FEIXES LEXICAIS E VISÕES DE MUNDO: UM ESTUDO SOBRE *CORPUS*[1]

Tania M.G. Shepherd (UERJ),
Sonia Zyngier (UFRJ),
Vander Viana (PUC-Rio)

RESUMO: O presente trabalho analisa os feixes lexicais, segundo Biber, Conrad e Cortes (2004), presentes em redações de crianças de quinta e sexta séries de duas comunidades — rural e urbana — com o objetivo de extrair possíveis perfis dos referidos grupos, em termos de suas visões de mundo e de suas organizações textuais. Para tanto, foram compilados dois corpora de 12.205 e 14.662 palavras respectivamente, dos quais extraíram-se feixes lexicais formados de quatro itens, subsequentemente analisados em termos de estrutura e função. Os perfis traçados após analise sugerem que, diferentemente dos participantes da área rural, o grupo urbano não faz uso de padrões repetidos para descrever seu mundo, representando-o sem tomar como base expressões do coletivo.

PALAVRAS-CHAVE: identidade; linguagem; cultura; análise de *corpus*; feixes lexicais; redações.

1) INTRODUÇÃO

Um dos conceitos mais discutidos desde a Grécia Antiga centra-se no papel da linguagem na formação e na expressão da identidade. A noção de identidade como produto de discurso, de linguagem em uso, também não é recente. Tem origem na visão de linguagem como objeto relacional, em que palavras geram significado, não devido a alguma característica intrínseca de objetos, pessoas, processos ou eventos, mas sim por meio de articulações co-construídas através de interações cotidianas.

Na última década, estudos principalmente na área de Análise do Discurso e Analise Crítica do Discurso têm observado a interface entre linguagem e construção de identidade. Wodak et al. (1999) analisam uma sociedade através de seus documentos escritos, enquanto Schiffrin (1994) investiga a construção de identidades individuais através de narrativas pessoais. O presente trabalho não segue nenhuma dessas perspectivas anteriores. O objetivo do mesmo é verificar como o perfil social de determinados grupos pode estar contido na escolhas que estes fazem através do uso de feixes lexicais.

A pesquisa aqui desenvolvida se baseia em algumas premissas analíticas já bem estabelecidas. Primeiramente, acredita-se, como Fairclough (2003, p. 15), que, em suas interações sociais, os seres humanos revelam modos de avaliar e falar sobre o que fazem e sobre quem são. Igualmente, como Teliya et al. (1998), também se entende que qualquer estudo sobre questões culturais deve envolver a análise do léxico empregado nas interações, desde unidades lexicais compostas por palavras simples até estruturas polilexicais mais complexas. Estas descrições, portanto, podem levar à compreensão de relações existentes entre linguagem, identidade e cultura.

As unidades lexicais a serem analisadas aqui tomam por base a distinção proposta por Sinclair (1991, p. 109-110) entre blocos lexicais pré-existentes (*idiom principle*) e de livre escolha (*open-choice principle*)[2]. De acordo com Sinclair (1991), ao se comunicarem, os indivíduos lançam mão de "frases semi-construídas que, na realidade se constituem em uma única escolha", além de usarem repertórios de escolhas individuais. Segundo o autor, qualquer texto resulta do entrelaçamento desses dois princípios. Em outras palavras, ou recorremos a unidades (palavras ou sintagmas) já ouvidos/lidos e internalizados ou fazemos escolhas complexas de natureza léxico-gramatical, mas livres.

Portanto, para investigar os perfis lingüísticos de grupos socialmente diferentes pode-se fazê-lo através dos padrões que evidenciam o principio 'idiomático' proposto por Sinclair (1991). Desta forma, o presente estudo explora a relação intrínseca entre linguagem e prática social através de um *corpus* de 195 textos escritos por crianças da quinta e sexta séries de duas escolas públicas. Uma das escolas situa-se na Favela da Maré, na zona urbana da cidade do Rio de Janeiro e a outra está localizada na região agropecuária de Tocantins, no estado de Minas Gerais[3]. Cada

grupo de composições deve conter padrões comuns de escolhas polilexicais que podem refletir os perfis dessas crianças com relação ao que percebem sobre si próprias e sobre o mundo que as cerca.

