TRATAMENTO DE *CORPORA* INFORMATIZADOS POR PROGRAMAS DE ANÁLISE LINGUÍSTICA PARA ESTUDOS DO PORTUGUÊS FALADO DE SÃO PAULO[1]

Zilda Maria Zapparoli

Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo (USP)

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)

Endereço postal: Rua Gaspar Moreira, 31 – CEP 05505-000 São Paulo – SP – Brasil

Endereço eletrônico: zmz@usp.br

Resumo − Neste artigo, dedicamo-nos a estudos do português falado de São Paulo a partir do tratamento de c*orpora* informatizados por programas especiais de análise linguística. Os *corpora* são gerados no *Sistema CorPor*, Sistema de Banco de Dados Relacional, que contém Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo (São Paulo, Campinas, Itu), organizadas, relacionadas e armazenadas em função de anotações linguísticas e extralinguísticas. Com vistas a investigações do comportamento do léxico, na exploração do *subcorpus* do português falado culto de São Paulo, aplicamos programas de computador desenvolvidos especialmente para análise linguística e que reúnem recursos computacionais, matemáticos e estatísticos. Nessa perspectiva, o trabalho insere-se no campo da *Linguística Informática*, apoiando-se em áreas que partilham a crença nos resultados positivos advindos da interação entre Linguística e Informática.

[1] ZAPPAROLI, Zilda Maria. Tratamento de *corpora* informatizados por programas de análise lingüística para estudos do português falado de São Paulo. *Boletim da Academia Galega da Língua Portuguesa.* Santiago de Compostela: Academia Galega de Língua Portuguesa, n. 3, 2010, p. 87-112. ISSN 1888-8763.

Palavras-chave − Linguística Informática, Tecnologias Informatizadas nos Estudos Linguísticos, Projeto CorPor, Sistema CorPor, Sistema de Banco de Dados Relacional, Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo, *Corpora* Eletrônicos do Português Falado de São Paulo.

Abstract – The aim of this article is to study the Portuguese language as spoken in São Paulo with the use of electronically processed c*orpora* and linguistic analysis programs. The *corpora* are generated in the *CorPor* System, a Relational Database System that contains orthographic and phonetic information about the Portuguese language as spoken in the State of São Paulo (São Paulo City, Campinas, Itu), organized, listed and stored taking into account linguistic and extralinguistic annotations. In order to carry out investigations into the lexicon, in the exploration of this particular *subcorpus* − the variety of Portuguese spoken in São Paulo by educated speakers −, we make use of computer programs designed specifically for linguistic analysis, which combine computational, mathematical and statistical resources. This study, therefore, belongs in the field of *Linguistic Informatics*, drawing support from the various areas that share the belief in the positive results of the interaction between Linguistics and Informatics.

Index Terms: Linguistic Informatics, Data Processing Technologies in Linguistic Studies, Corpor Project, Corpor System, Relational Database System, Databanks of Phonetic and Orthographic Information about The Portuguese Language as Spoken in São Paulo, Electronic Corpora of the Portuguese Language as Spoken in São Paulo.

INTRODUÇÃO

No presente artigo, dedicamo-nos a tratamento de c*orpora* informatizados por programas especiais de análise linguística para estudos do português falado de São Paulo. Os *corpora* são gerados no *Sistema CorPor*, Sistema de Banco de Dados Relacional, que inclui Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo (São Paulo, Campinas, Itu), compatíveis com os sistemas computacionais atuais, a partir dos dados coletados para a tese de doutorado (1980) e das informações contidas nas Bases então geradas em sistemas de computadores de grande porte.

As informações das Bases estão organizadas, relacionadas e armazenadas em função de anotações linguísticas (especificidades da língua oral, categorias de encontros

fônicos intervocabulares) e extralinguísticas que foram controladas na seleção dos 216 informantes que forneceram material linguístico para a constituição da amostra (região de origem, sexo, escolaridade, faixa etária e nível socioeconômico) e na produção dos diálogos (formal e informal).

Neste trabalho, a título de estudos de dados autênticos do português falado paulista da década de 1970 e com vistas a investigações do comportamento do léxico, utilizamos programas para análise linguística na exploração do *subcorpus* do português falado culto de São Paulo.

1 PRESSUPOSTOS TEÓRICO-METODOLÓGICOS

Numa dimensão mais ampla, o trabalho insere-se na área da *Linguística Informática*. A *Linguística Informática*, como linha de investigação científica, propõe-se, de um lado, à *utilização de recursos da Informática na Linguística* para o armazenamento, processamento e recuperação quantitativa e qualitativa de informações linguísticas; de outro, à *utilização de recursos da Linguística na Informática* para o desenvolvimento de sistemas que exigem equipes multidisciplinares, nas quais se incluem linguistas, como sistemas de tradução automatizada, sistemas de ensino de línguas naturais a distância, sistemas de produção e reconhecimento de línguas naturais.

Nessa perspectiva, o trabalho parte da utilização de *recursos da Informática na Linguística* para a composição e exploração de *Corpora* do Português em Sistema de Banco de Dados, cujos resultados servirão de subsídios às áreas que se servem de *recursos da Linguística na Informática*, a exemplo do processamento automático da língua portuguesa.

Ainda, concebendo a *Linguística Informática* como abrangendo as diferentes áreas em que as tecnologias informatizadas estão relacionadas aos estudos da linguagem – *Linguística de Corpus, Linguística Computacional e Processamento de Língua Natural* –, o trabalho enquadra-se mais particularmente nos propósitos da *Linguística de Corpus* em uma de suas preocupações, que constitui a condição *sine qua non* para a sua existência – construção de *corpora* eletrônicos a partir de textos e discursos reais. A *Linguística de Corpus* é vista, aqui, mais do que um simples instrumento de trabalho, por acreditar-se que o emprego das tecnologias informatizadas – base da *Linguística de Corpus* – na exploração de grandes quantidades de dados da língua em uso pode trazer informações inéditas sobre as línguas naturais.

O trabalho fundamenta-se, também, nos quadros teóricos: (a) da *Linguística Descritiva*, em sua preocupação com o que é dito ou escrito, por quem, onde e quando; (b) da *Linguística Aplicada*, em sua concepção atual – que vai além de sua aplicação ao ensino/aprendizagem de línguas –, enquanto área multidisciplinar, dedicada às situações de uso da língua e, pois, ao desenvolvimento de pesquisas a partir da análise de *corpora*; (c) da *Linguística Conversacional*, no cuidado em respeitar, na transcrição dos dados, as características específicas do discurso oral, evitando-se, na medida do possível, as normas tradicionais da linguagem escrita; (d) da *Fonética* e da *Fonologia*, nos critérios que nortearam a transcrição e o exame do comportamento fonológico dos encontros fônicos que se dão na juntura lexical, ou seja, nos limites de duas ou mais fronteiras de palavras; (e) da *Sociolinguística*, nas variáveis extralinguísticas que foram controladas na seleção dos informantes que forneceram material linguístico para a constituição dos *corpora*, com consequente oferecimento dos perfis dos dialetos e situações de uso contemplados através da técnica baseada em *corpus*; (f) da *Lexicologia*, nos critérios que orientaram a questionada definição e delimitação da palavra para a segmentação do enunciado nos seus constituintes léxicos, bem como a constituição de léxicos; (g) *da Linguística Textual e da Análise do Discurso Oral*, na descrição lexical quantiqualitativa, que oferece subsídios para estudos do texto e do discurso.

2 CORPUS DE ESTUDO

Os textos orais transcritos submetidos ao tratamento computacional foram produzidos por falantes cultos paulistanos – superior completo –, de ambos os sexos, de diferentes faixas etárias, nas situações formal (entrevista) e informal (conversação) de interação dialógica.

O *corpus* de estudo constitui-se de quatro variáveis – T1, T2, T3, T4 –, reunidas de acordo com o sexo e com as condições extraverbais de produção do diálogo – entrevistas e conversações:

- T1 – SPFSCF – seis informantes paulistanos, filhos de pais paulistanos – SP, do sexo feminino – F, com nível superior completo – SC, em situação formal de diálogo – F;
- T2 – SPFSCI – seis informantes paulistanos, filhos de pais paulistanos – SP, do sexo feminino – F, com nível superior completo – SC, em situação informal de diálogo – I;

- T3 – SPMSCF – seis informantes paulistanos, filhos de pais paulistanos – SP, do sexo masculino – M, com nível superior completo – SC, em situação formal de diálogo - F;
- T4 – SPMSCI - seis informantes paulistanos, filhos de pais paulistanos - SP, do sexo masculino - M, com nível superior completo - SC, em situação informal de diálogo - I.

3 PROCEDIMENTOS METODOLÓGICOS

Para estudos do português falado culto de São Paulo, aplicamos dois programas de computador desenvolvidos especialmente para análise linguística e que reúnem recursos computacionais, matemáticos e estatísticos – *WordSmith Tools* e *Stablex PC*.

