), seguidas da correlação entre N e o número de vizinhos de menor frequência (NVmF) e da correlação entre N e o número de vizinhos de maior frequência de ocorrência (NF). No que diz respeito à correlação das medidas de vizinhança ortográfica com a frequência de ocorrência das palavras do português do Brasil, o número de vizinhos de menor frequência de ocorrência (NVmF) apresentou uma correlação positiva com a frequência logarítmica (FLog) que foi de moderada (para palavras de quatro e cinco letras) à baixa (para palavras de seis letras), enquanto o número de vizinhos de maior frequência de ocorrência (NF) apresentou uma correlação negativa com FLog que também foi de moderada (para palavras de quatro letras) à baixa (para palavras de cinco e seis letras). No caso das palavras da língua inglesa, as correlações apresentaram o mesmo padrão (diminuíram com o aumento no número de letras), mas apresentaram maior magnitude que as correlações da língua portuguesa nas correlações entre NVmF e FLog e menor magnitude nas correlações de NF e FLog. Por fim, para ambas as línguas as correlações entre o número de vizinhos ortográficos (N) e a frequência logarítmica (FLog) e as correlações entre o número de posições em que vizinhos podem ser gerados (NPV) e FLog foram pequenas.

Discussão

Considerando-se inicialmente a questão da definição de vizinhança ortográfica, torna-se necessário que se considere a relação entre N, NF e TLN. Como pode ser observado nas análises desenvolvidas (especialmente, as referentes às tabelas 1 e 2), as correlações entre N e TLN e entre NF e TLN são ínfimas, o que torna muito difícil a proposta de que TLN possa explicar os efeitos de N e NF. Dessa feita, é provável que os efeitos de N e NF não sejam relacionados aos efeitos de TLN. Outra questão que coloca em dificuldades a proposta de se considerar TLN uma "melhor" medida de vizinhança ortográfica é que apenas 3,86% das palavras de quatro a seis letras no português do Brasil têm vizinhos formados pela transposição de duas letras, enquanto mais de 54% dessa mesma classe de palavras têm vizinhos ortográficos de acordo com a métrica tradicional (Coltheart e cols., 1977), sejam eles de maior frequência

de ocorrência ou não. Desse modo, no caso do português do Brasil, pode-se concluir que os efeitos de N e NF são muito provavelmente independentes dos efeitos de TLN e que, caso essas variáveis tenham algum efeito real (o que parece ser o caso de acordo com os trabalhos de Justi & Pinheiro, 2006, 2008), elas afetam um grupo muito mais amplo de palavras do que TLN. Portanto, no caso do português do Brasil, pode-se considerar N e NF medidas mais relevantes de vizinhança ortográfica, pelo menos, de um ponto de vista empírico.

Outra questão que merece análise é o quanto as diferenças na distribuição estatística de N e NF na língua portuguesa poderiam explicar diferenças nos efeitos dessas variáveis em relação à língua inglesa. O argumento levantado por Siakaluk e cols. (2002) é o de que, em uma língua em que a maioria das palavras tem vizinhos ortográficos de maior frequência de ocorrência, seria contraintuitivo o desenvolvimento de um mecanismo de processamento da informação que prejudicasse a identificação da maioria das palavras da língua. A ideia de Siakaluk e cols. é que, como a maioria das palavras do inglês tem vizinhos ortográficos de maior frequência (note-se que ao se considerar as palavras de quatro a seis letras a porcentagem de palavras da língua inglesa com vizinhos de maior freqüência cai para 47,85%, mas, de qualquer forma, continua bastante alta), o efeito de NF deveria ser facilitador e não inibitório. Em certo sentido, pode-se dizer que este estudo trouxe evidências que entram em conflito com a proposta de Siakaluk e cols., visto que a diferença entre a língua portuguesa e a língua inglesa foi sempre muito pequena no que diz respeito à distribuição estatística das variáveis de vizinhança ortográfica, e as correlações entre essas variáveis tenderam a apresentar a mesma magnitude e padrão em ambas as línguas (ver tabelas 3 e 4, por exemplo). É possível argumentar, porém, que, embora pequenas, existem diferenças entre a distribuição dos vizinhos ortográficos nas línguas em questão. Por exemplo, ainda que nas palavras de quatro letras as distribuições das variáveis de vizinhança ortográfica se sobreponham consideravelmente (diferenças em que o valor de *d* é de até 0,2, indicam uma sobreposição de 85% nas distribuições – Cohen, 1988), no caso das palavras de cinco e seis letras, a distribuição das variáveis N, NPV e NVmF apresentaram uma menor sobreposição.