2. Pesquisa sobre Escrita Infantil e Lingüística de *Corpus*

O texto escrito por jovens, principalmente em língua inglesa, já foi investigado sob a ótica da Lingüística de *Corpus*. Berber Sardinha e Shimazumi (1996), por exemplo, investigaram as características dos textos de alunos do Reino Unido na faixa etária de 15 anos, a partir de uma amostra dos exames escritos denominados APU (*Assessment of Performance Unit*). A hipótese do trabalho foi a de que essas características emergiriam ao se contrastar as redações destes jovens com textos de um jornal, escritos por profissionais da imprensa.

Com o auxílio de um programa computacional, Berber Sardinha e Shimazumi (1996) extraíram palavras isoladas, bigramas e trigramas[4] dos dois grupos de texto e compararam o nível de formalidade entre eles. A maior contribuição de seu estudo consistiu na verificação da presença idiossincrática do pronome de primeira pessoa do singular na escrita dos adolescentes, bem como da conjunção '*because*', caracterizando uma ênfase no 'eu' como ponto de partida para a articulação de opiniões e uma tentativa de fornecer justificativas para essas opiniões.

Outra contribuição sobre discurso escrito de jovens através da Lingüística de *Corpus* foi a de Sampson (2003), que examinou a linguagem oral e escrita produzida por crianças cuja língua materna era a inglesa. Em seu estudo, Sampson (op. cit.) rotulou sintaticamente todo o *corpus* para medir o nível de complexidade dos períodos e orações. Seu objetivo final foi comparar a escrita infantil com a fala e escrita de adultos a fim de testar a validade de seu etiquetador. Apesar de ter percebido uma tendência por parte das crianças examinadas a usarem mais palavras do que os adultos[5], Sampson (op. cit.) não focalizou a natureza do léxico de seus *corpora*.

Já utilizando a Lingüística de *Corpus* como instrumental para investigar a escrita de jovens, poucos são os trabalhos que se concentram

na extração e análise de grupos polilexicais ou de expressões pré-fabricadas em *corpora* dessa natureza (cf. STUBBS, 2001). Os fatores que dificultam este tipo de pesquisa são a terminologia e o foco analítico. Os grupos polilexicais têm sido rotulados de 'fórmulas' (WRAY, 2002), 'rotinas', 'lexemas frasais' (MOON, 1998), 'molduras colocacionais' (RENOUF e SINCLAIR, 1991) e 'n-gramas' (SINCLAIR, 2004). Nos estudos sobre inglês como segunda língua ou língua estrangeira, são também utilizados os termos 'padrões pré-fabricados' (GRANGER, 1998b) e 'phrasicon' (DE COCK et al., 1998) entre outros. Além da falta de consenso com relação à terminologia, estes estudos não chegam a um consenso nem quanto ao número de itens lexicais que devem fazer parte das seqüências estudadas, nem quanto aos aspectos a serem analisados: se forma, função ou ambos.

Apesar destes problemas, os pesquisadores acima concordam que qualquer usuário da língua escrita ou falada recorre a expressões que podem conter duas ou mais palavras com um significado único – à semelhança do "idiom principle" (SINCLAIR, 1991, p. 109) mencionado acima.

Para o presente estudo utilizamos o trabalho sobre grupos polilexicais de Biber (2004) e Biber, Conrad e Cortes (2004). Estes autores usaram quatro *corpora* com características específicas e, a partir de critérios de freqüência e distribuição, extraíram seqüências formadas por quatro palavras, a que denominaram '*lexical bundles*', ou feixes lexicais.

Em sua descrição, Biber, Conrad e Cortes (op. cit., p. 382) distinguem três tipos de feixes: *Tipo 1*, com fragmentos de sintagmas verbais; *Tipo 2* com orações subordinadas, ou em fragmentos ou em sua totalidade; e o *Tipo 3* com fragmentos de sintagmas nominais ou preposicionais. Em termos de função os autores distinguem também três funções: de *posicionamento*, de *referência* e de *organização discursiva*, cada uma das quais com subdivisões, conforme esquematizado, respectivamente, nas Figuras 1, 2 e 3 abaixo.

