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 DOS PADRÕES DO INTERNETÊS"

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

[1]PUC-SP, LAEL.Rua Monte Alegre, 1800 - Perdizes, SP: zmgg@uol.com.br
[2] PUC-SP, LAEL. Rua Monte Alegre 1800 - Perdizes, S.P: tony4@uol.com.br

Resumo: o presente trabalho apresenta uma investigação em um corpus coletado de blogs, escritos por jovens que utilizam o internetês para a comunicação. Para a análise utilizamos a ferramenta WordSmith Tools, a qual nos auxiliou com o levantamento dos padrões léxico-gramaticais do item *'td'*, com sentido de 'tudo' e 'todo'. A investigação pretende verificar se há padronização na linguagem da internet e se esses padrões contribuem para os sentidos do item *'td'*.

Palavras Chaves: Lingüística de Corpus, padrões, internetês.

Área de Conhecimento: Lingüística, Letras.

Introdução

Com advento da Internet, os usuários de computadores começaram a fazer parte de uma rede de conexão que revolucionaria a comunicação por meio da linguagem, integrando povos e interagindo com o resto do mundo, utilizando a escrita como fonte principal de comunicação, dando origem a um novo fenômeno lingüístico: 'o internetês'. 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". Assim, as abreviações típicas do internetês não comprometem a língua que é formada por regras e leis combinatórias (sintaxe e gramática). Possenti afirma que as abreviações são interessantes, pois privilegiam as consoantes, isto é, o nome das consoantes muitas vezes 'supre' as vogais que não são escritas. Como exemplo, destacamos o item '*kd'* (k = ka; d = de), neste caso, o nome da letra '*k'* supre a vogal '*a',* o mesmo ocorrendo com a letra *'d'* que é suprida a vogal *'e'.* Neste contexto, podemos considerar que o internetês é o resultado de uma comunicação rápida e instantânea, constituída de uma economia de caracteres digitados e uma despreocupação com as normas ortográficas e gramaticais da Língua Portuguesa.

Esse contexto tecnológico não favoreceu apenas a criação de uma nova grafia, mas também propiciou o desenvolvimento de pesquisas baseadas em corpus, ou seja, a Lingüística de Corpus (LC). Segundo Berber Sardinha (2004:3) a LC "ocupa-se da coleta e exploração de corpora, ou conjunto de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador". A grande vantagem desta área de estudo é o uso de computador permitindo a execução de tarefas complexas, oferecendo um grande potencial de pesquisa e análise, as quais só são possíveis através de programas computacionais.

Desta forma, os trabalhos desenvolvidos com a utilização do arcabouço teórico e do instrumental metodológico fornecidos pela Lingüística de Corpus revelam uma vasta quantidade de evidências lingüísticas advindas de corpora, permitindo lingüistas, tradutores, lexicógrafos e muitos outros profissionais a desvendar novas formas de percepção da língua.

Corpus de Estudo

Neste quadro teórico-metodológico, nosso objeto de estudo é um corpus do gênero *Blog;* esse gênero foi escolhido a fim de melhor compreender a prática da escrita dos jovens que utilizam a internet para a comunicação. De acordo com Marcuschi (2005: 61), "a princípio os *blogs* eram listas de *links* e *sites* interessantes que poderiam ser consultados, bem como notas de atalhos para navegação". Atualmente, os *blogs* estão sendo utilizados para anotações diversas como: poema, crítica literária, letras de música, exposição de idéias, opiniões políticas, enfim tudo que é dialógico no ambiente virtual. Como exemplo, podemos destacar o *Blog* 'Idéias e

Coisas Malucas[1]' que publica fotos, textos, filmes, músicas, e discussões atuais.

O gênero *blog,* focado nesse trabalho, é um diário público, acessado por jovens que se comunicam e expressam seus sentimentos, pensamentos e opinam sobre assuntos diversos. Esse gênero é importante para a sociedade, pois operam em um contexto definido, apresentam o uso do internetês, é realizado por forças sociais e tecnológicas, constituindo um conjunto de dados lingüísticos autênticos em linguagem natural (Marcushi, 2005).