De autoria de Mike Scott, Universidade de Liverpool, o programa *WordSmith Tools* (1998) é publicado pela Oxford University Press e distribuído via *World Wide Web* (http://www.lexically.net/wordsmith). Em sua quinta versão, disponível para PC/Windows 2000 ou superior, incluindo Windows 7, disponibiliza diversos recursos – *Wordlist*, *Concord*, *Keywords*, *Splitter*, *Text Converter*, *Dual Text Aligner*, *Viewer* – para tarefas específicas de análises de textos.

O programa *Stablex* (STA – de statistique, TAB – de tableaux, LEX – de lexique e T...EX – de texte), de autoria de André Camlong e Thierry Beltran, Universidade de Toulouse II, inicialmente desenvolvido para *Macintosh* (Toulouse, Teknea, 1991), conta, atualmente, com a sua versão PC (São Paulo, Pirus Tecnologia, 2004). Os seus recursos – geração de léxicos, indexação, extração de sequências e concordâncias, lematização, tratamento estatístico – foram desenvolvidos em função de um modelo de análise lexical, textual e discursiva – *método matemático-estatístico-computacional de análise de textos* de André Camlong. Trata-se, por conseguinte, da aplicação de um programa que serve de ferramenta para um método de análise de textos.

O método é fundado na matemática e na estatística paramétrica (estatística descritiva); possibilita o estudo descritivo, objetivo e indutivo do texto; permite a análise quantiqualitativa do léxico, que indica apontamentos para a análise textual e discursiva. Nele, o texto é o ponto de referência: as operações estatísticas partem do texto e, por sua vez, refletem o texto.

Dedicamo-nos mais ao emprego do *Stablex*, não por privilegiarmos um programa em detrimento do outro, mas pelo fato de o *Stablex* responder, de modo especial, aos interesses deste estudo.

Dentre as vantagens em função dos estudos que desenvolvemos, destacamos o fato de as listas originárias da análise de textos pelo *Stablex* exibirem o léxico do *corpus* e dos textos que o integram – e não apenas do *corpus* como um todo, caso do *WordSmith* –, o que facilita uma visão contrastiva do todo – *corpus* – em relação às partes – textos que integram o *corpus* – e das partes em relação ao todo, bem como das partes entre si, um dos objetivos de nossos estudos.

O *WordSmith* assume importância fundamental no tratamento de *corpora* extensos, voltados à construção de dicionários e de glossários, aos exames dos padrões linguísticos, aos estudos ligados à tradução e ao gênero.

3.1 Preparação dos Textos

A fim de que o *corpus* possa receber tratamento estatístico, a aplicação dos programas pressupõe uma preparação[2] dos textos em função dos objetivos do estudo – tarefa manual, que exige tempo e cuidado.

Os procedimentos utilizados foram os seguintes:

- reconstituição de sintagmas[3] – para preservar a unidade, o trabalho de geração de léxicos é precedido por uma tarefa manual de reconstituição dos sintagmas (unidades lexicais compostas – conjunto de palavras que funcionam como uma unidade lexical única; duas ou mais palavras com um único referente) nos textos, através da substituição dos espaços em branco por traços de união. Por exemplo: nomes de pessoas e de cidades – *Getúlio-Vargas, Juscelino-Kubitscheck-de-Oliveira, Zélia-Borges, São-Paulo, Rio-Grande-do-Sul* –; designação de escolas, cursos, corporações, repartições, edifícios, estabelecimentos, aeroportos, igrejas, exposições, congressos – *Universidade-de-São-Paulo, Editora-Abril, Associação-dos-Professores-de-Francês, Aeroporto-de-Congonhas* –; títulos de livros, jornais, revistas, artigos e produções artísticas, literárias e científicas em geral (filmes, peças, músicas, telas, teses etc.) – *Quando-é-Preciso-Ser-*

---

[2] Com base em Camlong, 1996:9-12.

[3] Na geração de léxicos computacionais, os critérios de composição é uma das questões delicadas. O trabalho foi orientado por critérios operacionais de fixidez de formas compostas, tais como institucionalização, uso, restrições distribucionais, cristalização, ou seja, de combinações fixas – dois ou mais vocábulos que ocorrem juntos com frequência, que têm significado próprio e que funcionam como unidades lexicais compostas, diferentemente de sintagmas formados livremente.

Homem, Sonhos-de-um-Sedutor –; números compostos – quarenta-e-quatro, mil-novecentos-e-quarenta-e-oito, duas-e-meia –; expressões fixas diversas – graças-a-Deus, nossa-mãe, fim-de-semana, ponto-de-vista.

- reconstituição de *a-gente* no emprego de pronome de 1ª pessoa do plural, enquanto referenciador textual – envolvimento do interlocutor (pronome de solidariedade) –, dado o seu papel no estudo das categorias de pessoa. *A-gente* é, pois, considerado como um item lexical. Esse procedimento de dicionarizar como um bloco facilita o estudo das concordâncias com a-gente: 3ª pessoa do singular e 1ª pessoa do plural.
- reconstituição de *não-é* com valor fático, enquanto marcador conversacional.

3.2 Recursos Empregados e Resultados Alcançados

Apresentamos, a seguir, recursos empregados e resultados da aplicação da abordagem de análise quantiqualitativa do léxico ao *corpus* de estudo, com tabelas, figuras e gráficos correspondentes. A partir dos resultados, tecemos comentários sobre itens ou conjuntos de itens.

3.2.1 Levantamento lexical com constituição de léxicos de frequência

A partir dos textos utilizados para a análise, geram-se, inicialmente, três léxicos de frequência, através de uma operação que consiste no levantamento automático exaustivo dos itens lexicais das variáveis, em que, respeitada a distribuição entre as variáveis, os itens lexicais são classificados por ordem alfabética, por ordem crescente de frequência e por ordem decrescente de frequência. Os léxicos relacionam, na primeira coluna, a ordem de classificação; na segunda, os itens lexicais; na terceira, a frequência total e, nas demais, a frequência por variável, ou seja, a distribuição da frequência total pelas variáveis em estudo.

Segue amostra do léxico delta (início e fim da listagem): a classificação por ordem decrescente de frequência põe em relevo o vocabulário gramatical nas altas frequências (*eu, de, que, é, não, o, e, a, um, muito, para, em* etc.), seguido pelo vocabulário nocional, em que se destaca o temático (*tenho, dizer, gosto, acho, casa* etc.), com os hápax no final da listagem. O léxico delta mostra a preferência do ponto de vista da massa lexical (frequência de emprego).

Tabela 1 – Léxico delta

Ordem	Léxico	Total	T1	T2	T3	T4
1	eu	504	158	157	108	81
2	de	370	127	70	90	83
3	que	368	97	106	84	81
4	é	272	56	73	46	97
5	não	268	66	78	49	75
6	o	248	59	62	63	64
7	e	212	54	39	55	64
8	a	168	38	41	54	35
9	um	156	37	35	51	33
10	muito	135	78	29	19	9
11	para	130	34	35	32	29
12	em	118	36	17	29	36
13	uma	105	25	28	26	26
14	mas	103	26	31	21	25
15	então	91	26	19	17	29
16	porque	91	24	30	15	22
17	né	88	15	35	17	21
18	na	86	19	24	21	22
19	você	81	5	36	6	34
20	com	76	20	22	14	20
21	se	75	12	20	9	34
22	do	72	29	16	12	15
23	mais	69	23	20	10	16
24	tenho	68	16	19	21	12
25	tem	66	19	16	14	17
26	no	64	18	21	13	12
27	lá	60	9	26	9	16
28	por	59	16	21	15	7
29	me	59	21	16	12	10
30	dizer	58	16	13	16	13
31	assim	58	33	14	8	3
32	sei	57	15	23	8	11
33	também	54	21	12	15	6
34	vou	53	17	8	25	3
35	como	52	12	15	14	11
36	ele	51	5	26	0	20
37	gosto	50	38	6	6	0
38	acho	50	20	12	13	5
39	os	49	7	7	21	14
40	aqui	49	12	9	7	21
41	da	49	14	16	8	11
42	casa	47	23	2	16	6
43	quer	47	13	12	10	12
44	sabe	47	9	30	0	8
45	a-gente	44	11	16	12	5
46	está	42	7	25	3	7
47	ou	42	14	6	19	3
48	agora	42	11	12	7	12
49	isso	41	9	16	4	12
50	só	41	7	15	4	15
51	ela	39	2	28	0	9