A função de *posicionamento* se apresenta em duas grandes subcategorias: *posicionamento epistêmico* e *atitudinal/modal*. Esta última função ainda se subdivide em *desejo*, *obrigação/direcionamento*, *intencionalidade/predição* e *habilidade*. As expressões que expressam *posicionamento* se apresentam no plano pessoal ou impessoal, com exceção da subcategoria 'desejo', que se apresenta somente na forma pessoal.

Expressão de posicionamento	epistêmico		Pessoal
			Impessoal
	atitudinal/ modal	Desejo	Pessoal
		Obrigação / Direcionamento	Pessoal
			Impessoal
		Intencionalidade / Predição	Pessoal
			Impessoal
		Habilidade	Pessoal
			Impessoal

Figura 1
Categorias de posicionamento e subdivisões

A segunda categoria funcional, denominada *referencial*, é formada de seqüências com a função de identificar algo enquanto um todo, parte de um todo, ou mesmo sua característica preponderante. A referência, por sua vez, pode ser focada ou imprecisa e direcionada para um aspecto determinado (tempo, lugar ou texto).

Expressão de referência	Identificação / Foco	
	Imprecisão	
	Especificação de atributos	Especificação de quantidade
		Atributos tangíveis
		Atributos intangíveis
	Tempo / Lugar / Texto	Referência a tempo
		Referência a lugar
		Dêixis
		Referência múltipla

Figura 2
Categorias referenciais e subdivisões

A terceira categoria funcional, a de *organização discursiva*, se constitui de feixes que organizam o discurso de duas formas: introduzindo novas seções ou elaborando as seções anteriores, conforme resumido abaixo.

Organizadores discursivos	Introdução de tópico / foco
	Elaboração/ esclarecimento de tópico

Figura 3
Organização discursiva e subdivisões

A seção abaixo explica como a lista de funções e tipos extraídos dos *corpora* examinados por Biber, Conrad e Cortes (op. cit.) pode ser usada, ainda que com alguns ajustes, para classificar os feixes encontrados nos textos escritos pelas crianças da cidade de Tocantins (MG) e da Favela da Maré (RJ), permitindo, assim, uma radiografia das expressões lexicais dos dois conjuntos de textos e uma descrição do perfil destes participantes.

3) Procedimentos metodológicos

Conforme descrito na Seção 1, o presente estudo focaliza a escrita de crianças da quinta e sexta séries do ensino fundamental de duas áreas distintas. Após a coleta, as redações foram digitadas a fim de se compilar dois *corpora*. O primeiro *corpus*, com redações de tema livre da cidade de Tocantins, foi rotulado como **ARu**, ou área rural. O segundo, com de redações das crianças oriundas da Favela da Maré, foi intitulado **AUr**, ou área urbana.

Com o auxílio do programa *WordsmithTools* (Scott, 1999), calculou-se o número total de palavras em cada *corpus*. **ARu** contém 14.662 palavras enquanto **AUr** totaliza 12.205. Cada uma das composições constitui um arquivo distinto, rotulado com informações sobre local de coleta, sexo e idade do escritor. Buscou-se manter a fidelidade aos textos originais, mas como nenhum dos dois grupos domina regras ortográficas da

língua portuguesa, houve problemas metodológicos quando da digitação das composições coletadas. Para que o programa pudesse ler as seqüências de palavras mais freqüentes em ambos os *corpora*, os erros ortográficos tiveram de ser corrigidos. Assim sendo, palavras como, por exemplo, 'prefiriu', 'aumoçar', 'estáva', foram substituídas por 'preferiu', 'almoçar' e 'estava' respectivamente.

