VARIAÇÃO DE PALAVRAS-CHAVE ENTRE GÊNEROS DE UM CORPUS JORNALÍSTICO

Carlos H. KAUFFMANN (LAEL-PUC/SP)

ABSTRACT: *This study examines from a Corpus Linguistics perspective (Berber Sardinha, 2004) main keywords occurring across daily press genres in Brazilian Portuguese. The keywords observed, e.g. verbal lemmas and personal pronouns, may indicate recurrent presence and/or absence in different selected genres.*

KEYWORDS: *keyword; corpus; press genre.*

0. Introdução

Com os recursos teórico-metodológicos proporcionados pela Lingüística de Corpus (Berber Sardinha, 2004), este estudo pretende comparar, por meio da análise da freqüência de palavras e de palavras-chave, diversos gêneros textuais presentes na imprensa escrita diária do Brasil contemporâneo. O objetivo é observar o comportamento lingüístico da distribuição da freqüência vocabular entre grupos de textos previamente selecionados por critérios estabelecidos pela comunidade discursiva e seu contexto sociocultural do qual são expressão (Swales, 1990; Bhatia, 1993), de forma a vislumbrar como a materialização do gênero no texto condiciona as opções léxico-gramaticais do autor/falante. Estudos que utilizam procedimentos estatísticos como a análise discriminante para a identificação de gêneros textuais (Biber 1993, por exemplo) poderão ser beneficiados pelos resultados deste trabalho preliminar.

A investigação utiliza um corpus inicial que equivale ao volume de textos publicados em uma semana nas edições de um jornal diário de circulação nacional (“Folha de S.Paulo”), como detalhado no Anexo 1 –ao todo, 1.431 textos (493.780 palavras). Uma classificação manual de gêneros foi efetuada a seguir pelo pesquisador, separando do corpus grupos de textos dos seguintes gêneros textuais: reportagem, notícia, entrevista, editorial, crítica, resenha, coluna de notas, carta, chamada (textos de primeira página),

artigo e crônica. Tais grupos serão chamados doravante de subcorpora de estudo.

Utilizaremos como definição de palavra-chave a propugnada por Berber Sardinha (1999b): é aquela "cuja freqüência em um corpus de estudo é estatisticamente diferente à sua em um corpus de referência", podendo ser atribuído a ela um valor positivo (quando a freqüência no corpus de estudo é superior à do corpus de referência) ou negativo (quando a freqüência no corpus de referência é superior à do corpus de estudo). Para a obtenção das palavras-chave dos subcorpora de estudo, portanto, foi utilizado um corpus de referência para fornecer uma medida comum de contraste de freqüência lexical. O corpus de referência (CR) escolhido foi o Banco de Português, do projeto DIRECT (LAEL/PUC-SP), que reúne um conjunto de 221.957.953 palavras de diversos gêneros dos modos oral e escrito.

1. Fundamentos teóricos

Uma importante corrente da Lingüística, ligada aos estudos iniciados por Firth, tomou vulto nos últimos 40 anos (Stubbs, 1993; Monaghan, 1979), com o aparelhamento tecnológico que permitiu a análise de largas quantidades de textos por meio do computador. O exame da língua em uso, em contraposição a modelos racionalistas baseados na intuição do pesquisador, como os produzidos por Chomsky e Saussure (cf. Stubbs, 1993), propõe uma abordagem empírica baseada no corpus, a partir da qual se podem extrair padrões de natureza léxico-gramatical (Sinclair, 1991). As áreas lexicográfica e pedagógica foram grandemente beneficiadas com a aplicação desse conhecimento no desenvolvimento de dicionários, gramáticas, métodos de ensino e novos materiais didáticos (cf. Kennedy, 1998 e Hunston, 2002).

A descrição lingüística proposta compartilha de uma mesma visão probabilística da linguagem, no âmbito da Lingüística de Corpus (Berber Sardinha, 2000, 2004). A análise de palavras-chave em um corpus lingüístico empregada utiliza técnicas provenientes de estudos como Berber Sardinha (1999a e 1999b). Utilizando vários recursos de natureza estatística, Biber (1988, 1993) estudou a língua inglesa tendo como ponto de partida um critério externo de classificação por gênero (ou registro, na terminologia empregada posteriormente), com intuito de realizar uma pesquisa de base empírica e exploratória na identificação de padrões lingüísticos subjacentes ao gênero.