Ordem	Léxico	Total	T1	T2	T3	T4
2090	atualizada	1	0	1	0	0
2091	treinar	1	0	1	0	0
2092	branco	1	0	0	1	0
2093	tomo	1	0	0	1	0
2094	totalmente	1	0	1	0	0
2095	brasile	1	0	1	0	0
2096	total	1	0	0	1	0
2097	transferir	1	0	0	0	1
2098	base	1	0	0	0	1
2099	toc-toc	1	0	0	0	1
2100	augusta	1	0	0	1	0
2101	auditório	1	1	0	0	0
2102	atuando	1	1	0	0	0
2103	treze	1	0	0	0	1
2104	atualizados	1	0	0	1	0
2105	barrada	1	0	1	0	0
2106	banho	1	0	0	1	0
2107	banque	1	1	0	0	0
2108	chamado	1	0	0	0	1
2109	chateado	1	0	0	1	0
2110	chateia	1	0	0	0	1
2111	chato	1	0	0	0	1
2112	chave	1	0	0	0	1
2113	chamam	1	0	0	0	1
2114	centro	1	0	0	1	0
2115	chamar	1	1	0	0	0
2116	terá	1	0	0	1	0
2117	central	1	0	0	0	1
2118	cérebro	1	0	1	0	0
2119	chamada	1	0	0	1	0
2120	teoplastos	1	0	0	0	1
2121	tentativa	1	0	0	1	0
2122	chamaremos	1	1	0	0	0
2123	caindo	1	1	0	0	0
2124	transmitir	1	0	0	1	0
2125	teatrais	1	0	0	1	0
2126	barba	1	0	0	1	0
2127	tratamentos	1	0	1	0	0
2128	trazem	1	0	0	1	0
2129	autores	1	1	0	0	0
2130	aviões	1	0	0	1	0
2131	auxiliava	1	0	0	0	1
2132	avaliar	1	1	0	0	0
2133	avisa	1	0	1	0	0
2134	caderneta	1	0	0	1	0
2135	tomando	1	0	0	1	0
2136	trazer	1	0	0	1	0
2137	ba	1	0	1	0	0
2138	avisam	1	0	1	0	0
2139	bahia	1	0	0	0	1
2140	dificuldade	1	0	1	0	0

A observação dos dados fornecidos pelos léxicos de frequência não oferece informação relativa ao valor do item lexical na construção do texto e do discurso. São dados brutos – puramente quantitativos –, que mostram, apenas, o recenseamento dos itens lexicais com sua frequência absoluta (número real de ocorrências) no *corpus* e nas variáveis, não sendo, pois, suficientes para uma descrição científica do *corpus* em virtude da diferença de extensão de cada texto.

3.2.2 Criação da Tabela de Distribuição de Frequências – TDF

A TDF[4] é gerada a partir do léxico delta e conserva, na matriz, apenas os dados numéricos da população lexical recenseada nos textos, representando, pois, a massa lexical, o *status* da população estudada (cálculo aritmético – tratamento quantitativo). A massa lexical, reproduzida na TDF sob a forma de matriz na escala funcional graduada de 1 a 504 (a mais alta frequência de emprego), é um numero absoluto de ocorrências ou de frequências de emprego que afetam todos os itens lexicais do *corpus* recenseado e classificadas segundo critérios fatoriais e categoriais.

Apresenta-se, a seguir, a TDF das quatro variáveis, onde se lê: na primeira coluna, a ordem de classificação de 1 a 75, correspondendo à ordem decrescente da segunda coluna; na segunda, as frequências de emprego arranjadas por ordem decrescente: da mais elevada – 504 para o vocábulo *eu*, conforme Tabela Delta, que lista os vocábulos classificados por ordem decrescente – à mais baixa – os vocábulos de frequência 1 (ou hápax[5]); na terceira, o número de vocábulos referente a cada frequência, conforme a contagem feita na Tabela Delta; na quarta, o número total de ocorrências da linha, que é o produto dos dados das duas colunas precedentes, ou seja, o produto das frequências pelo número correspondente de vocábulos, que é também a soma dos efetivos registrados nas colunas seguintes; nas quatro colunas seguintes, a distribuição das ocorrências, isto é, o número de ocorrências de cada variável. A parte superior da tabela traz os totais e os valores de "p" (probabilidade de ocorrência de cada item lexical em cada variável) e de "q" (probabilidade contrária).

[4] Sobre os procedimentos utilizados para a geração da TDF, consultar Camlong, 1996:28.
[5] Hápax são itens lexicais com frequência 1 numa variável e 0 nas demais.

Tabela 2 − Tabela de Distribuição de Frequências - TDF

Textos: 4			Totais:	11357	3084	2923	2654	2696
Linhas: 75			p		0,272	0,257	0,234	0,237
			q		0,728	0,743	0,766	0,763
Ordem	***Frequência***	***Número***	***Ocorrência***		***T1***	***T2***	***T3***	***T4***
1	504	1	504		158	157	108	81
2	370	1	370		127	70	90	83
3	368	1	368		97	106	84	81
4	272	1	272		56	73	46	97
5	268	1	268		66	78	49	75
6	248	1	248		59	62	63	64
7	212	1	212		54	39	55	64
8	168	1	168		38	41	54	35
9	156	1	156		37	35	51	33
10	135	1	135		78	29	19	9
11	130	1	130		34	35	32	29
12	118	1	118		36	17	29	36
13	105	1	105		25	28	26	26
14	103	1	103		26	31	21	25
15	91	2	182		50	49	32	51
16	88	1	88		15	35	17	21
17	86	1	86		19	24	21	22
18	81	1	81		5	36	6	34
19	76	1	76		20	22	14	20
20	75	1	75		12	20	9	34
21	72	1	72		29	16	12	15
22	69	1	69		23	20	10	16
23	68	1	68		16	19	21	12
24	66	1	66		19	16	14	17
25	64	1	64		18	21	13	12
26	60	1	60		9	26	9	16
27	59	2	118		37	37	27	17
28	58	2	116		49	27	24	16
29	57	1	57		15	23	8	11
30	54	1	54		21	12	15	6
31	53	1	53		17	8	25	3
32	52	1	52		12	15	14	11
33	51	1	51		5	26	0	20
34	50	2	100		58	18	19	5
35	49	3	147		33	32	36	46
36	47	3	141		45	44	26	26
37	44	1	44		11	16	12	5
38	42	3	126		32	43	29	22
39	41	2	82		16	31	8	27
40	39	1	39		2	28	0	9
41	37	2	74		19	16	27	12
42	36	1	36		13	7	6	10
43	35	1	35		7	3	23	2
44	34	3	102		46	17	21	18
45	33	1	33		10	9	4	10
46	32	5	160		53	27	37	43

Ordem	*Frequência*	*Número*	*Ocorrência*	*T1*	*T2*	*T3*	*T4*
47	30	2	60	14	7	30	9
48	29	3	87	7	30	15	35
49	28	2	56	19	11	13	13
50	27	2	54	33	10	8	3
51	26	5	130	41	25	45	19
52	24	2	48	10	6	10	22
53	23	2	46	7	13	18	8
54	22	5	110	41	15	29	25
55	21	5	105	30	34	22	19
56	20	4	80	17	20	25	18
57	19	1	19	5	7	4	3
58	18	6	108	34	27	27	20
59	17	7	119	43	19	41	16
60	16	11	176	51	49	38	38
61	15	5	75	21	21	15	18
62	14	9	126	27	33	40	26
63	13	9	117	38	12	34	33
64	12	13	156	46	42	33	35
65	11	11	121	37	34	26	24
66	10	5	50	13	11	15	11
67	9	13	117	28	19	30	40
68	8	21	168	47	38	59	24
69	7	27	189	34	56	39	60
70	6	44	264	59	53	68	84
71	5	63	315	79	95	69	72
72	4	89	356	97	111	71	77
73	3	132	396	87	113	88	108
74	2	367	734	166	187	169	212
75	1	1211	1211	326	281	307	297

Um exame paralelo entre a TDF e o Léxico Delta permite a identificação dos itens lexicais. Assim: 504 *eu*, 370 *de*, 368 *que*, 272 *é*... (leitura vertical da primeira e segunda colunas do Léxico Delta) até 326 vocábulos próprios de T1, 281 de T2, 307 de T3, 297 de T4 (leitura horizontal da última linha da TDF).

Observe-se que, se considerados o início e o fim da tabela, a relação entre as colunas 2 e 3 inverte-se: no início, há uma quantidade menor de itens lexicais para uma frequência de emprego elevada; no fim, um número elevado de itens lexicais para uma frequência de emprego baixa, decorrendo, daí, a importância dos itens lexicais de frequência 1 e dos hápax. É importante reportar-se aos léxicos para a identificação dos itens lexicais de mesma frequência de emprego. Por exemplo, linha 15 da TDF: há 2 itens que ocorrem 91 vezes no *corpus*. Pelo léxico delta, identificamos quais são os itens que têm essa mesma frequência no *corpus* – *então* e *porque* (número igual de ocorrências no *corpus*, mas diferente nas variáveis).

A observação dos dados da TDF não oferece, também, informação relativa ao valor dos itens lexicais nos textos. Os dados brutos fornecidos pela TDF, dado retratarem, simplesmente, a distribuição dos itens lexicais pelas variáveis, não são suficientes para uma descrição científica do *corpus* e, por isso, não permitem fazer nenhuma comparação, nem formular nenhuma hipótese. Nessa perspectiva, é preciso transformar a TDF em TDR.