O reduzido conhecimento das regras de pontuação por parte dos participantes também criou problemas na fase de análise dos dados. A decisão de manter a pontuação original levou o programa a identificar seqüências de palavras que, na verdade, não constituem grupos polilexicais analisáveis. Um exemplo é a seqüência 'de bicicleta gosto de', proveniente das seguintes frases não pontuadas: "Gosto de brinca *de bicicleta gosto de* estuda matemática, ciência e história [...]". Tais grupos polilexicais, oriundos de falta de pontuação adequada foram, então, excluídos da análise realizada.

Em relação ao número de itens dos feixes lexicais a serem analisados, optou-se primeiramente por trabalhar com aqueles compostos por três palavras. Contudo verificou-se a existência de um grande número de feixes sobrepostos. Por exemplo, se o parâmetro escolhido fosse feixes compostos de três palavras, o feixe 'Rio de Janeiro' se sobreporia a um outro, 'no Rio de'. De forma a reduzir os vários casos de sobreposição, decidiu-se trabalhar com unidades compostas por quatro itens, seguindo o procedimento adotado por Biber, Conrad e Cortes. (op.cit.) para a língua inglesa.

Para o presente estudo, foram adotados critérios de freqüência e dispersão. Para ser considerado um feixe, um grupo de quatro itens lexicais deveria aparecer pelo menos três vezes e em três redações distintas. Tal procedimento excluiu da análise os usos idiossincráticos de seqüências lexicais.

Uma vez levantados os feixes lexicais, os mesmos foram classificados em termos de estrutura e função, utilizando-se as classificações estrutural e funcional propostas por Biber, Conrad e Cortes (op. cit.).

4) ANÁLISE

A análise da freqüência dos feixes lexicais por *corpus* sugere duas realidades distintas, como ilustra a Figura 4.

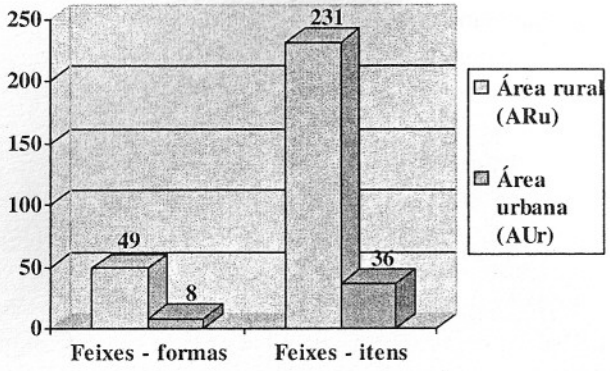

Figura 4
Distribuição de feixes nos *corpora* de pesquisa

Os dois *corpora* diferem claramente em termos do número absoluto de feixes lexicais (231 em **ARu** e 36 em **AUr**) e também em relação ao número de feixes diferentes (49 em **ARu** e somente 8 em **AUr**). No entanto, em números relativos, a diferença é mínima visto que as crianças da área rural utilizam cada feixe 4,71 vezes enquanto as da área urbana o fazem 4,5 vezes.

Em termos de estrutura, os *corpora* diferem em termos de distribuição das três categorias estruturais propostas por Biber, Conrad e Cortes (op.cit.), a saber, Tipo 1 (fragmentos de sintagmas verbais), Tipo 2 (fragmentos de orações subordinadas) e Tipo 3 (fragmentos de sintagmas nominais e/ou preposicionais). Em **ARu**, aproximadamente um em cada dois feixes lexicais é do Tipo 1. Além disso, 77,06% são dos Tipos 1 e 2. A grande freqüência destes dois tipos estruturais pode indicar que as crianças da área rural utilizam padrões oracionais repetidos ao redigirem suas composições.