Esse seria um argumento até aceitável se não fossem essas diferenças na direção oposta à predita por Siakaluk e cols. (2002). Afinal essas diferenças indicam que as palavras do português do Brasil têm mais vizinhos ortográficos do que as palavras da língua inglesa. A predição de Siakaluk e cols.(op.cit.), contudo, diz respeito mais diretamente à proporção de palavras que tem vizinhos ortográficos de maior frequência de ocorrência e, no caso dessa variável, (%NF na tabela 3) o valor de *d* médio foi de aproximadamente 0,2 indicando uma sobreposição muito alta das distribuições e, novamente, na direção oposta à predita. Destarte, o principal problema para o argumento de Siakaluk e cols.(op.cit.) é que, no caso do português do Brasil, a maioria das palavras (54,11%) também tem vizinhos ortográficos de maior frequência de ocorrência e, ainda assim, o efeito de NF parece ser inibitório (Justi & Pinheiro, 2006; 2008). Portanto, fica difícil sustentar que diferenças na proporção de palavras com vizinhos de maior frequência de ocorrência no português e no inglês poderiam explicar as diferenças nos efeitos de vizinhança ortográfica. Além disso, fica a lição de que nem sempre o nosso sistema

cognitivo opera de maneira ótima e de acordo com nossas intuições, pois, no caso do português do Brasil, esse sistema parece prejudicar o processamento da maioria das palavras.

Conclusão

O presente estudo avaliou uma possível explicação para diferenças encontradas entre diferentes línguas nos efeitos de vizinhança ortográfica. Essa avaliação foi apenas parcial, uma vez que a comparação traçada foi entre o português do Brasil e a língua inglesa. No entanto, essa comparação é bastante relevante, pois estudos realizados nas duas línguas em questão têm produzido resultados opostos para os efeitos de N e de NF. A despeito da proposta de Siakaluk e cols. (2002) de que seria contraintuitivo que em línguas em que a maioria das palavras tem vizinhos ortográficos de maior frequência de ocorrência NF apresentasse um efeito inibitório, esse parece ser o caso no português do Brasil. Por fim, as análises desenvolvidas evidenciaram que tanto no português quanto no inglês a maioria das palavras tem vizinhos de maior frequência de ocorrência e, por isso, diferenças no efeito de NF entre as línguas não podem ser explicadas supondo-se diferentes proporções de palavras com vizinhos de maior frequência nas duas línguas. Além disso, as análises desenvolvidas não permitem afirmar que existem grandes diferenças entre o português do Brasil e o inglês, no que diz respeito às outras variáveis de vizinhança ortográfica, o que torna a proposta de que diferenças na distribuição estatística dessas variáveis entre as línguas poderiam modular o efeito de vizinhança ortográfica bastante difícil de ser sustentada.

Outra preocupação deste trabalho foi quanto à possibilidade dos efeitos de N e de NF serem modulados pela relação dessas variáveis com o número de vizinhos transpostos. No caso do português do Brasil, as análises desenvolvidas evidenciaram que a correlação entre N e TLN e entre NF e TLN foi ínfima, o que torna muito improvável que os efeitos de N e NF sejam derivados do efeito do número de vizinhos transpostos. Soma-se a isso que, de um ponto de vista empírico e prático, N e NF parecem ser variáveis mais importantes, uma vez que mais de 54% das palavras têm vizinhos ortográficos enquanto apenas 4% das palavras têm vizinhos formados pela transposição de duas letras. Dessa forma, mesmo que se considerem N e NF como apenas "uma boa primeira aproximação" (Grainger, 2008) do conceito de similaridade ortográfica, no caso do português do Brasil, talvez essas variáveis sejam a aproximação mais relevante.

Além das hipóteses investigadas, é interessante ressaltar que o presente estudo traz dados importantes sobre características psicolinguísticas de uma amostra bastante razoável de palavras da língua portuguesa. Assim sendo, ele contribui diretamente para estudos futuros que queiram investigar efeitos de vizinhança ortográfica ou para o controle experimental dessas variáveis quando estas não forem o foco da pesquisa (os dados de vizinhança ortográfica que foram gerados para as 8465 palavras podem ser obtidos mediante contato por correio eletrônico com os autores desse estudo).