Esta pesquisa, igualmente, tem uma meta de trabalho similar, com a adição de observar linhas de conduta sugeridas em EAGLES (1996) na seleção dos subcorpora.

Importante considerar aqui a conceituação específica de gênero (Bhatia, 1993; Swales, 1990), que o circunscreve a dada organização da cultura e em função de um propósito social reconhecido. A coerção a que é submetido o texto limita as escolhas do autor/falante em termos lingüísticos e influencia, entre outras áreas, o léxico: "Cada seleção textual coage as escolhas lexicais possíveis, e é nessa combinação entre escolhas lexicais e textuais efetuadas por escritores ou falantes que a sua criatividade é expressa." (Hoey, 1991:217)

Para a seleção dos textos que formou os subcorpora de estudo, buscou-se subsidiariamente o amparo teórico dos estudos de Jornalismo (Marques de Melo, 1994, 1992, 1998), no âmbito da Comunicação Social, e definições de gênero existentes em manuais de redação de jornais. Com tais critérios de apoio em mente, efetuou-se a seleção de textos das páginas dos diários em subgrupos, de acordo com categorias que se configuram culturalmente como gêneros de cunho jornalístico.

2. Procedimentos metodológicos

O desenho do corpus utilizado neste estudo determina a sua representatividade e lhe proporciona certo equilíbrio (Hunston, 2002:28): a fonte primária do corpus são as edições diárias de um jornal que compõem uma semana na regularidade (segunda-feira, terça-feira, etc.), mas não são seqüenciais. Trata-se de uma "semana construída", cujas edições diárias foram aleatoriamente escolhidas entre as 365 publicadas em um ano (cf. anexo 1). Tal medida evita "distorção excessiva na amostragem" (Kennedy, 1998:75). Percebe-se na Tabela 1 que, na divisão em gêneros textuais, o peso correspondente à reportagem é majoritário, correspondendo a mais da metade de textos e de *tokens* (ocorrências). É secundado pelo gênero artigo, se usado o critério de volume de *tokens*, ou pela notícia, se for considerado o volume de textos. Se é possível classificar o corpus inicial total como de média extensão (Berber Sardinha, 2000:346), o mesmo não podemos afirmar dos subcorpora de menor peso relativo de *tokens*, como chamada, crônica e resenha.

Os textos divididos em gêneros após um processo de seleção manual geraram agrupamentos de subcorpora, assim dispostos e equilibrados[1]:

Gênero	**Número de textos**	***tokens* (ocorrências)**	**% *tokens* (total)**
reportagem	645	247.393	53,5%
artigo	81	52.655	11,4%
notícia	290	50.400	10,9%
coluna de notas	69	39.737	8,6%
entrevista	24	23.115	5,0%
crítica	47	15.860	3,4%
carta	14[2]	10.318	2,2%
editorial	19	6.195	1,3%
resenha	11	5.825	1,3%
chamada	72	5.692	1,2%
crônica	11	5.509	1,2%
Total	1.283	462.699	100%

Tabela 1: Composição dos subcorpora de estudo

Conclui-se que, em relação a gêneros de baixa representatividade, esta pesquisa deve ser tomada como preliminar e sugestiva, já que é apenas produto dos dados observados no corpus.

A ferramenta básica empregada no estudo foi a suíte WordSmith Tools (versão 3.0), notadamente os módulos WordList e KeyWords. As listas de palavras-chave obtidas nos subcorpora de estudo, produto de sua comparação com o CR, foram ordenadas alternadamente, para uma análise prévia (que não será detalhada neste artigo), de acordo com os critérios de maior chavicidade (*keyness*), freqüência no subcorpus de estudo e freqüência no CR. Tais listas apontam vários pontos de vista a partir dos quais as palavras-chave mais salientes emergem, ensejando listas comparáveis de palavras que têm relevância nos subcorpora.

Buscou-se especialmente observar nas listas da análise prévia a utilização dos verbos mais comumente empregados em cada subcorpora de estudo, bem como a presença ou a ausência de itens gramaticais (isto é, que

compõem categorias gramaticais tradicionalmente estabelecidas, como artigos, preposições, conjunções, advérbios, entre outros) e lexicais de cunho não-específico. O intuito é destacar palavras caracterizadoras de gêneros textuais, deixando de lado palavras pertencentes a categorias como a dos substantivos próprios, que delimitam semanticamente de forma muito estreita a análise e individualizam excessivamente os textos.