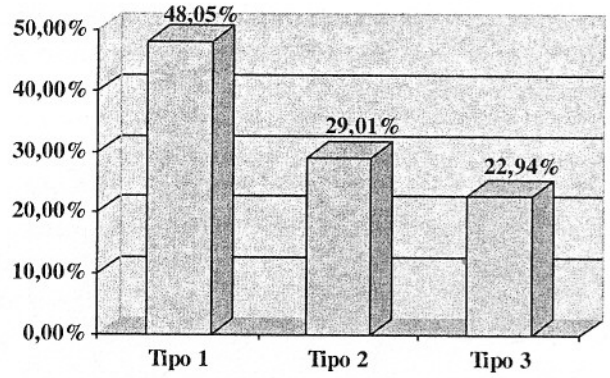

Figura 5
Distribuição estrutural de feixes lexicais em ARu

Em **AUr**, os feixes do Tipo 3 são os mais freqüentes, conforme mostra a Figura 6 abaixo:

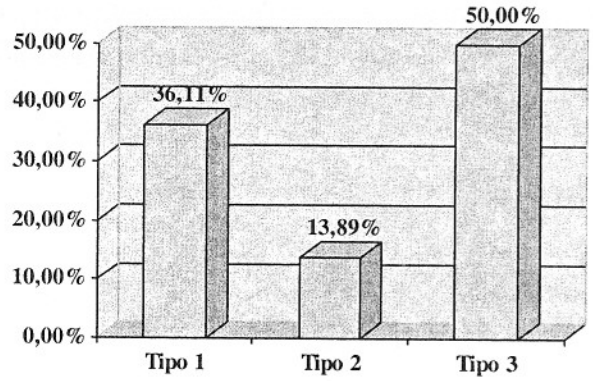

Figura 6
Distribuição estrutural de feixes lexicais em AUr

Considerando a natureza dos feixes dos Tipos 1 e 2, verifica-se que as crianças da área rural se expressam coletivamente através de padrões formados no nível da oração. Por outro lado, os feixes lexicais empregados pelas crianças da área urbana, distribuídos em termos de tópicos (50% de feixes do Tipo 3) e processos verbais (50% de feixes dos Tipos 1 e 2), não sugerem preferência por um determinado tipo.

Em termos de função, as crianças da área rural parecem priorizar feixes referenciais em detrimento de feixes que expressam atitude e que organizam o discurso, como aponta a Figura 7:

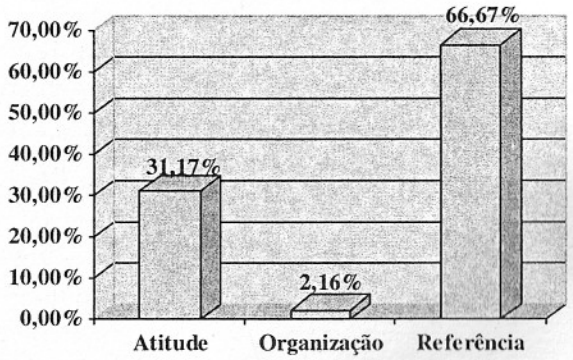

Figura 7
Distribuição funcional de feixes lexicais em ARu

No *corpus* relativo à área urbana, contudo, aparecem somente dois tipos funcionais de feixes: organizadores discursivos e referenciais, sendo que o último totaliza quase 89% das instâncias analisadas, como indica a Figura 8 abaixo.

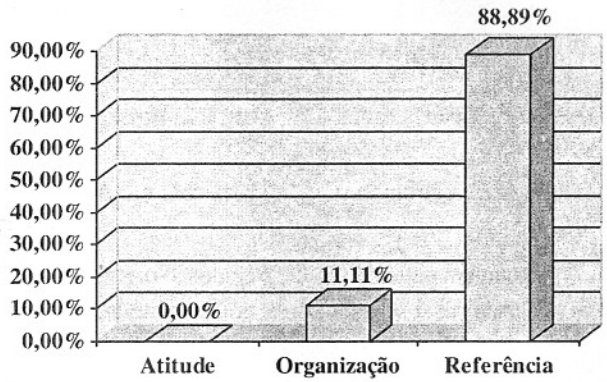

Figura 8
Distribuição funcional de feixes lexicais em AUr

Uma vez comparados os dois *corpora* em termos de função, nota-se a ausência de feixes atitudinais em **AUr**. O fato de não haver uma padronização na expressão de posicionamento não sinaliza, necessariamente, ausência de atitude, mas sim uma preferência pela expressão de posicionamento de forma individualizada.

