IDENTIDADES SOCIAIS E LINGUÍSTICA DE CORPUS: UM ESTUDO DE TRÊS CONTEXTOS SOCIAIS

Tânia SHEPHERD
Universidade Estadual do Rio de Janeiro (UERJ)
Sônia ZYNGIER
Universidade Federal do Rio de Janeiro (UFRJ)

RESUMO

Este trabalho analisa contrastivamente escolhas lexicais feitas por alunos brasileiros da quinta série, a partir de três corpora eletrônicos contendo, cada um, 85 redações de temática livre. O trabalho provoca uma reflexão sobre o estudo léxico a partir da Linguística de Corpus para a compreensão das relações entre linguagem e identidades coletivas.

ABSTRACT

This paper investigates the lexical choices made by Brazilian 5th graders in three electronic corpora, each consisting of 85 unprompted compositions. The study indicate insights into the analyses of lexis through Corpus Linguistics in order to provide an insight into the potential interface between language and collective identity.

PALAVRAS-CHAVE

Análise crítica do discurso. Identidade. Léxico. Linguística de Corpus.

KEY-WORDS

Critical Discourse Analysis. Corpus Linguistics. Identity. Lexis.

Introdução

A Linguística de Corpus vem despertando interesse entre analistas do discurso pelo fato de lhes possibilitar observações empíricas em grande escala e de lhes dar condições de aliá-las a interpretações críticas, de forma transparente, mais objetiva e replicável. Desta forma, uma interface entre a Linguística de Corpus e a Análise Crítica do Discurso (A.C.D.), por exemplo, pode tornar esta última mais sistematizada: uma possível resposta à constante crítica feita por certos teóricos sobre interpretações baseadas tão somente nas intuições do analista e com base em poucos dados linguísticos (ver debate entre Widdowson, 1995 e 1996; Fairclough, 1996 e Toolan, 1997 e Widdowson, 2000).

Estudos que utilizam corpora eletrônicos trazem inúmeros benefícios para o analista. Primeiro, o computador possibilita a extração de padrões de linguagem de uso coletivo que, usando-se tão somente a leitura e análise manual de textos isolados, seriam difíceis de extrair. Além disso, como afirma Adolphs (2006:10), a análise que segue os princípios da Linguística de Corpus pode trazer à luz questões de natureza ideológica, como o caso da presente investigação.

De fato, análises de questões de identidades sociais são em geral realizadas através de métodos qualitativos (ver Wodak et. al., 1999) e se baseiam em textos coletados a partir de entrevistas e grupos de enfoque. Entretanto, os estudos sobre identidades sociais podem partir da comparação da frequência e tipo de léxico usado por determinados grupos, sendo que as unidades analíticas enfocadas podem ser tanto unidades lexicais isoladas, como fraseologias típicas (ver Shepherd et.al., 2006 e 2007). Investigações deste cunho, portanto, podem levar a uma compreensão das relações existentes entre a linguagem usada por determinados grupos, sua identidade e, consequentemente, sua cultura.

No presente estudo, propomos extrair possíveis padrões identitários a partir de análise de corpora eletrônicos constituídos de redações de crianças de três escolas com perfis distintos. A partir do estudo do léxico extraído desses corpora com auxílio de software, e utilizando alguns conceitos da Linguística de Corpus como o de 'colocados' e 'palavras-chave' (ver Sinclair, 1991; Hunston, 2002 e Berber-Sardinha, 2004), objetiva-se verificar o perfil sociocultural de cada um dos grupos de redações examinadas.

1. Pressupostos teóricos e objetivos do trabalho

O estudo de identidade é uma área complexa. Esta afirmação tem por base o uso das mais diversas metáforas para entender o fenômeno. Por exemplo, para Wodak et al. (1999:13-15), ao se tratar da questão de identidade, está-se andando em "gelo escorregadiço". Outra metáfora utilizada por esses autores se refere à identidade como "selva semântica". Já para Ricoeur (1992), ela se compara à imagem do DNA, no sentido de não se poder falar em uma identidade, mas na sua forma plural. Em outras palavras, assim como um mesmo indivíduo pode ter em seu DNA marcas genéticas que o fazem pertencer a vários conjuntos, sua identidade sócio-cultural atravessa vários sistemas (Ricoeur, 1992). O que estas metáforas têm em comum é a impossibilidade de se chegar a um consenso sobre o que é 'identidade' e sobre como definir o termo. Em última instância, qualquer tentativa de definição do que seja identidade parte de uma correlação entre linguagem e cultura.