3. Apresentação e análise dos dados coletados

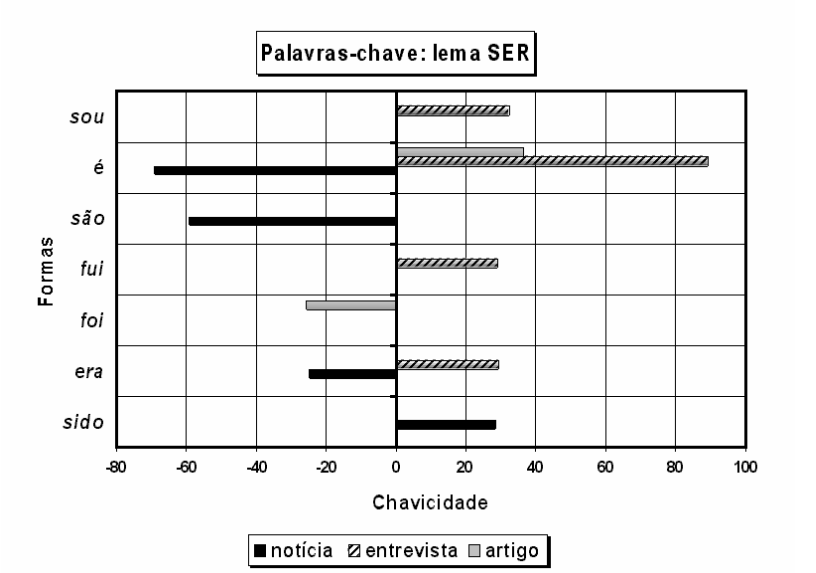

Gráfico 1: Palavras-chave do lema SER

Os gráficos mostrados nesta seção resumem os principais resultados obtidos na etapa de análise prévia. É necessariamente uma lista não-exaustiva de palavras-chave, com várias exclusões, em razão do espaço disponível. Foram privilegiadas as palavras-chave comuns aos subcorpora dos gêneros mais representativos: reportagem, artigo, notícia, coluna de notas e entrevista.

Outro critério de pesquisa foi a observação de polarizações entre as palavras-chave. Algumas delas apresentam chavicidade (*keyness*) positiva *e* negativa entre os diferentes subcorpora. Assim, é possível observar o comportamento de aproximação e distanciamento entre gêneros no que tange às palavras-chave de cada grupo.

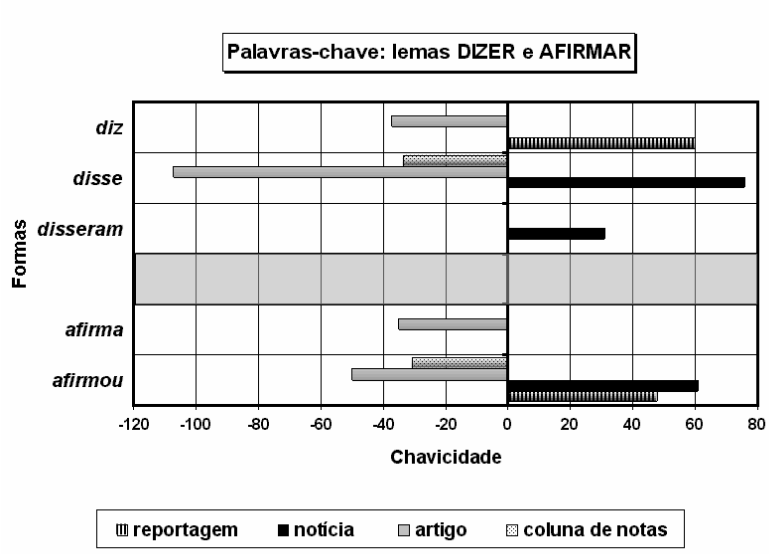

Gráfico 2: Palavras-chave dos lemas DIZER e AFIRMAR

Os dois primeiros gráficos mostram a situação de três verbos recorrentes no corpus: os lemas[3] dos verbos SER, AFIRMAR e DIZER, nas formas em que se manifestaram como palavras-chave. Eles mostram que alguns gêneros têm um comportamento diferenciado em relação a outros por meio de algumas palavras-chave bastante freqüentes. O terceiro gráfico enfoca algumas formas salientes de pronomes pessoais.