Estudos dos Padrões

O presente trabalho propõe enfocar a questão da padronização de um item em internetês por meio da exploração de um corpus coletado de *blogs.*

De acordo com Berber Sardinha (2004:39) os padrões de uma palavra podem ser definidos "como todas as palavras e estruturas com as quais são regularmente associados e que contribuam para seu significado".

O estudo da padronização obtém o apoio teórico no princípio idiomático (*idiom principle*), segundo o qual "o usuário de uma língua tem a sua disposição um grande número de frases pré ou semiconstruídas que constituem em escolhas únicas, muito embora pareçam analisáveis em segmentos" (Sinclair, 1991). Assim, quando falamos em padrão léxico-gramatical, pressupomos que haja um espaço comum formado pelo léxico e pela sintaxe, colocando em dúvida, desta forma, a dicotomia entre léxico e gramática. Em outras palavras, a escolha de um item lexical da língua, implica na diminuição das escolhas dos itens lexicais e das categorias gramaticais que podem compô-lo. Da mesma forma, a escolha de uma classe gramatical também limita a escolha possível de classes gramaticais e itens lexicais que podem segui-la.

Nesta visão, a LC descreve as probabilidades de certos itens ocorrerem em co-textos específicos, e desse modo, a separação entre os níveis do léxico e da gramática torna-se supérflua, sendo uma questão de conveniência analítica, sem apoio empírico (Sinclair, 1991).

Metodologia

Para análise e o levantamento dos padrões, foram executados os seguintes procedimentos: Primeiramente coletamos um corpus de *blogs* em sites da internet que apresentavam comentários de usuários e participantes em internetês. Em seguida, buscamos a ferramenta WordSmith Tools (Scott 1997) que nos forneceu 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. Ao observarmos os *types* mais freqüentes, percebemos que o item *'td'* indicava vários sentidos, quais sejam: todo, toda, tudo, todas e todos. Optamos, então por analisá-lo, com a finalidade de buscar evidências que constatem se o léxico é padronizado, observando os contextos de ocorrência, utilizando a lista de concordância e os colocados. "Concordância é uma lista de ocorrências de um item específico, dispostas de tal modo que a palavra de busca ou palavra nódulo (a que se centra a pesquisa) aparece centralizada e rodeada pelas palavras que ocorrem junto a elas no corpus" (Berber Sardinha, 2004). Igualmente importante é a observação dos colocados. "Colocados são associações entre as palavras que acompanham (tanto à direita como à esquerda) a palavra alvo da pesquisa" (Berber Sardinha, 2004). Com esses procedimentos poderemos comprovar se as palavras se submetem às regularidades do tipo de associação, mesmo tratando-se do internetês. Esses pressupostos vão ao encontro da declaração: "na língua, a menos que se prove o contrário, todas as palavras possuem padronização, escolhendo os padrões a que se associam, privilegiando alguns vizinhos e preterindo outros" Berber Sardinha (2004:221).

Resultados

Para análise do uso em contexto do item *'td'*, recorremos a ferramenta 'Concord' do WordSmith Tools que nos forneceu a lista de concordância e os colocados mais freqüentes. Observando as concordâncias, chegou-se a um total de 760 linhas ou 0,55%[4] relativas ao item no corpus. Após analisar as ocorrências e diferir os sentidos, obtivemos os seguintes resultados:

- *'td'* (tudo) apresentou 75% (571/760) do total das ocorrências no corpus.
- *'td'* (todo) apresentou 18% (137/760) do total de ocorrências no corpus.
- *'td'* (toda) apresentou 5% (36/760) do total de ocorrências no corpus.
- *'td'* (todos e todas) apresentaram a minoria de ocorrências, isto é, 2% para os dois sentidos (16/760).

[1] http://ideiasecoisasmalucas.blogspot.com/

[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).

[4] Valor calculado sobre o total de palavras (tokens) do corpus.