Desta forma, o conceito de identidade é relacional, porque ela é ao mesmo tempo coletiva e individual. É relacional também porque pode ser auto-atribuída ou atribuída pelo 'outro'. A identidade se gera e se reproduz através do discurso, porque se trata de um "complexo de crenças, opiniões, comportamentos e manifestações afetivas comuns *internalizadas* no decorrer da *socialização*" (Wodak et al., 1999: 28, nosso grifo).

Apesar da variedade de metáforas para definir a área, os estudos sobre identidade apresentam uma relação marcante com linguagem como instrumento de cultura. Neste trabalho partimos do pressuposto de que uma análise da cultura pode ter como ponto de entrada as unidades lexicais, quer sejam palavras ou suas combinações (Teliya et al., 1998).

Portanto, o presente trabalho busca verificar como crianças de áreas diferentes expressam suas identidades enquanto grupos sociais, ou seja, como constroem suas identidades discursivamente. Para tanto, utilizam-se métodos da Linguística de Corpus. Esta perspectiva permite o foco no léxico para buscar possíveis padrões que vão sendo estabelecidos pelos participantes ao usarem a língua para construir um texto. A ênfase do trabalho se situa na frequência e na relevância de determinados padrões utilizados.

2. Metodologia

Há um número considerável de estudos sobre a escrita de crianças em língua materna. Entretanto, as investigações baseadas em corpora são esparsas. Podemos citar Sardinha & Shimazumi (1996), que estudaram a escrita de jovens aprendizes cuja língua materna é o inglês britânico, analisando uma amostra da prova de avaliação denominada APU (Assessment of Performance Test), em contraste com um corpus composto de escrita de adultos. O foco do estudo foram itens lexicais individuais. Outros estudos sobre corpora eletrônicos foram realizados por Sampson (2003; 2006), que observou sequências de duas palavras (bigramas) em textos escritos por crianças na década de 60. Mais recentemente, Shepherd, Zyngier e Viana (2006 e 2007) estenderam a unidade de investigação para analisar 'feixes lexicais', denominação dada por Biber et al. (2004) para sequências de três ou mais palavras constantes em um corpus, com frequência e distribuição pré-determinadas.

O presente estudo partiu da compilação de corpora específicos. Foram feitas três visitas a escolas distintas e coletadas redações de alunos da quinta série, sob as mesmas condições. Foi pedido que os alunos fizessem uma redação de temática livre na sala de aula, resultando em coleta de 85 composições de cada escola, as quais foram digitadas e identificadas em termos de escola, sexo e idade do aprendiz. As escolas escolhidas para a coleta foram: uma escola pública em área de risco do Rio de Janeiro (a Comunidade da Maré), uma escola particular no centro da mesma cidade e uma escola pública na localidade agrícola de Tocantins, em Minas Gerais. Partiu-se da premissa que o perfil das escolas, com três realidades distintas**,** traria à tona visões de mundo particulares.

Após a coleta de dados, todas as composições foram digitadas e apenas os erros de grafia foram eliminados, a fim de evitar problemas quando da leitura do corpus pelo programa WordSmithTools. Cada um dos corpora recebeu um rótulo: G1 foi dado ao corpus de redações dos participantes da comunidade da área de risco, G2 para o corpus de redações dos participantes da comunidade agrícola e G3 para o corpus dos participantes mais privilegiados economicamente.

3. Análise

Como ponto de entrada nos corpora foi extraída a densidade lexical a partir do número total de palavras (types) e do número de palavras diferentes (tokens). A Tabela 1 a seguir mostra uma "radiografia" de cada corpus, uma espécie de levantamento lexical de como cada corpus se caracteriza:

TABELA 1: Características lexicais de cada corpus

	G1	G2	G3
No. total de palavras	10.146	12.669	6.873
No. de palavras diferentes	2.098	2.153	1.549
Densidade lexical	20.68%	16.99%	22.97%

Nota-se que o G3 foi o grupo com menor número de palavras. No entanto, é o grupo que aponta maior densidade lexical, ou seja, o que apresenta maior variedade no uso do léxico. Isto pode ser entendido da seguinte forma: apesar de escreverem menos, os participantes de G3 se expressam com maior individualidade. Com o G2 se dá justamente o contrário. As redações desse grupo mostram menos recursos lexicais para a expressão de suas ideias.

Como segundo passo, buscou-se extrair a lista de frequência lexical contendo as sessenta palavras mais frequentes em cada corpus. Esta estratégia permitiu que se verificasse quais os itens lexicais (não gramaticais) em comum entre os três grupos, ou seja, quais seriam as palavras de conteúdo.