3.2.3 Criação da Tabela de Desvios Reduzidos – TDR

Criada a partir da TDF, na TDR[6] todos os itens lexicais são medidos com a mesma unidade, a do desvio padrão[7], em relação a um centro de gravidade – o ponto de equilíbrio –, em que a média é reduzida a zero[8] (cálculo algébrico – tratamento quantiqualitativo). A TDR dá, pois, o *peso lexical*, um valor algébrico do desvio reduzido (Z).

Segue a TDR das quatro variáveis, onde se lê: na primeira coluna, a ordem de classificação de 1 a 75, correspondendo à ordem decrescente da segunda coluna; na segunda, o número de frequências de emprego, arranjadas por ordem decrescente (extraídas da TDF); na terceira, o número total de desvios reduzidos na linha (ou das quatro variáveis), correspondentes às frequências de emprego; nas quatro colunas seguintes, o número de desvios reduzidos de cada variável em relação às frequências de emprego. No cabeçalho da tabela, calculam-se a soma dos desvios reduzidos (Σ Zs) do total e das variáveis, o desvio reduzido médio (Z Médio), que é a soma dos desvios reduzidos dividida pelo número de linhas, o quadrado do Z Médio (Khi2) e a normalidade da distribuição lexical. Enquanto as leituras verticais, coluna por coluna, referem-se à especificidade de cada variável a partir de uma descrição do conjunto, as leituras horizontais, linha por linha, dizem respeito à comparação entre as variáveis, permitindo a comparação das diferenças ou das preferências de emprego de elementos ou de grupos de elementos de um texto a outro. A visão combinada dos dois eixos – vertical e horizontal – fornece os traços comuns e os traços distintivos entre os textos que integram o *corpus*.

---

[6] Para a confecção das tabelas de desvios reduzidos, consultar Camlong, 1996:34-35.

[7] Medida de dispersão dos valores de uma variável em torno de sua média – medida universal por excelência.

[8] Em matéria de estatística descritiva, o centro de gravidade é reduzido a zero.

Tabela 3 − Tabela de Desvios Reduzidos – TDR

		Total	*T1*	*T2*	*T3*	*T4*
	∑(Zs)	*-0,319*	*3,728*	*5,430*	*-0,112*	*-9,365*
	Z Médio	-0,004	0,050	0,072	-0,001	-0,125
	Khi2	0,023	0,002	0,005	0,000	0,016
	Normalidade	1,000				
Ordem	***Frequência***	***Desvio Red.***	***T1***	***T2***	***T3***	***T4***
1	504	-0,178	2,117	2,780	-1,029	-4,046
2	370	-0,056	3,101	-3,000	0,434	-0,591
3	368	-0,023	-0,343	1,346	-0,246	-0,779
4	272	0,085	-2,435	0,415	-2,517	4,622
5	268	-0,003	-0,931	1,261	-1,967	1,634
6	248	0,065	-1,191	-0,266	0,757	0,765
7	212	0,097	-0,551	-2,445	0,886	2,207
8	168	0,085	-1,322	-0,395	2,687	-0,885
9	156	0,085	-0,965	-0,943	2,752	-0,759
10	135	-0,345	8,000	-1,131	-2,552	-4,662
11	130	0,005	-0,257	0,309	0,336	-0,383
12	118	0,042	0,819	-2,815	0,310	1,728
13	105	0,031	-0,771	0,218	0,337	0,246
14	103	-0,012	-0,436	1,012	-0,715	0,127
15	182	-0,024	0,096	0,366	-1,845	1,358
16	88	0,009	-2,132	3,012	-0,898	0,028
17	86	0,036	-1,055	0,460	0,230	0,402
18	81	0,068	-4,246	3,851	-3,395	3,858
19	76	-0,016	-0,165	0,640	-1,019	0,528
20	75	0,081	-2,172	0,184	-2,327	4,395
21	72	-0,102	2,504	-0,682	-1,344	-0,579
22	69	-0,079	1,154	0,617	-1,742	-0,107
23	68	0,027	-0,672	0,416	1,464	-1,181
24	66	-0,008	0,298	-0,278	-0,414	0,385
25	64	-0,047	0,174	1,295	-0,578	-0,938
26	60	0,002	-2,117	3,118	-1,532	0,533
27	118	-0,086	1,026	1,396	-0,125	-2,382
28	116	-0,153	3,653	-0,606	-0,682	-2,518
29	57	-0,072	-0,142	2,524	-1,665	-0,788
30	54	-0,067	1,939	-0,591	0,766	-2,181
31	53	0,034	0,805	-1,772	4,095	-3,093
32	52	0,019	-0,661	0,513	0,606	-0,438
33	51	-0,009	-2,786	4,123	-3,944	2,598
34	100	-0,271	6,935	-1,770	-1,032	-4,404
35	147	0,090	-1,283	-1,101	0,321	2,153
36	141	-0,106	1,271	1,485	-1,383	-1,479
37	44	-0,026	-0,321	1,612	0,612	-1,929
38	126	-0,040	-0,444	2,154	-0,094	-1,656
39	82	-0,014	-1,556	2,500	-2,913	1,955
40	39	-0,060	-3,093	6,579	-3,449	-0,097
41	74	0,050	-0,286	-0,810	2,667	-1,521
42	36	-0,036	1,208	-0,864	-0,950	0,570
43	35	0,139	-0,952	-2,323	5,920	-2,506
44	102	-0,131	4,074	-2,095	-0,664	-1,446
45	33	-0,032	0,407	0,202	-1,527	0,886
46	160	-0,007	1,698	-2,564	-0,073	0,932

Ordem	*Frequência*	*Desvio Red.*	*T1*	*T2*	*T3*	*T4*
47	60	0,125	-0,666	-2,493	4,875	-1,591
48	87	0,123	-4,008	1,866	-1,351	3,615
49	56	-0,023	1,140	-1,043	-0,027	-0,092
50	54	-0,229	5,610	-1,213	-1,485	-3,140
51	130	0,012	1,124	-1,697	3,030	-2,445
52	48	0,100	-0,985	-2,098	-0,415	3,598
53	46	0,085	-1,820	0,391	2,526	-1,012
54	110	-0,024	2,386	-2,903	0,742	-0,249
55	105	-0,061	0,326	1,557	-0,585	-1,359
56	80	0,067	-1,188	-0,151	1,666	-0,260
57	19	-0,028	-0,082	1,107	-0,239	-0,814
58	108	-0,039	1,011	-0,175	0,401	-1,275
59	119	-0,017	2,202	-2,438	2,857	-2,639
60	176	-0,045	0,544	0,638	-0,557	-0,670
61	75	-0,024	0,165	0,448	-0,689	0,053
62	126	0,074	-1,445	0,116	2,222	-0,819
63	117	0,055	1,295	-3,830	1,455	1,135
64	156	-0,043	0,655	0,339	-0,654	-0,382
65	121	-0,057	0,847	0,594	-0,489	-1,009
66	50	0,031	-0,184	-0,604	1,108	-0,289
67	117	0,103	-0,784	-2,350	0,581	2,656
68	168	0,034	0,239	-0,925	3,599	-2,880
69	189	0,090	-2,833	1,224	-0,888	2,587
70	264	0,142	-1,756	-2,104	0,917	3,085
71	315	-0,015	-0,828	1,795	-0,614	-0,368
72	356	-0,075	0,039	2,349	-1,527	-0,935
73	396	0,067	-2,320	1,274	-0,539	1,653
74	734	0,129	-2,765	-0,161	-0,220	3,276
75	1211	0,073	-0,184	-2,017	1,630	0,643

Na TDR, o cálculo do Z é feito a partir da TDF, portanto, por frequências observadas no *corpus*. Na TDR, tem-se o Z por item, quando há um único item para uma dada frequência, ou o Z por conjunto de itens que se relacionam pela mesma frequência no *corpus*, quando há mais de um item para a mesma frequência. Neste caso, para discriminar-se o valor do Z de cada item, usa-se a técnica da discriminação (desagrupamento de itens), recurso disponível na macro do *Excel* que acompanha o programa *Stablex*.

A estatística é essencialmente um instrumento de medida contrastiva, em que o todo se define em relação às partes e as partes, em relação ao todo.

Observemos o valor do pronome *eu*, item lexical que encabeça o léxico delta e a TDR.

Tabela 4 – Valor lexical do pronome *eu*

	Item	*Total*	*T1*	*T2*	*T3*	*T4*
	EU					
Frequência		504	158	157	108	81
Valor		-0,178	2,117	2,78	-1,029	-4,046

Acusa-se um desvio reduzido negativo de -4,046 para a quarta variável e de -1,029 para a terceira variável, em oposição a um valor positivo para a primeira e segunda variáveis – 2,117 e 2,780 respectivamente –, o que significa que o emprego do pronome pessoal *eu* é deficitário nos diálogos cujos informantes são do sexo masculino em relação ao emprego que tem nas outras duas variáveis, cujos falantes são do sexo feminino.

Com base na tabela (extraída da última linha da TDR) e no gráfico seguintes, consideremos o valor lexical dos hápax.