O verbo ser é o mais freqüente verbo do corpus estudado e – com boa dose de certeza – por extensão, da língua portuguesa. O Gráfico 1 indica que, em especial, o gênero entrevista utiliza mais freqüentemente certas

formas do lema SER, de primeira pessoa (*sou*, com índice de chavicidade 36,2; *fui*, 29,1), e terceira pessoa (*é*, 89,3; *era*, 29,3). Por sua vez, o gênero notícia evita as formas *é* (chavicidade negativa de -69,2), *são*[4] (-59,1) e *era* (-24,9), caracterizando uma tendência de comportar-se em oposição à entrevista. O gênero artigo divide-se em movimentos contraditórios que podem indicar um uso mais freqüente do presente do indicativo (pela presença de *é*, 36,5), em oposição ao pretérito perfeito (dada a chavicidade negativa da forma *foi*, -25,6). Digno de menção também é a chavicidade positiva da forma *sido* (28,7) em notícia, sugerindo uma presença maior do condicional.

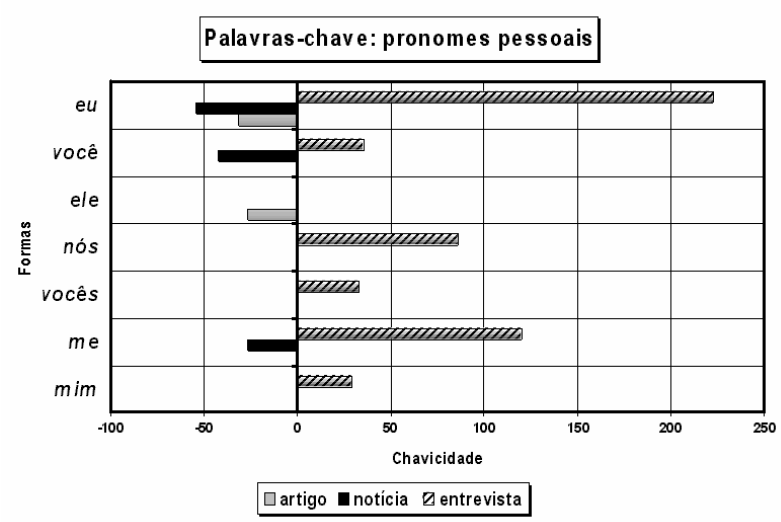

Gráfico 3: Palavras-chave dos pronomes pessoais

O propósito de apresentar em um mesmo ambiente os lemas DIZER e AFIRMAR (Gráfico 2) é destacar a semelhança de comportamento entre os gêneros notícia e reportagem em verbos que possuem uma função similar no texto jornalístico. Ambos apresentam altos índices de chavicidade positiva nos dois verbos: em notícia preponderam as formas *disse* (75,7), *disseram* (31) e *afirmou* (60,9), enquanto em reportagem predominam as formas *diz*

(59,8) e *afirmou* (47,6). Em tendência oposta, os dados observados indicam uma recusa sistemática de uso por parte do gênero artigo a essas formas. Exemplo disso é a grande chavicidade negativa da forma *disse* (-107,4). De um modo menos enfático, o gênero coluna de notas também segue o comportamento do gênero artigo.

O Gráfico 3 mostra o quanto os pronomes pessoais são desproporcionalmente usados entre os gêneros presentes no jornal. Percebe-se um claro predomínio do gênero entrevista na utilização de pronomes pessoais – o maior exemplo é a surpreendente chavicidade positiva alcançada pela forma *eu*, de índice 223,3, acompanhada dos átonos de primeira pessoa *me* (120,1) e *mim* (29). Em menor escala, surgem as formas *nós* (86,4), *você* (35,4) e *vocês* (33,1). Em contraste, aparecem dois gêneros que indicam significativamente, ainda que menos marcadamente, a ausência da formas reta e átona de primeira pessoa, a notícia (*eu*, -54,7; *me*, -26,7) e o artigo (*eu*, -31,7). Duas formas apresentam também chavicidade negativa (-26,7): *ele*, em artigo, e *me*, em notícia.

4. Conclusão

A freqüência ou a ausência relativa de determinados itens lexicais entre diferentes gêneros textuais é um dos fenômenos que podem ser melhor compreendidos com o estudo comparativo de palavras-chave.

Foram observados padrões na utilização recorrente de formas entre os gêneros analisados, como por exemplo, a presença de pronomes de primeira pessoa (*eu*, *me*, *mim*, *nós*) em entrevistas. Entre os verbos que revelam uma freqüência estatística diferenciada, estão principalmente o lema SER (formas *é* e *era* freqüentes em entrevista e pouco freqüentes em notícia) e os lemas DIZER e AFIRMAR (formas *disse*, *diz*, *disseram* e *afirmou*, freqüentes em notícia e reportagem, mas pouco freqüentes em artigo e coluna de notas).