Outro fator que deve ser ressaltado é que mesmo com 31,17% de feixes atitudinais, **ARu** contém tão somente uma única subcategoria de atitude: desejo. Em outras palavras, o único tipo de atitude que as crianças da área rural verbalizam de forma padronizada refere-se a seus desejos como em 'gosta muito de brincar', 'gosto muito da minha' e 'gosto da minha família'.

Com relação aos feixes organizacionais, os dois *corpora* apresentam perfis semelhantes. Apesar de desiguais em números, uma análise mais detalhada revela que as crianças recorrem a um mesmo feixe em versões diferentes em cada *corpus*: 'era uma vez uma' (**ARu**) e 'era uma vez um' (**AUr**).

Já os feixes referenciais, os mais freqüentes (66,67% em **ARu** e 88,89% no em **AUr**) apresentam alguma diferença em termos de distribuição de categorias. Em **ARu**, reúnem-se em torno de identificação/foco, lugar e posse[6], conforme ilustrado na Figura 9 abaixo:

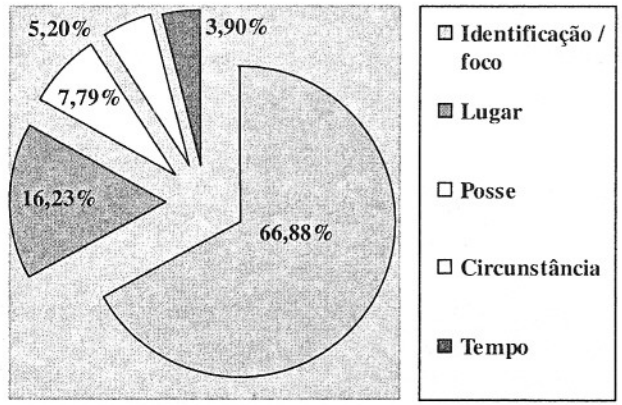

Figura 9
Classificação dos feixes referenciais em ARu

Em **AUr**, essa distribuição não acontece, já que apresenta somente quatro subcategorias: identificação/foco, lugar, tempo e circunstância. Assim como em **ARu**, a subcategoria identificação/foco é a mais freqüente. No entanto, a subcategoria de **lugar** aparece em quase o dobro.

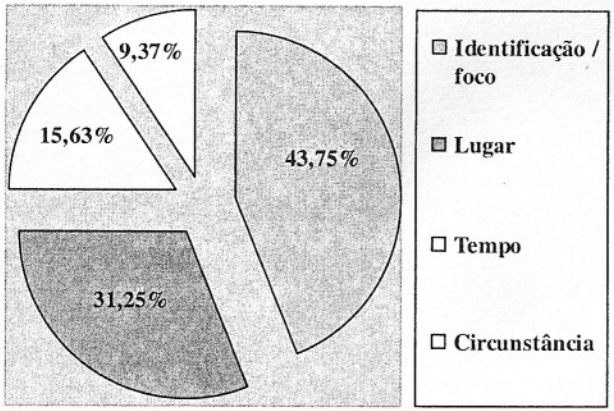

Figura 10
Classificação dos feixes referenciais em AUr

Esta distribuição dos feixes referenciais sugere que as crianças da área rural privilegiam a identificação de tópicos de forma padronizada. A configuração dos feixes lexicais nas composições dos participantes da área urbana mostra ausência da noção de posse e mais foco na questão de lugar.

Esta constatação sugere que a escrita das crianças da área rural se atém mais ao "idiom principle" (cf. SINCLAIR, 1991) no que se refere ao que gostam e ao que não gostam. As crianças da área urbana recorrem ao mesmo principio quando se referem a lugares e ao ambiente à sua volta.

5) Conclusão

Neste ponto cabe lembrar que feixes lexicais são construtos artificiais sistematizados pelo computador. Não constituem em si uma unidade lingüisticamente coerente, embora possam apontar para a existência

de combinações e funções lingüisticamente relevantes. Este trabalho buscou analisar textos produzidos por crianças de dois contextos diversos. A escolha dos contextos teve por base verificar como a produção escrita livre dos participantes revela, através do léxico por eles utilizado, sua visão de mundo; mais especificamente, como este mundo pode ser percebido através de escolhas lexicais padronizadas.

