XI Encontro Latino Americano de Iniciação Científica e VII Encontro Latino Americano de Pós-Graduação, da Univap 2007 (XI INIC/ VII EPG)

"A LINGÜÍSTICA DE CORPUS NA ANÁLISE DO INTERNETÊS"

Gonzalez, Z. M. Gutierrez [1], Prado, C. P., Berber Sardinha, A. P.[2]

[1]PUC-SP, LAEL.Rua Monte Alegre, 1800 - Perdizes, SP: zmgg@uol.com.br
[2]DERJ–DIRETORIA ENSINO JACAREÍ – Rua barão de Jacareí, 187 ciniraprado @terra.com.br
[3]PUC-SP, LAEL. Rua Monte Alegre 1800 - Perdizes, S.P: tony4@uol.com.br

Resumo: o presente trabalho pretende investigar a possível economia de toques (keystrokes) na grafia do internetês, por meio de um corpus constituído de blogs de jovens que utilizam a internet para a comunicação. Assim, submetemos o corpus à duas análises. Primeiro, investigamos a eliminação de toques na grafia dos itens em internetês, comparando-os com a grafia da norma padrão. Segundo, verificamos quantos toques comumente são utilizados na formação dos itens em internetês. Para tanto, recorremos aos preceitos teóricos e o intrumental metodológico oferecidos pela Lingüística de Corpus (Biber et al., 1998; Berber Sardinha, 2004); além disso, verificamos em que frequência ocorre essa possivel economia nos itens, nos servindo da visão sobre sistemas probabilísticos (Halliday, 1991,1992; Berber Sardinha, 200; Biber et al., 1998).

Palavra Chave: Lingüística de Corpus, internetês, toques (keystrokes);

Área do conhecimento: Língua Portuguesa, Lingüística de Corpus;

Introdução

Um fenômeno recente da língua portuguesa é o aparecimento de uma nova maneira de grafar palavras, muito usada em meios eletrônicos de comunicação, como chats, blogs e mensagens de texto via celular. O resultado desta comunicação rápida e instantânea é uma nova grafia, constituída de uma economia de caracteres digitados e uma despreocupação com as normas ortográficas e gramaticais da Língua Portuguesa. Esse fenômeno trata-se do internetês, termo que será utilizado no decorrer desse trabalho.

Alguns especialistas em estudos da linguagem encaram o internetês como uma variedade de língua, como tantas outras. Segundo Possenti (2006) "Uma coisa é a grafia, outra, a língua. Não há linguagem nova, só técnicas de abreviação. As soluções gráficas são até interessantes, pois a grafia cortada é a vogal". Isso significa que as técnicas de abreviação com eliminação de vogais e consoantes não comprometem a língua, que é formada por regras e leis combinatórias (sintaxe e gramática). Ao contrário, as abreviações trazem soluções práticas, com intuito de agilizar a comunicação. Por exemplo, a palavra 'gente' nos meios digitais é grafada como *gnt,* assim o nome da letra 'g' sonoriza /ge/ e o nome da letra 't' sonoriza /te/. "Dessa forma, ao suprimir principalmente as vogais, o nome das consoantes substitui o 'som' das vogais que não são escritas" (Possenti, 2006). Ao observarmos esse fenômeno, a impressão geral é a de que as pessoas que se utilizam do internetês o fazem com o intuito de digitar a menor quantidade possível de caracteres, possivelmente com a intenção de economizar tempo.

Neste contexto, propomo-nos investigar a economia de toques (keystrokes) na formação de palavras do internetês, focando o nível léxico-gramatical da língua através da exploração de um corpus constituído por blogs.

Dentre as várias definições de 'blog' obtidas, a que consideramos como a mais adequada para fins de uma pesquisa sobre o internetês é a de Schittine (2004:186): "o blog, ou seja, o diário íntimo na internet é um híbrido de vários tipos de escrita. Dividido entre vários estilos, ele se aproxima de uns, se afasta de outros, mas acaba tendo um pouco de cada um deles". Atualmente, os blogs são utilizados para anotações diversas, como poemas, críticas literárias, letras de música, exposição de idéias, opiniões políticas; enfim, qualquer texto que possa ser considerado dialógico no ambiente virtual. O gênero blog focado neste trabalho é um diário público, acessado por jovens que se comunicam, expressam seus sentimentos e pensamentos, bem como opinam sobre assuntos diversos.

Lingüística de Corpus

O desenvolvimento de pesquisas lingüísticas baseadas em corpus está intimamente ligado à evolução tecnológica, mais especificamente à existência do computador eletrônico. Assim, corpus é definido como: "um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise". (Sanches, 1995 apud Berber Sardinha, 2004: 18).