Após a observação das linhas de concordância, selecionamos para a análise apenas as ocorrências que correspondiam aos sentidos 'tudo' e 'todo', pois apresentaram uma freqüência maior de ocorrências. Em seguida foram analisados os agrupamentos[5] (*clusters*) selecionados, com a finalidade de classificá-los em porções de acordo com a sua freqüência. Essa observação incluiu a necessidade da descrição dos sentidos associados com as colocações recorrentes.

Discussão

Ao observarmos as linhas de concordância, a primeira constatação é que há um padrão freqüente que é formado por um colocado que sucede *'td*' que apresenta o sentido de 'todo', neste caso, referimo-nos a '*mundu*/mundo'. Conforme mencionado, das 137 linhas de ocorrências obtidas no corpus com sentido de 'todo', o padrão '*td* + *mundu/*mundo' abrange 73% (100/137) das ocorrências. Esse padrão vem imediatamente antecedido pela preposição '*pra'* (para) em 22% (30/137) das ocorrências. Para ilustrar, apresentamos abaixo algumas linhas de concordância:

624	feliz anu novu pra	td	mundu ai viu... Meu ESPE
625	entaum mil bjus pra	td	mundu amu v6 má... tudu
626	Feliz Natal pra	td	mundo! Bjos especiais pra
627	Entaum Bjusss* pra	td	mundo!! e especialll hehe
628	m..e. isso eh pra	TD	mundo! Haha..Bom vo indu
629	eira...du LP...!! Pra	td	mundu q entra aki.....
630	(ou seja praticamente	td	mundu) Mais ctz q uns
631	E um feliz 2005 pra	td	mundo!!!! Morzinhuu

Quadro I – linhas de concordância do item 'td' (todo).

A explicação para o constante uso do padrão '*td* + *mundu*/mundo' e '*pra* + *td* + *mundu*/mundo' é clara; os usuários referem-se a todos que acessam o *blog* para redigir uma mensagem, buscando a interatividade com os participantes, desejando-lhes algo 'positivo'. Vale salientar que 't*d* + m*undu*/mundo', normalmente, é aplicado ao "mundo particular" ou a todos que fazem parte do convívio dos blogueiros. Desse modo, o sentido não se aplica a todas as pessoas do mundo.

Um outro agrupamento freqüente é '*td* + dia' e 'dia + *td*' que indica idéia de tempo. Esses agrupamentos são formados pelo colocado 'dia', como mostra o quadro a seguir:

645	..Bia e Suellen (q m atura	td	dia...brigado pela sua
646.	tem o q te fla...agente c fla	td	dia hahahaha....bjus..
647	erdade foi uma eternidade..	td	dia a mesma coisa...
658	dormi o dia praticamento	td!	E hj vi as minhas vidas!
733	fessores, poder fikr o dia	td	sem ter q se preoucupar
734	q vo fica em casa u dia	td	sem faze nda !!hehehe
739	q fiquem me lembrando	td	dia .... "Sou aquela q
741	io pra k. Fez mo frio o dia	td	!! sabado o marcos veio

Quadro II – linhas de concordância do item *'td'* (todo).

Com o levantamento identificamos dois padrões típicos para o colocado 'dia': anteposto e posposto ao nódulo. O padrão de posposição '*td* + dia' apresentou 8% (10/137) das ocorrências no corpus. Esse padrão indica sentido de um tempo extenso, exprimindo a idéia de vários dias (vide linhas 645, 646, 647, 739). Já o padrão 'dia + *td*', o colocado 'dia' vem anteposto ao nódulo e está sempre precedido pelo artigo definido 'o/*u*' (o/*u* + dia + *td*). Esse padrão apresentou 6% (8/137) das ocorrências no corpus, dando-nos a idéia de tempo limitado, expressando sentido de apenas um dia (vide linhas 658, 733, 734, 741).

Os temas que abordam a temporalidade, destacando o colocado 'dia', referem-se freqüentemente aos relatos do cotidiano.

Em seguida, apresentaremos os colocados referentes ao item '*td'* que apresentaram o sentido de 'tudo'. Há uma maior incidência de padrões para esse sentido, já que a palavra 'tudo' tem um sentido mais abrangente. Abordaremos aqui, somente os dois padrões mais freqüentes encontrados no corpus.