As diferenças encontradas tanto no número de feixes lexicais de cada grupo quanto nos tipos e funções destes mesmos feixes produzidos por cada um podem ser interpretadas de duas maneiras. A primeira está ligada à interface entre a escolha dos feixes e a atividade de processamento. Quando enunciadores se utilizam uma fórmula, recorrem a um repositório pré-existente, já utilizado pela comunidade a qual pertencem, ao invés de construírem um todo a partir de partes individuais. Se tomarmos Wray (2002, p. 74) como base para interpretação, o uso de seqüências padronizadas de formações polilexicais pode ser explicado em termos da economia de esforço para produzi-las e processá-las.

A segunda interpretação se baseia no fato de que, dadas as oportunidades de se expressar livremente, uma mesma comunidade pode compartilhar seqüências repetidas. Este achado corrobora a posição de Wray (2002, p. 74), quando afirma que "o inventário que um indivíduo guarda de seqüências holísticas é amplamente influenciado pelos padrões de uso corrente em uma determinada comunidade de falantes". No entanto, de acordo com Halliday (1985 et seq.), esta relação é dialética, ou seja, o repertório de padrões de uma comunidade de falantes pode também alimentar o inventário pessoal de cada membro desta comunidade. Os indivíduos selecionam para serem usadas seqüências que guardaram como um conjunto.

Como, então, estas duas interpretações podem explicar as diferenças encontradas nas produções escritas dos dois grupos de participantes deste estudo? Sugere-se, aqui, que a economia de processamento é a explicação menos plausível para a presença ou ausência de feixes lexicais. De acordo com os dados obtidos, as crianças do meio rural representam o mundo em termos de padrões ou seqüências repetidas, o que, em troca, os ajuda a expressar seu posicionamento e foco com relação a tópicos de interesse mútuo. Assim, reproduzem seqüências que podem ter ouvido na comunidade e incorporado como suas. Estes participantes parecem estar afinados com seu contexto, refletindo a forma como o grupo vê e sente o

mundo, fazendo planos, expressando desejo, posse e identificando suas famílias. Em contrapartida, a linguagem utilizada pelos participantes da área urbana revela um aspecto mais individualizado e, o que é mais intrigante, não manifesta posicionamento coletivo através de feixes lexicais, refletindo talvez uma visão de individualidade.

Os resultados do estudo não podem ser generalizados para se descrever a visão de mundo e a organização textual de crianças pobres de áreas rurais e urbanas brasileiras, necessitando ainda de um corpus de referência que se concentre nos textos de crianças de classes A e B. Entretanto, o presente estudo sugere que uma análise da forma e função de feixes lexicais pode revelar muito sobre formas preferidas de organização textual e visões coletivas de mundo.

ABSTRACT: This research analyses lexical bundles (cf. Biber, Conrad and Cortes 2004), found in essays written by children in the Brazilian 5th and 6th years (average age 12-14) from a rural and an urban community. The research objective was to identify possible profiles emerging from their writing in terms of their world views and their textual organization. To this end, two corpora were compiled consisting of 12, 205 and 14,662 words each. From these same corpora four-word lexical bundles were extracted and subsequently analyzed in terms of structure and function. The analyses reflect two different profiles. In contrast to the writing by the children from the rural area, those from the urban milieu failed to express a collective world view by means of repeated textual and functional patterns.

KEY-WORDS: identity; language; culture; corpus analysis; lexical bundles, writing.

6) REFERÊNCIAS BIBLIOGRÁFICAS

BERBER SARDINHA, T. *Lingüística de corpus*. São Paulo: Manole, 2004.

BERBER SARDINHA, T.; SHIMAZUMI, M. Approaching the assessment of performance unit archive of schoolchildren's writing from the point of view of corpus linguistics. In: TEACHING AND LANGUAGE CORPORA (TALC), 2., 1996. Disponível em: <http://www2.lael.pucsp.br/~tony/1998apu_talc.pdf>. Acesso em: 14 mar. 2005.