A partir da observação de padrões de natureza lingüística, como os possibilitados pelo exame da freqüência comparada de palavras-chave, portanto, podem ser identificados elementos semelhantes e diferenciadores de corpora estabelecidos originalmente por critérios externos ao nível lingüístico.

NOTAS:

1. O total de textos coletados nos subcorpora de estudo é inferior ao corpus de estudo descrito no anexo 1 porque foram excluídos no processo de seleção alguns textos que não se encaixavam nos gêneros aqui analisados (por exemplo, efemérides, comentário e análise).
2. Os "textos" do gênero carta, ressalte-se, são constituídos pelo conjunto de cartas publicadas na seção diária específica, não tendo sido contadas individualmente.
3. Lema, segundo Sinclair (1991:173), é o agrupamento de formas semelhantes ou variantes de uma mesma extração lexical.
4. Ressalve-se que é uma forma ambígua, que pode significar não apenas uma forma verbal, mas também um substantivo ou um adjetivo.

ANEXO 1

Consideramos todos os textos veiculados da "Folha de S.Paulo" nas edições local e nacional do dia, descontados os textos repetidos ou reduzidos de uma edição para a outra. Foram excluídos alguns elementos, como: tabelas, infográficos, cotações, fragmentos de obras literárias, frases, textos-legenda, publicidade, informe publicitário, propaganda legal, horóscopo, obituário, expediente e classificados. As edições utilizadas, todas do ano de 2003, circularam em: 17 de janeiro (segunda-feira); 03 de junho (terça-feira); 06 de agosto (quarta-feira); 27 de março (quinta-feira); 30 de maio (sexta-feira); 19 de julho (sábado); e 19 de outubro (domingo).

REFERÊNCIAS BIBLIOGRÁFICAS

BHATIA, V. *Analysing Genre: Language Use in Professional Settings*. London: Longman, 1993.

BERBER SARDINHA, T. Usando o WordSmith Tools na investigação da linguagem. *Direct Papers*, São Paulo, v. 40, 1999a.

_____. O banco de palavras chave. *Direct Papers*, São Paulo, v. 39, 1999b.

_____. Lingüística de corpus: histórico e problemática. *D.E.L.T.A.*, São Paulo, v.16 (2), p. 323-367, 2000.
_____. *Lingüística de Corpus*. São Paulo: Manole, 2004.
BIBER, D. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.
_____. Using register-diversified corpora for general language studies. *Computational Linguistics*, v. 19, n. 2, p. 219-241,1993.
EAGLES. *Preliminary recommendations on corpus typology*. EAG-TCWG-CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale, 1996. Disponível em <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>. Acesso em: 15/08/2004.
FOLHA DE S.PAULO. *Manual Geral da Redação*. São Paulo: Folha de S.Paulo, 1984 e 1987 (2. ed).
_____. *Novo Manual da Redação*. São Paulo: Folha de S.Paulo, 1992.
_____. *Manual da Redação*. São Paulo: Publifolha, 2001.
HOEY, M. *Patterns of Lexis in Text.* Oxford: Oxford University Press, 1991.
HUNSTON, S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.
KENNEDY, G. *An Introduction to Corpus Linguistics*. London: Longman, 1998.
MARQUES DE MELO, J. (org.). *Gêneros Jornalísticos na Folha de S.Paulo*. São Paulo: FTD, 1992.
_____. *A Opinião no Jornalismo Brasileiro*. Petrópolis: Vozes, 1994.
_____ (org.). Gêneros e formatos na comunicação massiva periodística: um estudo do jornal "Folha de S.Paulo" e da revista "Veja". In: *Anais do Congresso Brasileiro de Ciências de Comunicação*, 21. Recife: Intercom, 1998.
MONAGHAN, J. *The Neo-Firthian Tradition and its Contribution to General Linguistics*. Tübingen: Max Niemeyer Verlag, 1979.
SINCLAIR, J. *Corpus, Concordance, Collocation***.** Oxford: Oxford University Press, 1991.
STUBBS, M. British traditions in text analysis. In: BAKER, M., FRANCIS, G. & TOGNINI-BONELLI, E. (org.), *Text and Technology – In Honour of John Sinclair* (p. 1-33). Philadelphia e Amsterdam: John Benjamins, 1993.
SWALES, J. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 1990.