Por fim, é importante considerar também que o presente estudo investigou apenas possíveis diferenças entre a língua inglesa e o português do Brasil no que diz respeito à distribuição das variáveis de vizinhança ortográfica. Mas, outra fonte de diferenças relevantes para o efeito de vizinhança ortográfica diz respeito a diferenças na consistência do mapeamento grafema-fonema entre as línguas (Andrews, 1997; Ziegler & Perry, 1998). Esse tipo

de análise vai muito além dos objetivos do presente estudo, mas seria interessante que análises de corpus linguístico como a realizada neste trabalho fossem desenvolvidas tendo como foco essa questão. Por exemplo, qual a proporção de palavras no português e no inglês que tem um mapeamento grafema-fonema irregular ou inconsistente? Essa proporção varia de acordo com o número de letras e/ou com a frequência das palavras? Existiria uma relação entre vizinhança ortográfica e consistência no mapeamento grafema-fonema? Embora possam ser traçadas diferenças entre as línguas (Seymour, 2005), enquanto essas não forem mais bem qualificadas empiricamente, talvez seja difícil testarem-se hipóteses mais específicas quanto à relevância dessas diferenças.

Referências

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: activation or search? *Journal of Experimental Psychology: Learning, Memory & Cognition*, 15, 802-814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 18, 234-254.

Andrews, S. (1996) Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, 35, 775-800.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4, 439-461.

Balota, D.; Pilotti, M.; & Cortese, M. (2001) Subjective frequency estimates for 2938 monosyllabic words. *Memory & Cognition*, 29, 639-647.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 23, 857-871.

Cohen, J. (1998) *Statistical Power for Behavioral Sciences* 2nd ed. New York: Academic Press.

Cohen, J. (1994) The earth is round ($p < .05$). *American Psychologist*, v.49, p.997-1003.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In: S. Dornic (Ed.) *Attention and performance VI* (pp. 535-555). Hillsdale: Erlbaum.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37, 65-70.

Grainger, J. (2008). Cracking the orthographic code: an introduction. *Language and Cognitive Processes*, 23, 1-35.

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., O'Reagan, K., Jacobs, A., & Segui, J. (1989). On the role of competing word units in visual word recognition: the neighborhood frequency effect. *Perception & Psychophysics*, 45, 189-195.

Huntsman, L., & Lima, S. (1996). Orthographic neighborhood structure and lexical access. *Journal of Psycholinguistic Research*, 25, 417-429.

Huntsman, L., & Lima, S. (2002). Orthographic neighbors and visual word recognition. *Journal of Psycholinguistic Research*, 31, 289-306.

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. & Pinheiro, A. (2008). O efeito de vizinhança ortográfica em crianças brasileiras: estudo com a tarefa de decisão lexical. *Interamerican Journal of Psychology*, 42, 559-569.

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

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception & Performance*, 24,767-779.

Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: a review. *Psicológica*, 21, 327-340.

Perea, M., Carreiras, M., & Grainger, J. (2004). Blocking by word frequency and neighborhood density in visual word recognition: a task-specific response criteria account. *Memory & Cognition*, 32, 1090-1102.

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 faixe pré-escolar e séries iniciais do 1^{o} grau*. São Paulo: Associação Brasileira de Dislexia.

Sears, C., Hino, Y., & Lupker, S. (1995). Neighborhood size and neighbourhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, 16, 65-76.

Seymour, P. (2005). Early Reading Development in European Orthographies. In: M. Snowling & C. Hulme (Eds.) *The Science of Reading: A Handbook* (pp. 296-315). Oxford: Blackwell.

Siakaluk, P., Sears, C., & Lupker, S. (2002) Orthographic neighborhood effects in lexical decision: the effects of nonword orthographic neighborhood size. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 661-681.

Snodgrass, J., & Minzer, M. (1993). Neighbourhood effects in visual word recognition: facilitatory or inhibitory? *Memory & Cognition*, 21, 247-266.

Whitney, C. & Lavidor, M. (2005). Facilitative orthographic neighborhood effects: the SERIOL model account. *Cognitive Psychology*, 51, 179-213.

Ziegler, J., & Perry, C. (1998). No more problems in Coltheart's neighborhood: resolving neighborhood conflicts in the lexical decision task. *Cognition*, 68, B53-B62.