Para esse trabalho optamos por utilizar o arcabouço teórico e metodológico oferecidos pela Linguística de Corpus (doravante LC) uma área de estudo que segundo Berber Sardinha (2004:3): "ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística". Tal área, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador.

Portanto, a LC é uma área do conhecimento que estuda a linguagem por meio da utilização de grandes quantidades de dados empíricos relativos ao efetivo uso da linguagem, com o auxílio de computador.

A LC e a visão probabilística da Linguagem

Um dos maiores expoentes da visão probabilística é Halliday (1991, 1992), a partir de estudos sobre probabilidades em sistemas lingüísticos em meados de 1950, quando preparava uma gramática da língua chinesa. Dez anos mais tarde, o autor volta-se para a língua inglesa, levando em conta as probabilidades de ocorrência de traços gramaticais ao tomar como amostras 2.000 orações de diferentes registros. Seu principal objetivo era o de descrever os sistemas não somente como escolhas *a* ou *b* ou *c*, mas o de verificar como as escolhas *a* ou *b* ou *c* estão ligadas a certas probabilidades de ocorrência (Berber Sardinha, 2006). Para Halliday, era evidente que alguns traços do sistema, tais como as 'polaridades' negativa/positiva ou os 'tempos verbais' presente/passado, poderiam ser mensurados. As 2.000 orações tomadas como amostra não formaram, porém, um corpus suficientemente grande que permitisse a conclusão do seu trabalho. Para um resultado fidedigno, centenas de milhares de orações – e não somente milhares de orações – seriam necessárias.

Em 1991, alguns pesquisadores - entre eles Halliday - dedicaram-se à compilação do Cobuild, grande corpus eletrônico da língua inglesa, projeto desenvolvido pela Universidade de Birmimgham[1]. Finalmente, foi então possível concluir o trabalho que permanecera inacabado. Para tanto, Halliday comparou os resultados das 2.000 orações iniciais aos grandes dados constantes do corpus Cobuild. O resultado foi surpreendente pois, ao analisar as 'polaridades' no corpus não-eletrônico, o autor encontrou a razão 0.9 em sentenças positivas e 0.1 em negativas; no corpus eletrônico, em contrapartida, Halliday obteve 0.87 para as positivas e 0.13 para as negativas. Quanto à análise dos tempos verbais, o primeiro corpus revelou 0.5 sentenças no presente e 0.5 no passado, ao passo que, na amostra do Cobuild, chegou-se a 0.4955 no presente e a 0.5041 no passado, sendo que tais variações (diferenças) foram consideradas como diferenças não-significativas (Bod, 2003:12 *apud* Berber Sardinha, 2006). Tal resultado apontou para uma tendência de distribuição de probabilidades das polaridades correspondente a 0.9 para 0.1 (sentenças positivas e negativas), e a 0.5 para 0.5 para os tempos verbais. Halliday nomeou tais distribuições 'skew' (enviesada) para os valores 0.9 para 0.1, e 'equiprobable' (equiprovável) para os resultados próximos de 0.5 (Halliday 1993:9). Esses resultados também podem ser expressos em porcentagem (Berber, 2006 *apud* Bod, 2003:12). Desta forma, e tomando como exemplo os valores encontrados por Halliday (1993), as polaridades referentes às sentenças podem ser representadas como 90% positivas e 10% negativas; quanto aos tempos verbais, as polaridades revelam-se semelhantes, ou seja, 50% no presente e 50% no passado.

As pesquisas acima descritas mostram, dessa forma, que o fato de as probabilidades fazerem parte dos sistemas lingüísticos altera o entendimento do conceito de 'escolha' na língua em uso, revelando, em conseqüência, que a "livre escolha" é geralmente pouco provável (Halliday, 1993). Se assim fosse, os usuários de uma língua poderiam 'escolher' sentenças negativas às positivas, o que não ocorre normalmente. Halliday (1993) destaca, ainda, dois pontos importantes a serem considerados na pesquisa probabilística: (1) As probabilidades mudam de acordo com o tempo, e essa mudança ocorre diacronicamente, de forma dinâmica; (2) os padrões de probabilidade variam de acordo com o número de diferentes situações. Por exemplo, a probabilidade de alguns traços em registros específicos normalmente varia se for comparada à língua em sua totalidade.

[1] Grã-Bretanha.

As pesquisas realizadas por Halliday (1991,1992) somam-se inúmeras outras que confirmam o potencial das probabilidades no enriquecimento das descrições lingüísticas (Biber et al, 1998), refletido no desenvolvimento de vários projetos em andamento, muitos deles voltados para o processo de ensino e aprendizagem.