TABELA 2: Listagem de frequência lexical

	G1 - Maré	G2 - Tocantins	G3 - Centro do Rio
1	E	E	E
2	A	DE	DE
3	QUE	EU	EU
4	DE	A	SOU
5	O	QUE	É
6	EU	GOSTO	TENHO
7	PARA	MINHA	GOSTO
8	É	O	QUE
9	NÃO	É	MEU

continuação tabela 2	**G1 - Maré**	**G2 - Tocantins**	**G3 - Centro do Rio**
10	COM	MUITO	MUITO
11	OS	MEU	A
12	UM	UM	NÃO
13	NA	PARA	MAS
14	AS	DA	O
15	ELA	NA	UM
16	DA	NÃO	MINHA
17	MUITO	COM	ANOS
18	ELE	TEM	UMA
19	VOU	UMA	ADORO
20	BRASIL	DO	NO
21	GOSTO	ELE	EM
22	SUA	OS	COM
23	ATÉ	MAS	NA
24	GENTE	SOU	NOME
25	LÁ	ELA	ME
26	VIOLÊNCIA	NO	PARA
27	TE	EM	DO
28	NO19	SE	QUANDO
29	SE	PAI	MEUS
30	UMA	MAIS	MAIS
31	MAS	TAMBÉM	AMIGOS
32	MINHA	MEUS	FUTEBOL
33	MAIS	TENHO	JOGAR
34	POR	VOU	DA
35	CASA	MÃE	SÃO
36	QUANDO	ESCOLA	TAMBÉM
37	PORQUE	VIDA	LEGAL
38	TEM	ESTUDAR	OS
39	DO	AS	BEM
40	EM	QUANDO	AMO
41	DIA	ANOS	SER

continuação tabela 2	**G1 - Maré**	**G2 - Tocantins**	**G3 - Centro do Rio**
42	MEU	FAMÍLIA	OLHOS
43	MÃE	CASA	AS
44	ESCOLA	SER	COMO
45	PESSOAS	ME	DIA
46	ESTAVA	PORQUE	FAZER
47	ELES	VOCÊ	VER
48	FOI	DIA	POUCO
49	ERA	MORO	VEZES
50	TAMBÉM	POR	VOU
51	COISA	CHAMA	CASTANHOS
52	MUNDO	CIDADE	ESTOU
53	COMO	GOSTA	ETC
54	ISSO	COLEGAS	MÃE
55	DEPOIS	NOME	POR
56	ME	SÃO	NASCI
57	TER	IRMÃO	ACHO
58	VIDA	ERA	ÀS
59	PA49I	FOI	FAMÍLIA
60	VOCÊ	LÁ	PAI

A tabela 2 acima mostra uma interessante distribuição de frequência lexical. As palavras mais frequentes, como era de se esperar, são igualmente as mais frequentes na língua portuguesa como um todo (preposições, o 'que' integrante, artigos etc.). Mais abaixo, começam as palavras ditas de conteúdo (substantivos, adjetivos). A partir dessas palavras de conteúdo compartilhadas pelo menos por dois corpora (gosto, sou, pai, mãe, irmãos, colegas, etc), foram geradas as primeiras hipóteses, mais especificamente, de que o discurso dos participantes girava em torno de determinadas temáticas. Foram elaboradas perguntas a partir do léxico compartilhado.

- Quem sou eu?
- Como sou?
- Do que gosto?
- Qual meu núcleo familiar?
- Com quem me relaciono?
- Onde me situo?

Uma vez estabelecidos os pontos em comum, e como terceira estratégia, buscou-se, então, distinguir quais seriam as palavras-chave, ou seja, aquelas palavras típicas de cada corpus e que não são compartilhadas com outros corpora de referência. Para isso, verificou-se a chavicidade (*keyness*) de cada um dos corpora, conceito definido por Scott (1996) como sendo "as palavras de um corpus cuja frequência ... é maior do que em outro corpus de referência" . A Tabela 3 mostra o que foi encontrado:

TABELA 3: Palavras-chave

G1	G2	G3
violência	gosto	sou, tenho, gosto, adoro
crianças	minha	nome
Geiza	raposa	futebol, amigos
Brasil	Tocantins	castanhos, olhos

O que transparece nos dados da Tabela 3 é a palavra "violência" em primeiro lugar nos escritos dos participantes de G1, seguido de "crianças", "Geiza" (nome de uma das participantes) e "Brasil". Nota-se que o imaginário das crianças aqui gira em torno de seu país, com a situação de risco em que se encontram. O pronome "eu" (pronomes possessivos) não aparecem como termos-chave no corpus da Maré. Já os escritos dos participantes de G2, o grupo de Tocantins, revela o mesmo número de palavras-chave, quatro, sendo que "gosto" e "minha" mostram um interesse por si. Em especial, o pronome possessivo 'minha' como palavra-chave do G2 revela a posse como item-chave. 'Tocantins'

também é palavra-chave de G2, o que pode sugerir uma visão mais restrita acerca do espaço em que vivem. O termo "raposa" aparece por terem alguns alunos decidido contar a fabula da raposa e das uvas.