Tabela 5 – Valor lexical dos hápax

Valor lexical dos hápax					
	Total	*T1*	*T2*	*T3*	*T4*
Frequência	1211	326	281	307	297
Valor	0,073	-0,184	-2,017	1,63	0,643

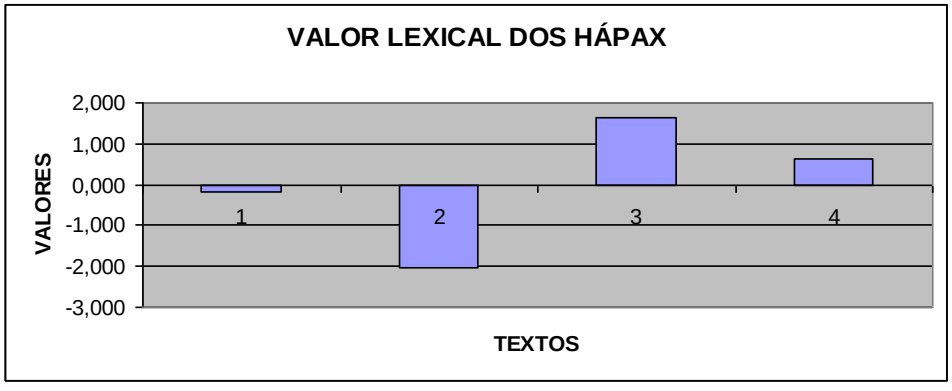

Figura 1 - Valor lexical dos hápax

Regra geral, a frequência = 1/1 (hápax) coincide, em média, com o desvio reduzido = +1,96 ~ +2, que, por sua vez, coincide com o final do vocabulário

preferencial. O *corpus* em estudo foge dessa regra geral, visto que os hápax, com valor inferior a =1,96, ocupam a zona do vocabulário básico.

Nota-se que, no discurso feminino (variáveis T1 e T2), os hápax apresentam valores negativos (T1 = -0,184 e T2 = -2,017), o que aponta para a repetição no sentido de reforço temático, em oposição ao discurso masculino (variáveis T3 e T4), em que os hápax, com valores positivos (T3 = 1,630 e T4 = 0,643), sinalizam busca do melhor item descritivo ou narrativo.

A exploração exaustiva dos valores da TDR, classificando-se o Z em ordem crescente ou decrescente pelo *corpus* e pelas variáveis, permite avaliações contrastivas do todo em relação às partes e das partes em relação ao todo e, a partir daí, a seleção de lemas que serão objetos de lematizações – reagrupamentos de itens pela técnica da lematização –, recurso também disponível na macro do *Excel* que acompanha o *Stablex*.

3.2.4 Criação de Léxicos Preferenciais (Tabela de Valores Lexicais) – LP

Obtido pela aplicação da técnica da discriminação do valor de cada item lexical em função do conjunto da variável e do *corpus*, o léxico preferencial dá a distribuição preferencial dos itens lexicais, ou seja, a ordenação dos itens lexicais por ordem decrescente de preferência de emprego no texto (ordem decrescente de desvios reduzidos); resulta, pois, do cálculo do desvio reduzido (Z) de todos os itens lexicais de cada variável.

Enquanto a TDR expõe o Z por frequência observada no *corpus*, o LP dá o Z por item lexical. Se há um único item para uma dada frequência, o valor do Z da TDR coincide com o valor do Z do LP. O LP é ideal para lematizações temáticas.

Em função dos pesos, fixados pela escala de valores dos desvios reduzidos dos itens lexicais, o LP decompõe-se em três zonas – ou variedades de vocabulários –, que correspondem às noções estatísticas de vocabulário preferencial, básico e diferencial.

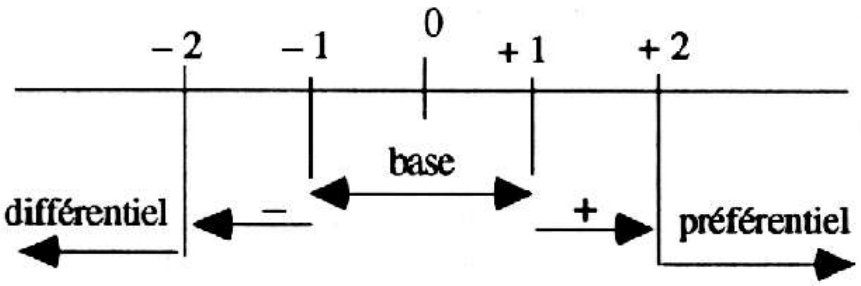

Figura 2 – Variedades de vocabulário (*apud* CAMLONG, 1996:127)

- vocabulário preferencial ($Z \geq +1,96$ ou $+2$) – inclui os itens de maior relevo ligados à temática e à articulação do discurso, indicadores de uma escolha privilegiada; o desvio reduzido positivo acusa um uso privilegiado;

- vocabulário básico, comum, vulgar, ou banalizado, cujo Z tende para zero (Z entre -1,96 e +1,96), em que ainda se podem distinguir o vocabulário básico com tendência positiva ($+1 \leq Z \leq +2$) e o vocabulário básico com tendência negativa ($-2 \leq Z \leq -1$), delimitando, mais especificamente, a zona centrada do vocabulário comum ($-1 \leq Z \leq +1$) – é o vocabulário próprio do *corpus*, imagem dos empregos correntes da amostra considerada; o desvio reduzido em torno da média reduzida a zero revela um uso normal;

- vocabulário diferencial ($Z \leq -1,96$ ou -2) – reflete uma escolha deficitária; o desvio reduzido negativo retrata um uso rejeitado.

Esses vocabulários permitem-nos destacar as temáticas na zona preferencial, as gramáticas específicas nas zonas preferenciais igualmente, assim como o léxico representativo da expressão banalizada, o suporte da temática, nas zonas básicas. Do mesmo modo, destaca-se, nas zonas básicas, o suporte gramatical básico.

Importa, ainda, observar outras variedades de vocabulário, para melhor discernir as características da fala estudada. Assim, além das variedades de vocabulários que correspondem à estratificação do léxico em zonas de acordo com os empregos preferenciais dos itens lexicais, podem-se considerar três outros tipos:

- vocabulário exclusivo (ou particular) – itens exclusivos de uma dada variável;

- vocabulário de exclusão – inexistente na variável considerada (em comparação com as outras variáveis); reúne, pois, itens lexicais ausentes numa variável, mas presentes em outra ou em outras variáveis. Pode possibilitar o estudo do silêncio em sua qualidade física e, a partir dele, do confronto entre ausência e presença, o estudo do silêncio no processo de significação;

- vocabulário específico – vocabulário de síntese que, em função de finalidades do estudo, obedece a critérios de reagrupamento de vocábulos (por associação léxica, semântica ou temática) pela técnica da lematização, com o cálculo do valor (peso) do novo vetor obtido.

O exame do léxico preferencial a partir dos vocabulários que o compõem destaca as características de emprego dos itens lexicais e os elementos fundamentais da estrutura temática e articuladora do discurso.

Apresentamos, a seguir, uma amostra do léxico preferencial da variável T1 (partes inicial e final de cada tipo de vocabulário).

Relacionam-se, nas listas, cinco tipos de colunas: a primeira refere-se à ordem de classificação; a segunda – do léxico – registra os itens lexicais; a terceira, a sua frequência no total do *corpus*; a quarta, a sua frequência na variável; e a quinta, o seu valor lexical – valor pelo desvio reduzido; assinalam-se os vocabulários em que se subdivide o léxico – preferencial, básico e diferencial – e destacam-se, em amarelo, os itens lexicais particulares ou exclusivos.

Tabela 6 - Léxico preferencial da variável T1 – SCFSPF

Ordem	*Léxico*	*Total*	*T1*	*Valor*
VOCABULÁRIO PREFERENCIAL				
1	**muito**	135	78	8,000
2	**gosto**	50	38	7,766
3	**trabalho**	27	23	6,780
4	**assim**	58	33	5,093
5	**ahn**	22	15	4,327
6	**televisão**	26	16	3,942
7	**profissional**	8	7	3,838
8	**teatro**	16	11	3,741
9	**campo**	5	5	3,662

Ordem	*Léxico*	*Total*	*T1*	*Valor*
10	**fora**	9	7	3,415
11	**bastante**	13	9	3,411
12	**às**	34	18	3,381
13	**casa**	47	23	3,357
14	**horário**	4	4	3,276
15	**depende**	4	4	3,276
16	**varia**	4	4	3,276
17	**pais**	4	4	3,276
18	**olha**	16	10	3,179
19	**de**	370	127	3,101