Metodologia

Com intuito de verificar até que ponto a abreviação de palavras efetivamente reverte em um maior ganho de tempo, realizamos os seguintes procedimentos; solicitamos à ferramenta computacional WordSmith (Scott, 1997), que nos fornecesse a lista de palavras contidas no corpus e suas respectivas freqüências. Esse procedimento possibilitou-nos a verificação de 138.021 *tokens*[2] e 15.552 *types*[3] no corpus. Para esse trabalho optamos por um 'recorte', analisando as 50 formas mais freqüentes do corpus de estudo. O próximo passo foi confrontá-las com o dicionário digitalizado 'Houaiss' com a finalidade de determinar quais palavras são grafadas de acordo com a norma padrão e quais palavras são grafadas em internetês. Porém, ao confrontarmos os itens sob investigação com o dicionário selecionado para a pesquisa, os resultados não foram satisfatórios, pois os sentidos de alguns itens encontrados no corpus não correspondiam aos sentidos apresentados em verbetes; para exemplificar, destacamos a palavra *'num'* (não), na qual em verbetes encontram-se acepções tais como: "contração da preposição 'em' + artigo indefinido 'um', encontrar-se *n*. lugar - indicação de lugar" (Houaiss, 2004). Desse modo, para garantirmos a fidelidade dos dados, recorremos à ferramenta Concord do WordSmth Tolls que nos forneceu os contextos de ocorrências das 50 palavras. Assim, pudemos determinar quais palavras apresentaram características do internetês e quais palavras pertenciam a norma padrão.

Após obtermos as palavras em internetês, despendemos duas análises.

Primeiro, contabilizamos quantos toques e, por conseguinte, quantas letras e/ou acentos gráficos foram subtraídos de algumas palavras. Para melhor ilustrar o procedimento, tomemos os itens 'vc' (você) como exemplo: se observarmos o item 'você' na grafia padrão, temos uma palavra formada por 4 letras: 'v' (1), 'o' (2), 'c' (3), 'e' (4), mais um acento gráfico **(^)**; em internetês, a economia é de três toques, ou seja, duas letras e um acento gráfico, resultando em: *'v'* (1) e *'c'* (2). Portanto, as quatro letras e um acento gráfico da norma-padrão (cinco toques) são reduzidos a duas letras (ou dois toques) em internetês. É importante destacar que em muitos itens do internetês não ocorrem a eliminação de letras/acento, apenas ocorrem modificações na grafia. Para ilustrar essas modificações, tomemos o item *'naum'* (não), no qual em ambas grafias, padrão e internetês, são escritas com 4 toques.

Segundo, decidimos determinar a quantidade de toques necessários à composição desses itens, assim, optou-se por levantar o número de toques comumente utilizados na formação dessas palavras. Para melhor exemplificar o procedimento, tomemos as palavras 'todo' e 'toda' as quais, na formação do item no internetês, são reduzidas a dois toques - *'td'*; em um outro exemplo, a palavra 'você' em internetês é representada por apenas dois toques - 'vc'. Nos exemplos selecionados, a economia na digitação resultou na formação de itens com dois toques.

Resultados

Na primeira análise, pudemos constatar que 43% (21/50) das palavras estão grafadas de acordo com a norma padrão; por outro lado, 57% (29/50) estão grafadas em internetês. Para esse trabalho, selecionamos apenas o foco da investigasção, ou seja, as 29 formas em internetês.

Desse modo, após despendermos as análises mencionadas acima obtivemos, primeiramente, os resultados sobre a quantidade de toques reduzidos na passagem da grafia padrão para o internetês, como segue abaixo:

Toques grafia padrão	Toques Internetês	Nº itens internetês	Toques Eliminados
1	1	2	0
2	2	1	0
3	3	1	0
4	4	1	0
4	3	4	1
3	2	5	1
2	1	4	1
3	1	1	2
4	2	5	2
5	3	3	2
5	2	1	3
6	2	1	4

Quadro I – redução de toques na grafia do internetês.

Os resultados mostraram uma freqüência majoritária de 47% (13/29) dos itens analisados, parece haver uma grande tendência para a eliminação de um toque ou de um acento/letra. Segundo estão à eliminação de dois toques ou

2 Tokens é o número de itens (ou ocorrências); por exemplo, a frase 'o João viu o Pedro' possui cinco itens: (1) o, (2) João, (3) viu, (4) o, (5) Pedro. Portanto, a frase possui 5 tokens (Berber Sardinha, 2004:94).