Um terceiro quadro totalmente diverso se delineia com os participantes de G3. Aqui, há nove palavras-chave que caracterizam o grupo da escola particular. Além do olhar voltado para si (uso da primeira pessoa como em G2 nas terminações verbais), há um acúmulo de verbos ("sou", "tenho", "gosto", "adoro"). Os verbos *ser* e *ter* se caracterizam por expressar processos relacionais (ou processos de 'ser'); os verbos *gostar*, *adorar* caracterizam processos mentais (ou processos de experienciar internamente). O grupo busca uma identificação ("nome"), e tem como chave os conceitos de amigos e futebol, além da preocupação com a descrição física ("castanho", "olhos").

Em suma, enquanto a chavicidade de G1 mostra uma preocupação com o que está 'lá fora' (a violência, o Brasil), a chavicidade de G3 se concentra no eu e no círculo imediato dos alunos (quem são eles, do que gostam e características físicas).

Dando prosseguimento à análise e de volta às perguntas geradas a partir da listagem dos dados na Tabela 2 acima, buscou-se, então verificar quais seriam os colocados da palavra "sou", ou seja, as palavras que ocorrem no entorno da palavra de busca. Os resultados são apresentados na Tabela 4 abaixo:

TABELA 4: Colocados de "sou"

	G1	G2	G3
sou	ϕ nome	+ um/a menino/a	+ adjetivo
sou	+ nome	+menino/ garoto/pessoa	+ adjetivo
sou	+ nome	+ uma pessoa	+advérbio +adjetivo

O que chama a atenção na Tabela 4 é o fato de os participantes de G1 não se identificarem nominalmente, enquanto que os outros dois grupos claramente usam "sou+nome". Por outro lado, os participantes de G1 e G2 se identificam quanto a gênero e idade ("sou" + menino/a), o que não ocorre em G3. Outro aspecto relevante nesta tabela é a ocorrência de advérbios que aparecem em G3, com a função de maximizadores ("eu sou uma pessoa muito legal"). A posição deste grupo se traduz no uso de uma mesma expressão por dois participantes: "Eu sou mais eu".

Buscando observar a noção de posse dos participantes, verificaram-se os colocados de "tenho", dispostos na Tabela 5 abaixo:

TABELA 5: Colocados de "tenho"

G1	G2	G3
# anos	# anos	# anos
	numeral + amigos, irmão/irmã, pai, coelho, etc.	amigos, irmão, irmã
		olhos, cabelos + adjetivo

Segundo a Tabela 6, os participantes de G1 só usam "tenho" quanto à idade. Nada mais dizem de "ter". Já os outros dois grupos se representam menos desvalidos, mostrando ter amigos e família. Além disso, no grupo G3, de alunos de escola particular, nota-se o 'ter' apontado também para a descrição física. A partir destes resultados, buscou-se, então, verificar como os participantes definiam a família, conforme a Tabela 6 abaixo:

TABELA 6: Colocados de "mãe" e "pai"

G1	G2	G3
	tenho, gosto muito	amo, gosto, adoro
sua	minha	minha
assassinado	mãe	amigos

A Tabela 6 mostra que a noção de desvalimento se faz mais forte ainda em G1, que não só não fala em família, mas quando o faz, a referência é a terceiros ("sua"). Volta a entrar a questão da violência aqui, com a palavra "assassinado" como colocado de "pai de família". Já G3 estende a noção de família para incluir amigos. Neste sentido, então, buscou-se verificar os colocados de "amigo", conforme descrito na Tabela 7 abaixo:

TABELA 7: Colocados de "amigo"

G1	G2	G3
não tenho	Tenho	tenho adoro gosto
	meu/s	meu minha
verdadeiro		

Confirmando os resultados das tabelas anteriores, G1 usa "não tenho" na proximidade de "amigo". Para eles, amigo é algo abstrato, visto que dizem querer um "verdadeiro amigo", um ideal a ser atingido, mas que não existe na realidade. Os participantes de G2 meramente revelam ter amigos enquanto os de G3 mostram posicionamento assertivo. Além de "tenho" como colocado de amigos, essas crianças usam também "adoro" e "gosto" como colocados de amigo.