Ordem	Léxico	Total	T1	Valor
20	vezes	34	17	2,995
21	dos	17	10	2,936
22	entro	3	3	2,837
78	incrível	2	2	2,316
79	atrai	2	2	2,316
80	estudando	2	2	2,316
81	assistindo	2	2	2,316
82	disponho	2	2	2,316
83	pouco	28	13	2,293
84	hora	8	5	2,248
85	peça	6	4	2,176
86	difícil	6	4	2,176
87	série	4	3	2,151
88	possibilidade	4	3	2,151
89	trabalha	4	3	2,151
90	sair	4	3	2,151
91	interessante	4	3	2,151
92	amigas	4	3	2,151
93	raro	4	3	2,151
94	nenhum	4	3	2,151
95	eu	504	158	2,117
96	estou	32	14	2,111
97	ambiente	21	10	2,109
98	posso	11	6	2,043
99	acho	50	20	2,042
	VOCABULÁRIO BÁSICO			
100	também	54	21	1,939
101	ao	22	10	1,930
102	tipo	7	4	1,784
103	dentro	12	6	1,779
104	tanto	12	6	1,779
105	geral	12	6	1,779
106	família	12	6	1,779
107	minha	32	13	1,713
108	atual	5	3	1,651
109	papo	5	3	1,651
110	jornal	5	3	1,651
111	embora	5	3	1,651
112	ai	5	3	1,651
113	possa	5	3	1,651
114	momento	5	3	1,651
115	fizeram	5	3	1,651
116	problema	18	8	1,649
117	comer	1	1	1,638

Ordem	Léxico	Total	T1	Valor
118	supõe	1	1	1,638
119	somos	1	1	1,638
120	sílvio-santos	1	1	1,638
121	costumamos	1	1	1,638
122	solteira	1	1	1,638
803	exemplo	14	2	-1,083
804	aula	24	4	-1,155
805	o	248	59	-1,191
806	uhn	10	1	-1,220
807	vai	20	3	-1,222
808	sabe	47	9	-1,234
809	curso	30	5	-1,292
810	outra	16	2	-1,318
811	ainda	16	2	-1,318
812	a	168	38	-1,322
813	são-paulo	11	1	-1,347
814	falo	11	1	-1,347
815	só	41	7	-1,451
816	tinha	12	1	-1,466
817	mim	12	1	-1,466
818	nem	12	1	-1,466
819	fui	12	1	-1,466
820	está	42	7	-1,528
821	noite	13	1	-1,578
822	vez	13	1	-1,578
823	três	13	1	-1,578
824	bom	32	4	-1,864
825	pode	16	1	-1,880
	VOCABULÁRIO DIFERENCIAL			
826	ano	23	2	-1,990
827	os	49	7	-2,025
828	éh	29	3	-2,035
829	foi	29	3	-2,035
830	dá	18	1	-2,060
831	lá	60	9	-2,117
832	né	88	15	-2,132
833	se	75	12	-2,172
834	não-é	21	1	-2,307
835	é	272	56	-2,435
836	ele	51	5	-2,786
837	era	29	1	-2,870
838	ela	39	2	-3,093
839	você	81	5	-4,246

Verifiquemos o lugar ocupado pelo vocábulo *gente* (Z = -0,026), no emprego *a gente,* usado no lugar de *nós* e sem excluir ocorrências de *nós* (Z = -0,027)**.**

Tabela 7 – Valor lexical dos itens *a-gente* e *nós*

	Item	***Total***	***T1***	***T2***	***T3***	***T4***
	a-gente					
Frequência		44	11	16	12	5
Valor		-0,026	-0,321	1,612	0,612	-1,929
	nós					
Frequência		34	11	8	6	9
Valor		-0,027	0,681	-0,294	-0,788	0,374

Os vocábulos *a gente* e *nós*, ambos implantados no vocabulário fundamentalmente básico, no seu uso conhecido como plurais de modéstia, em vez do singular *eu***,** denotam um discurso centrado numa primeira pessoa que busca tornar o ouvinte partícipe do seu discurso, da sua argumentação. Têm, assim, o papel de instaurar a solidariedade entre os interlocutores e, pois, de aumentar a força argumentativa do discurso, dada a sua importância na composição discursiva – item lexical que tem força argumentativa pelo envolvimento dos interlocutores.

O uso do item a-gente é uma característica do discurso oral dos informantes, refletindo um uso normal no grupo, dado integrar o vocabulário básico de todos eles: vocabulário fundamentalmente básico em T1 e T3 (Z = -0,321 e Z = 0,612, respectivamente) – entrevista formal –, vocabulário básico com tendência positiva em T2 (Z = 1,612) e vocabulário básico com tendência negativa em T4 (Z = -1,929). Deve-se notar que, em T4, o item lexical reflete um uso deficitário, visto estar quase implantado no vocabulário diferencial.

3.2.5 Constituição de vocabulários específicos pela técnica da lematização

A técnica da lematização[9] (agrupamento de itens) é um procedimento de síntese parcial do léxico, que consiste na redução de itens lexicais a um único vetor centrado em torno de uma raiz temática – campo temático –, de um sema – campo semântico –, ou de um vocábulo-chave (por exemplo, reagrupamento pelas diferentes flexões de um vocábulo,

[9] Recurso disponível na planilha 5 da macro do Excel que acompanha o *Stablex*.

ou pela mesma categoria gramatical de nomes, adjetivos, verbos, pronomes etc.) – campo lexical. A determinação da frequência e do peso do novo vetor permite considerar o lugar que o conjunto de itens ocupa no *corpus* e na variável e, portanto, sua carga semântica, temática, argumentativa e discursiva. Enfocando-se o léxico preferencial de cada variável e recorrendo-se ao vocabulário específico fundado na lematização, podem-se, pois, focalizar blocos considerados mais importantes. Esse procedimento permite observar a escolha estratégica que o falante faz dos itens lexicais para a construção de seu discurso – é o estudo da composição lexical do texto a serviço da descoberta das perspectivas discursivas.

Tecemos comentários sobre o emprego do pronome pessoal *eu* nas quatro variáveis, tendo observado o seu uso deficitário nos discursos produzidos pelos informantes do sexo masculino. A aplicação do cálculo do Z ao agrupamento dos itens *a-gente* (= nós) e *nós*, conforme tabela 8, reforça a rejeição ao uso da 1ª pessoa por falantes do sexo masculino. Para esse cálculo, usa-se a regra da lematização.

Tabela 8 – Lematização dos itens *a-gente* e *nós*

Lematização						
	Itens	***Total***	***T1***	***T2***	***T3***	***T4***
	a-gente	44	11	16	12	5
	nós	34	11	8	6	9
Frequência		78	22	24	18	14
Valor		-0,038	0,209	1,016	-0,061	-1,202

Observemos o agrupamento dos itens *a-gente* e *nós* com outros pronomes de primeira pessoa.

Tabela 9 – Lematização de pronomes de primeira pessoa

Lematização						
	Itens	***Total***	***T1***	***T2***	***T3***	***T4***
	eu	504	158	157	108	81
	nós	34	11	8	6	9
	a-gente	44	11	16	12	5
	meu	36	13	7	6	10
	meus	16	7		7	2
	minha	32	13	7	9	3
	minhas	7	1	1	4	1
Frequência		673	214	196	152	111
Valor		-0,181	2,708	2,009	-0,480	-4,418

O agrupamento de pronomes de primeira pessoa corrobora a observação já feita no que diz respeito à rejeição ao emprego de primeira pessoa nos discursos masculinos, em especial, em situação informal de interação, em oposição à sua predileção nos discursos femininos, em especial, em situação formal.

Observemos, agora, o lugar ocupado pelo *não-é?* e pela sua variante *né?* – elementos fáticos característicos do discurso oral, enquanto marcadores conversacionais de confirmação da atenção do interlocutor.

Tabela 10 – Lematização dos itens *não-é?* e *né?*

Lematização						
	Itens	***Total***	***T1***	***T2***	***T3***	***T4***
	não-é	21	1	9	8	3
	né	88	15	35	17	21
Frequência		109	16	44	25	24
Valor		0,036	-2,929	3,494	-0,107	-0,422

O agrupamento desses dois itens pela técnica da lematização mostra que, no *corpus*, eles ocupam o vocabulário fundamentalmente básico; em T3 e em T4 – discursos produzidos por falantes do sexo masculino –, o vocabulário básico; nos discursos produzidos por falantes do sexo feminino, há uma oposição marcante na sua utilização – em T1 (diálogo formal), integram a zona do vocabulário diferencial, revelando emprego deficitário, e, em T2 (diálogo informal), a zona do vocabulário preferencial, denotando uso privilegiado.

A aplicação da técnica da discriminação permite-nos verificar o valor desses itens separadamente.

Tabela 11 – Discriminação dos itens *não-é?* e *né?*

	Item	***Total***	***T1***	***T2***	***T3***	***T4***
	não-é					
Frequência		21	1	9	8	3
Valor		0,064	-2,307	1,794	1,595	-1,018
	né					
Frequência		88	15	35	17	21
Valor		0,009	-2,132	3,012	-0,898	0,028

A observação do valor lexical do *não-é?* e do *né?* separadamente denota que as duas formas são rejeitadas no discurso de T1 – diálogo formal; T2 e T4 preferem a

forma sintética – diálogo informal, e T3, a analítica – diálogo formal. Ou seja, informantes do sexo feminino privilegiam o emprego da forma sintética em situação informal de interação dialógica, rejeitando as duas formas em situação formal; em informantes do sexo masculino, as duas formas fazem parte do seu vocabulário básico, com a preferência pela analítica em situação formal e pela sintética em situação informal.