Os colocados agrupam-se em vários conjuntos lexicais, sendo que o maior deles é formado pelo advérbio 'bem/*bom/baum'* posposto ao nódulo com 21% (120/571) das ocorrências. O quadro a seguir destaca alguns exemplos dos agrupamentos:

38	nenhuma comentou...mas	td	bem nunca + deixo msg
41	com a cara.mas ateh aih	td	bom.. pq eu (normalmente
83	vai me abandonar, mais	td	bem nem ligo msm vai
101	padrasto q eu adoro! mas	td	bem..mt legal e amanha
106	meio distante assim. +	td	bemmmmm.. bju pra
127	oalzinhu do meu bloguxo	td	bem com vcs? Ai espero
130	se inscrevam!! Bjusss!!	Td	baum?? comigo td..Eai
134	.....ninguem merece.. +	td	bem, bom agora to aki

Quadro III – linhas de concordância do item 'td' (tudo).

Estes conjuntos comumente aparecem com intuito de interagir com os participantes, cumprimentando os usuários ou concordando com uma dada situação.

Outro padrão encontrado no corpus é formado pelo verbo 'dar' anteposto ao nódulo. Os padrões '*da* + *td*' e 'dar + *td*' apresentaram 12% (69/571) das ocorrências no corpus. Esse padrão é comumente encontrado com o verbo 'vai', apresentando o sentido futuro.

[5] Esses agrupamentos foram formados pelas cinco palavras (colocados) antepostas ao nódulo e as cinco palavras (colocados) pospostas ao nódulo.

Já os agrupamentos '*q* + *de*/dê/*d'* antepostos ao nódulo apresentaram 16% (91/571) das ocorrências no corpus. Tal padrão tipicamente é usado indicando sentido condicional (complementando formas verbais 'espero' e 'tomara'), com a finalidade de exprimir vontade, desejo, confortar e elevar a auto-estima. Alguns exemplos são destacados abaixo:

58	nsada, mas espero q dê td certo. Semana q vem
61	..aii kredu.tomara q de td certu!. vo tem q tenta
84	sorti p/ vcs tomara q d td certuuuu.agora convence
110	na praia.espero q dê td certo, q não chova
118	Valença e tomara q de td certo!! Q essa chuva pare
119	ZER OQ NEH? VAI DA TD CERTO..VC VAI VER
138	Deus quiser vai dar td certo pq tadinhos dos

Quadro IV – linhas de concordância de item 'td' (tudo).

Conclusão

A padronização exibida pelos itens analisados nos oferece uma visão da linguagem utilizada nos meios digitais e nos revela muitas formas possíveis de uso do item '*td'.* As análises indicaram que o item *'td'* está associado a comentários referentes às particularidades do dia-a-dia e remetem às temáticas recorrentes, dentre as quais se destacam situações cotidianas vividas pelos blogueiros e a expressão de sentimentos. De modo geral, os blogueiros parecem utilizar tais expressões para interagir com a sociedade.

Ao contrário da opinião de muitos que criticam o internetês, afirmando que é uma grafia transgride regras gramaticais, não é o caso aqui observado; pelo que pudemos notar com relação aos padrões, as regras (sintaxe e gramática) não são quebradas, existindo apenas subtrações de letras/acentos na grafia das palavras. Essas subtrações não parecem alterar os sentidos dos agrupamentos e seus respectivos padrões, em relação a seus equivalentes da norma culta.

Portanto, os padrões verificados podem ser claramente remetidos a maneiras de expressão ratificadas pela norma padrão.

Bibliografia.

BERBER SARDINHA, A. P. (org.). *A Língua Portuguesa no Computador*. Campinas-SP: Mercado das Letras, Fapesp, 2005.

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

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

MARCUSCHI, L. A. e XAVIER, A. C. (orgs.). *Hipertexto e gêneros digitais: novas formas de construção do sentido.* Rio de Janeiro: Lucerna, 2005.

SINCLAIR , J. M. *Corpus, Concordance, Collocation***.** Oxford: Oxford University Press, 1991.

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