As tabelas acima, portanto, respondem em parte às perguntas geradas a partir da lista de frequência. Mostram que G1 não reflete uma preocupação com o "eu" ou com um nicho familiar. Situa-se num país, mas não indica sentir-se de posse de nada. Sua preocupação maior, caracterizada pelas palavras-chave, gira em torno da violência. Já os participantes de G3 voltam o olhar para si, descrevendo-se. Dizem-se possuidores de amigos e avaliam positivamente seu entorno. G2 se coloca numa posição intermediária: nem negligente de si e do ambiente, nem completamente voltado para si.

Conclusão

Este estudo pretendeu fazer um primeiro levantamento lexical a partir do que se pode inferir sobre a noção de identidade de três grupos em contextos diferentes. O que se pode afirmar a partir das observações acima é que enquanto o léxico dos aprendizes da Comunidade da Maré (G1) revela algum desvalimento emocional, o dos outros dois grupos de crianças inclui a apreciação por si e pelo próximo.

A partir dos escritos de G2 e G3 pode-se notar alguma segurança e autoestima, o que não transparece nos dados de G1. Pelo contrário, há uma tendência para a desvalorização pessoal (não ter amigos e definir um hipotético amigo verdadeiro), além de uma visão de família consistindo de mãe somente (pai de família ocorre em grande frequência com 'assassinado'). Esta análise nos permite afirmar provisoriamente que enquanto G1 se vê só em um mundo hostil, G2 e G3 se cercam da família e da cidade.

Além da validação do método da Linguística de Corpus, que, com apoio em dados empíricos, permitiu a percepção de realidades sociais diversas a partir da coleta de vozes individuais, este estudo ressalta a experiência de uma realidade sociológica de privação de núcleos básicos de cidadania em G1.

Trata-se de um estudo pioneiro que abre caminhos para novos desenvolvimentos. Para tanto, é preciso aumentar-se o corpus a fim de confirmar os achados deste trabalho ou revelar possíveis exceções. Caberia, também, incluírem-se outros contextos para verificar se os padrões ora encontrados se repetem em áreas com perfis semelhantes. Quanto ao tipo de dado coletado, poder-se-ia mudar o gênero utilizado, solicitando aos participantes que escrevessem cartas ou outro tipo de texto, ao invés de redações de tema livre, para ver se os padrões se repetem.

Em última instância, o presente trabalho ratifica a relação entre linguagem e identidade social. Esta conclusão não poderia ter sido atingida com tanta evidência, caso não se tivesse tido o apoio da análise de textos eletrônicos, que permitiu, a partir da coleta de vozes individuais, chegar-se às vozes coletivas e trazer à tona as implicações identitárias submersas nas entrelinhas do discurso.

Referências

ADOLPHS, S. **Introducing Electronic text Analysis:** a practical guide for language and literary studies. London: Routledge, 2006.

BERBER SARDINHA, T. **Linguística 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., CONRAD, D & CORTES, V. **"If you look at…lexical bundles in university teaching and textbooks"**. Applied Linguistics 25(3):371-405, 2004.

FAIRCLOUGH, N. **"A Reply to Henry Widdowson's 'Discourse Analysis: A Critical View"**. Language & Literature 5(1): 49-56, 1996.

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

RICOEUR, P. **Oneself as another. Chicago:** University of Chicago Press.

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. 177-93.

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

SHEPHERD, T.; ZYNGIER, S.; VIANA, V. **"Feixes lexicais e visões de mundo: um estudo sobre corpus"**. Matraga 13 (19): 125-140, 2006.
_____. **A Tale of Two Cities':** Lexical Bundles as Indicators of Linguistic Choices and Socio-cultural Traces. In: Jeffries, L. et al. **Stylistics and Social Cognition**, Amsterdam: Rodopi, 2007.

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

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.

TOOLAN, M. **"What Is Critical Discourse Analysis and Why Are People Saying Such Terrible Things About It?"** Language & Literature 6(2): 83-103, 1997.

WIDDOWSON, H. **"On the Limitations of Linguistics Applied"**. Applied Linguistics 21 (1): 3-25, 2000.
_____. **"Reply to Fairclough: Discourse and Interpretation:** Conjectures and Refutations". Language & Literature 5(1): 57-69, 1996.
_____. **"Review of Fairclough's Discourse and Social Change"**. Applied Linguistics 16(4): 510-516, 1995.

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