O agrupamento poderia incluir outros marcadores do tipo, como certo? sabe?, *olha*, *olhe.*

Os textos estão permeados pelos itens *ah*, *ahn, uhn, éh, eh, tsi, tsu, ih*, sem valor cognitivo, mas que, típicos da língua falada, marcam as condições de produção do discurso, no caso, sinalizando hesitações do falante. Trata-se, pois, de recurso utilizado para o planejamento do texto, enfim, estratégia em busca de orientação discursiva e de itens lexicais adequados ao foco temático:

Tabela 12 – Lematização de marcadores conversacionais de hesitação

Lematização						
	Itens	***Total***	***T1***	***T2***	***T3***	***T4***
	ah	26	7	7	7	5
	ahn	22	15	4	1	2
	uhn	10	1	4	3	2
	eh	2			2	
	éh	29	3	8	11	7
	ih	1	1			
	tsi	5	1	1	3	
	tsu	2	2			
Frequência		97	30	24	27	16
Valor		-0,026	0,835	-0,224	1,039	-1,677

Pelo agrupamento desses itens com o cálculo do desvio reduzido (valor) do bloco, observa-se que esses marcadores conversacionais de hesitação ocupam a zona básica no *corpus* e nas variáveis – vocabulário fundamentalmente básico no *corpus*, em T1 e em T2; vocabulário básico com tendência positiva em T3 e vocabulário básico com tendência negativa em T4.

Os truncamentos de palavras marcam, também, procedimentos de hesitação do falante – exem/ dé/ sta/... –, bem como a gaguez, indicada, na transcrição, por (...). Um

exame completo da hesitação no contexto interacional deve, pois, incluir os diversos elementos que marcam a hesitação do falante.

3.2.6 Extração de sequências textuais

A extração de sequências textuais (ou seja, de recortes discursivos, ou concordâncias, ou listas de palavras em contexto) – ferramenta do *Stablex* – faz-se por recurso aos textos, em função de finalidades de estudo. O recurso aos textos para a extração de sequências permite o estudo mais preciso da significação discursiva de itens ou conjunto de itens, bem como a confirmação dos itens que devem fazer parte do conjunto. A ferramenta *Concord* do programa *WordSmith* apresenta-se igualmente valiosa para a criação de concordâncias, com a vantagem de exibir listas de padrões de colocados, de fundamental importância para a determinação da norma de uso dos itens.

Na seção anterior, destacamos o lugar ocupado pelo item a gente (=nós). Observemos, agora, a tabela de concordâncias do item lexical a-gente, produzida pelo programa *WordSmith* (tabela 13) e recortes discursivos, extraídos pelo programa *Stablex*, para verificar as concordâncias com esse item – 3ª pessoa do singular ou 1ª pessoa do plural −, bem como o emprego fático e conativo que os informantes fazem dele: usado numa tarefa interacional, em que o locutor está preocupado em fazer-se compreender e em despertar a atenção do seu interlocutor, visando ao seu envolvimento.

Tabela 13 – Concordâncias do item lexical *a-gente*

Ordem	*Concordância*	*Palavra*	*Arquivo*	*%*
1	a reação deles e, éh, a-gente percebe que, h	2.894	c:\ws_spsc\corpor~1\2_spfsci.txt	97
2	ra, porque— não sei— a-gente já fica aqui o	924	c:\ws_spsc\corpor~1\1_spfscf.txt	30
3	toda experiência, se a-gente souber olhar o.	2.804	c:\ws_spsc\corpor~1\2_spfsci.txt	94
4	-graduação, realmente, a-gente não vive um	84	c:\ws_spsc\corpor~1\3_spmscf.txt	3
5	uco de sociologia...— a-gente troca, né?—; a	199	c:\ws_spsc\corpor~1\1_spfscf.txt	6
6	lusive, pensam mal de a-gente lá, né?; aqui...	265	c:\ws_spsc\corpor~1\2_spfsci.txt	9
7	brincadeira e, depois, a-gente entra na profu	2.787	c:\ws_spsc\corpor~1\1_spfscf.txt	89
8	a de coração, quando a-gente chega, encontra	2.538	c:\ws_spsc\corpor~1\2_spfsci.txt	85
9	ser gravado, não; mas a-gente já viveu muitas	2.425	c:\ws_spsc\corpor~1\2_spfsci.txt	82
10	mas teve época que a-gente podia dizer q	396	c:\ws_spsc\corpor~1\3_spmscf.txt	14
11	qui é... é interessante: a-gente ficou com vont	237	c:\ws_spsc\corpor~1\2_spfsci.txt	8
12	a; de vez em quando, a-gente vai, para não	1.859	c:\ws_spsc\corpor~1\1_spfscf.txt	59
13	io, mal... mal no café a-gente se encontra; e	1.392	c:\ws_spsc\corpor~1\4_spmsci.txt	49
14	i dar; é que, em casa, a-gente sabe... bo	1.051	c:\ws_spsc\corpor~1\4_spmsci.txt	37
15	a, origens de coisas— a-gente gosta; meu irm	239	c:\ws_spsc\corpor~1\1_spfscf.txt	8
16	professor, quer dizer, a-gente sempre tem q	115	c:\ws_spsc\corpor~1\3_spmscf.txt	4
17	fora os cursos que a-gente faz— especializ	856	c:\ws_spsc\corpor~1\1_spfscf.txt	28

18	?—; às vezes, à noite, a-gente vai, assim, lug	892	c:\ws_spsc\corpor~1\3_spmscf.txt	31
19	a luta tremenda, para a-gente conseguir... con	1.630	c:\ws_spsc\corpor~1\2_spfsci.txt	54
20	u quero dizer é isso: a-gente sai... éh se pa	2.774	c:\ws_spsc\corpor~1\2_spfsci.txt	93
21	eu não fiz. às vezes, a-gente tem até um p	2.593	c:\ws_spsc\corpor~1\2_spfsci.txt	87
22	cadeira— éh... então, a-gente vem na faculd	100	c:\ws_spsc\corpor~1\3_spmscf.txt	4
23	e textos, mesmo que a-gente não soubesse	1.763	c:\ws_spsc\corpor~1\2_spfsci.txt	59
24	o... pô/ outras vezes, a-gente dá um balanço	2.615	c:\ws_spsc\corpor~1\2_spfsci.txt	88
25	não, podia ser, porque a-gente também não s	2.665	c:\ws_spsc\corpor~1\2_spfsci.txt	89
26	xiste tanto filme, para a-gente ver; tenho vist	1.679	c:\ws_spsc\corpor~1\3_spmscf.txt	60
27	remo, não-é? é, no fim, a-gente ficava gostando	477	c:\ws_spsc\corpor~1\2_spfsci.txt	16
28	tidas e desastre, né?, a-gente chega em ca	813	c:\ws_spsc\corpor~1\3_spmscf.txt	28
29	r exemplo, atualmente, a-gente tem que dizer	369	c:\ws_spsc\corpor~1\3_spmscf.txt	13
30	o canal dois é difícil a-gente ver. ah o	1.856	c:\ws_spsc\corpor~1\3_spmscf.txt	67
31	isa que fizeram com a-gente, lá na Ford, el	581	c:\ws_spsc\corpor~1\4_spmsci.txt	20
32	le no Chefão— então, a-gente diz para o	385	c:\ws_spsc\corpor~1\3_spmscf.txt	14
33	trair simplesmente, se a-gente não começar	1.890	c:\ws_spsc\corpor~1\1_spfscf.txt	60
34	administração; então, a-gente aproveita um p	171	c:\ws_spsc\corpor~1\1_spfscf.txt	5
35	assim que desagrada, a-gente desliga. ol	1.916	c:\ws_spsc\corpor~1\1_spfscf.txt	61
36	e ê/ de início, quando a-gente pega esse tip	2.772	c:\ws_spsc\corpor~1\1_spfscf.txt	88
37	ma coisa dificílima de a-gente... uai! parec	1.943	c:\ws_spsc\corpor~1\2_spfsci.txt	65
38	exame vestibular que a-gente faz dá direito	832	c:\ws_spsc\corpor~1\4_spmsci.txt	30
39	inda. não, porque a-gente começa achar	2.714	c:\ws_spsc\corpor~1\4_spmsci.txt	97
40	, sei lá, as coisa que a-gente tem que fazer;	2.279	c:\ws_spsc\corpor~1\1_spfscf.txt	72
41	i lá se é maravilhosa; a-gente vive, não-é?; qu	2.304	c:\ws_spsc\corpor~1\2_spfsci.txt	77
42	éh... seria imprudente a-gente estar gravando	2.714	c:\ws_spsc\corpor~1\2_spfsci.txt	91
43	ema, ou?... sei lá, a-gente nunca tem u	352	c:\ws_spsc\corpor~1\3_spmscf.txt	13
44	bancário, assim, extra, a-gente acaba gastan	749	c:\ws_spsc\corpor~1\3_spmscf.txt	26