BIBER, D. Lexical bundles in academic speech and writing. In: LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004, p. 165-178.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at...: lexical bundles in university teaching and textbooks. *Applied Linguistics*, v. 25, n. 3, p. 371-405, 2004.

BIBER, D. et al. *Longman grammar of spoken and written English*. London: Longman, 1999.

DE COCK, S. et al. An automated approach to the phrasicon of EFL learners. In: GRANGER, S. (Ed.). *Learner English on computer*. London: Longman, 1998, p. 67-79.

COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998.

FAIRCLOUGH, N. *Analysing discourse: textual analysis for social research*. London: Routledge, 2003.

GRANGER, S. The computer learner corpus: a versatile new source of data for SLA research. In: _____ (Ed.). *Learner English on computer*. London: Longman, 1998a, p. 3-18.

GRANGER, S. Prefabricated patterns in advanced EFL writing: collocations and formulae. In: COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998b, p. 145-160.

GRANGER, S. (Ed.). *Learner English on computer*. London: Longman, 1998c.

HALLIDAY, M. *An introduction to functional grammar*. London: Edward Arnold, 1985.

HUNSTON, S. *Corpora in applied linguistics*. Cambridge: Cambridge University Press, 2002.

JORDÃO, S. et al. Violência no imaginário da criança. In: ZYNGIER, S.; VIANA, V.; FAUSTO, F. (Org.). *Venturas & desventuras: coletânea dos trabalhos do V ECEL*. Rio de Janeiro: Editora da Faculdade de Letras da UFRJ, 2005, p. 172-188.

LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004.

MCENERY, T.; WILSON, A. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1996.

MOON, R. Frequencies and forms of phrasal lexemes in English. In: COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998, p. 79-100.

RENOUF, A.; SINCLAIR, J. Collocational frameworks in English. In: AIJMER, K.; ALTENBERG, B. (Ed.). *English corpus linguistics*. London: Longman, 1991, p. 128-143.

SAMPSON, G. The structure of children's writing: moving from spoken to adult written norms. In: GRANGER, S.; PETCH-TYSON, S. (Ed.). *Extending the scope of corpus-based research*. Amsterdam: Rodopi, 2003, p. 177-193.

SCHIFFRIN, D. *Approaches to discourse analysis*. London: Blackwell, 1994.

SCOTT, M. *WordSmith tools*. Oxford: Oxford University Press, 1999.

SINCLAIR, J. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. Preface. In: LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004, p. 7-11.

STUBBS, M. *Words and phrases*. Oxford: Blackwell, 2001.

TELIYA, V. et al. Phraseology as a language of culture: its role in the representation of a collective mentality. In: COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998, p. 55-75.

WODAK, R. et al. The discursive construction of national identity. Edunburgh: Edinburgh University Press, 1999

WRAY, A. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press, 2002.

NOTAS

[1] Agradecemos ao Prof. John McH. Sinclair, que nos introduziu à noção de feixes lexicais, e o Prof. Michael Hoey pelos comentários. Ambos têm sido uma fonte constante de inspiração.

[2] Não há tradução consagrada para as expressões "idiom principle" e "principle of free-choice". As mesmas foram traduzidas por Berber-Sardinha e Walter Carlos Costa (comunicado pessoal) como princípio idiomático e da livre-escolha.

[3] Agradecemos à Suzana Jordão pela coleta das redações. O corpus coletado pertence ao banco de dados do Grupo de Pesquisa REDES.

[4] Bigramas e trigramas são os nomes dados pelos autores que trabalham com Lingüística de Corpus a conjuntos de dois e três itens lexicais respectivamente.

[5] Sampson cunhou o termo '*wordiness*' para descrever esta tendência.

[6] As categorias de 'posse' e 'circunstância' não constam na análise feita por Biber, Conrad e Cortes (2004). No entanto, estas categorias emergiram a partir dos dados da presente pesquisa.