3 Types é o número de formas (ou vocábulos). Na frase 'o João viu o Pedro' há quatro formas: duas formas 'o', uma forma 'João', uma forma 'viu, uma foram 'Pedros'. Portanto a frase possui 4 types (BerberSardinha, 2004:94).

duas letras/acento em 32% (9/29). Porém, dois casos apresentaram baixa freqüência de subtração de letras/acento, ou seja, a eliminação de 3 toques resultou 1% (1/29) e a eliminação de 4 toques 1% (1/29). Vale ressaltar que não houve eliminação de toques em 18% (5/29) dos itens, conforme mencionado anteriormente, a grafia do internetês em alguns itens não implica, efetivamente, na redução de palavras, somente ocorrem modificações na grafia.

Após a obtenção dos resultados sobre a eliminação de toques na grafia internáutica, decidimos verificar quantos toques comumente são utilizados para a formação desses itens. Destacaremos a seguir os resultados dessa investigação:

Toques na formação dos itens	Nº de itens
2	13
3	8
1	7
4	1

Quadro II – Toques na formação de itens em internetês.

Os dados apresentados acima indicam uma frequência maioral de 45% (13/29) para os itens grafados com 2 toques. Segundo, estão os itens grafados com 3 toques, nos quais a frequência apontou para 28% (8/29) dos itens. Em terceiro lugar, estão os itens foram grafados com apenas 1 toque com 24% (7/29). Por fim, apenas um item, entre os selecionados para o estudo, foi grafado com 4 toques 1% (1/29).

Discussão

Dentre as 50 palavras selecionadas para o estudo, pudemos verificar que há uma grande incidência de palavras grafadas em internetês.

Ao observarmos as ocorrências de eliminação de toques, na passagem da grafia padrão para o internetês, pudemos constatar que das 29 formas em internetês selecionadas para o estudo, em 81% (24/29) dos casos ocorreram reduções de toques. Esse resultado é considerado o mais freqüente em termos de ocorrências, já que a redução de toques ocorreu em 24 palavras.

Por outro lado, a análise possibilitou-nos perceber que em 63% (15/24) dos itens contidos na amostra, há uma tendência para as reduções ocorrerem nos vernáculos-padrão compostos por 4 e 3 toques.

É importante ressaltar que no internetês, há uma predominância de palavras grafadas com 2 e 3 toques, isto é, dos 29 casos observados 73% (21/29), das palavras foram grafadas com 2 e 3 letras/acento.

Conclusão

A partir da análise das amostras, pudemos concluir que há efetivamente uma economia de caracteres digitados na grafia do internetês, o que reverte a um maior ganho de tempo. Essa economia de caracteres digitados ocorre numa frequência maior, comparando-as com as palavras que apresentam modificações na grafia, sem redução de toques. Outro aspecto importante é a preferência pela formação de palavras grafadas com 2 e 3 toques, proporcionando um maior ganho de tempo na digitação e servindo-se "das máximas[4] de velocidade e brevidade" (Thurlow, 2006).

Dessa forma, a internet mudou a maneira de comunicação enfatizando as abreviações e por conseguinte a diminuição de toques digitados originando uma nova grafia - o internetês – que é um novo meio de comunicação linguistica que não acontece frequentemente na história da humanidade (Crystal, 2001).

Bibliografia.

BERBER SARDINHA, A. P. *Lingüística de Corpus*. Barueri-SP: Manole, 2004.

BIBER, D. *Corpus Linguistic – Investigating Language Structure and Use.* Cambridge University Press, 1998.

HALLIDAY, M. A. K. *Corpus Studies and probabilistic grammar.* In.: K. Aijmer & B. Alterberg (org). *English Corpus Linguistic: Studies in honour of Jan Svartivik.* (30-43). London Logman, 1991.

_____*Quantitative studies and probabilities in grammar.* In: M. Hoey (Ed.) **Data Description Discourse – Papers on the English language in Honour of Jonh McH Sinclair on his Sixtieth Birthday** (1-25). London: HarperCollins, 1992.

HOUAISS, A. *Dicionário eletrônico Houaiss da Língua Portuguesa.* Rio de Janeiro: Objetiva. Disponível no site: www.uol.com.br

THURLOW, Crispin & BROWN, Alex. *Generation Txt? The sociolinguistics of young people's text messaging.* University Washington, 2006.

POSSENTI, SÍRIO *Revista Língua Portuguesa - a Revolução do Internetês.* Segmento, nº 5, 2006. p. 24.

4 Serving the sociolinguistic 'maxims' of brevity and speed.