Recortes Discursivos:

T1:

_____*' a-gente'* _____

então, ***a-gente*** *aproveita um pouquinho das experiências um dos outros, né?*
a-gente *troca, né?*
a-gente *gosta*
fora os cursos que ***a-gente*** *faz*
a-gente *já fica aqui oito horas lendo*
a-gente *vai*
e ***a-gente*** *não começar entrar no mérito do... do... do nível cultural*
a-gente *desliga*
as coisa que ***a-gente*** *tem que fazer*
quando ***a-gente*** *pega esse tipo de trabalho*
a-gente *entra na profundidade do problema*
a-gente *percebe que, hoje, a mocidade é tão franca, tão aberta*
a-gente *ficou com vontade de voltar, não-é?*
inclusive, pensam mal de ***a-gente*** *lá, né?*
a-gente *ficava gostando do Brasil, por causa disso*
para ***a-gente*** *conseguir...*
mesmo que ***a-gente*** *não soubesse nada do...*

T2:
_____' a-gente' _____

*está uma luta tremenda, para **a-gente** conseguir...*
***a-gente** vive, não-é?*
*quando **a-gente** chega*
*às vezes, **a-gente** tem até um pouquinho de senso de culpa*
*porque **a-gente** também não sabe*
*seria imprudente **a-gente** estar gravando uma porção de... de detalhes*
***a-gente** sai...*
*se **a-gente** souber olhar o... o lado bom dela*

T3:
_____' a-gente' _____

*porque pós-graduação, realmente, **a-gente** não vive um ambiente, assim, universitário*
*éh... então, **a-gente** vem na faculdade, uma vez por semana*
*quer dizer, **a-gente** sempre tem que manter uma distância*
*sei lá, **a-gente** nunca tem um artista predileto*
*então, por exemplo, atualmente, **a-gente** tem que dizer que é o Marlon-Brando*
*então, **a-gente** diz para o Marlon-Brando*
*mas teve época que **a-gente** podia dizer que o Dustin-Hoffman*
*então, o dinheiro que eu tenho, assim, um... um saldo bancário, assim, extra, **a-gente** acaba gastando lá, porque é um pouco caro*
***a-gente** chega em casa chateado*
*às vezes, à noite, **a-gente** vai, assim, lugar aí que tem samba*
*não existe tanto filme, para **a-gente** ver*
*mas mesmo o canal dois é difícil **a-gente** ver*

T4:
_____' a-gente' _____

*e aquela outra pesquisa que fizeram com **a-gente**, lá na Ford*
*porque o exame vestibular que **a-gente** faz dá direito a chegar ao mestre de administração-de-empresa*
*é que, em casa, **a-gente** sabe...*
*mal no café **a-gente** se encontra*
*não, porque **a-gente** começa achar que tudo da mocidade era melhor, né?*

Nota-se que a concordância, no *corpus* em estudo, sempre se faz em terceira pessoa do singular, o que é esperado no discurso de falantes cultos. A pesquisa exaustiva a partir de informantes de outros níveis de escolaridade e socioeconômicos muito provavelmente exibirá concordância em primeira pessoa do plural.

O resultado da extração pode se transformar numa antologia, dicionário ou gramática de algum aspecto especial.

Seguem recortes discursivos dos itens *não-é?* e *né?*, extraídos do texto transcrito – variável T1 –, os quais exibem, também, marcadores conversacionais, truncamentos de palavras e gaguez:

Ah item difícil... —falar o quê, ***né?****— que que você quer que eu diga, assim?* ***Ahn****... bom, em casa... deixe eu ver... eu moro com meus pais que têm idade,* ***não é?****; é uma vida bastante pacata, bastante,* ***ahn****, metódica, vamos dizer assim;* ***ahn****... depois, a minha família tem dois irmãos casados, tenho sobrinhos; então, aquela vidinha, assim,* ***ahn****, muito****(...)*** *concentrada na família,* ***né?****; depois, tenho numerosas amigas que eu gosto muito, aliás;* ***ahn****... quer dizer, tenho as diversões, assim, comum, de cinema, teatro, bate-papo, essas coisa...* ***ahn****... no ambiente de... de escola, propriamente... bom, eu gosto muito do meu trabalho, gosto mesmo muito, e... e... e, depois de uns anos para cá, ando fazendo pesquisa****(...)*** *—francamente, me diverte, entende?; quer dizer, mais do que só gostar, me apaixona— e gosto do ambiente de trabalho; em geral, tenho colegas muito agradáveis; ultimamente, tivemos uma série de experiências muito positivas que nós chamaremos muito reforçadoras; estou cra/ muito animada; principalmente, agora, estou muito animada no meu trabalho. Não sei o que que você quer mais saber.*

Assim, eu... eu acho... eu... o indivíduo, quando escolhe a profissão por... por escolha, independente de influência de qualquer ***indivi/*** *qualquer pessoa, tem,* ***ahn****..., muito mais possibilidade de realizar se dentro do campo que escolheu. Por* ***exem/*** *eu acho que* ***dé/*** *dentro do* ***sta/*** *do status atual, biblioteconomia é um campo altamente explorável, com boa remuneração econômica e com grande,* ***ah****, possibilidade de atividades e especializações; é um campo novo, com poucos especialistas,* ***ih****, dentro da o... dentro de São-Paulo; quase todos eles são englobados pela Universidade-de-São-Paulo. acho assim, por exemplo, no momento, nós contamos, aqui na faculdade, com oito bibliotecários, todos de curso superior, dos quais quatro tem especialização em ciências biomédicas —inclusive eu tenho especialização—.* ***Ah*** *o ambiente de trabalho é ideal; não sei, porque não conheço,* ***uhn****...,* ***éh****..., nenhum; trabalhei um... durante dois anos como bibliotecária da...* ***tsi****... do Conselho-Regional-de-Contabilidade-do-Estado, mas era uma biblioteca independente, com um único profissional; então, você não pode avaliar bem a... o relacionamento; isso eu vim sentir mais aqui na universidade; eu acho um campo... por ser um campo muito novo, todo profissional é muito unido; acho ideal, um trabalho muito bom, e nós trabalhamos aqui, em equipe; embora nós tenhamos todas nós setores bem definidos, pela carga de trabalho ser muito grande, nós todas trabalhamos em comum acôrdo, a ponto de podermos qualquer uma substituir a outra, a qualquer momento. É, eu acho que deu, assim, uma amplitude de trabalho muito grande, o que, muitas vezes, não se verifica em outros campos,* ***né?****; nós, graças-a-Deus, não tivemos esse problema.*

Ah *isso é difícil de responder,* ***né? Ahn****... basicamente, veja bem, eu... eu... eu estudo, leio e trabalho; meu horário, não muito regular, porque as aulas variam;* ***ahn*** *quer dizer, às vezes, tenho aula... começa às duas;* ***ahn ahn ahn*** *às vezes, é de manhã; quer dizer... e variam muito o horário, mas, basicamente, a minha vida é... é... é estudar e... e dar aula,* ***né?***

CONSIDERAÇÕES FINAIS

Os estudos efetuados demonstram que programas especiais de análise linguística representam uma interface útil entre o homem e a máquina, especialmente em se tratando de sistemas desenvolvidos especialmente para aplicações linguísticas, como é o caso do *WordSmith* e do *Stablex*, que fazem mais do que simplesmente acelerar e facilitar o trabalho do pesquisador: fornecem, de forma confiável e segura, porque embasada por métodos e critérios científicos, indicadores para uma análise da estrutura temática e articuladora do discurso, algumas vezes, reforçando, outras orientando hipóteses prévias.

REFERÊNCIAS BIBLIOGRÁFICAS

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi (1998). *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Camlong, André (1996). *Méthode d'analyse lexicale textuelle et discursive*. Paris: C.R.I.C. & OPHRYS,.

Camlong, André (2004). *Stable version PC.* São Paulo: Pirus Tecnologia.

International Phonetic Association (1999). *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.

Scott, Mike (2007). *WordSmith Tools* version 5. Oxford: Oxford University Press.

Szyperski, C. (1998). *Component Software: Beyond Object-Oriented Programming*. Boston: Addison-Wesley.

Zapparoli Castro Melo, Zilda Maria (1980). *Análise do comportamento fonológico da juntura intervocabular no português do Brasil (variante paulista). Uma pesquisa linguística com tratamento computacional*. São Paulo. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística do Departamento de Linguística da Universidade de São Paulo.

Zapparoli, Zilda Maria (2006). Análise lexical, textual e discursiva: uma abordagem quantiqualitativa. In: I CONGRESO INTERNACIONAL, I., Universidade de Navarra, Pamplona, 2002. *Actas – I.* Pamplona: Arco / Libros,. pp. 835-849.

Zapparoli, Zilda Maria (1997). Considerações sobre a utilização de novas tecnologias na análise do léxico do português falado culto de São Paulo. Preti, Dino, org., *O discurso oral culto.* São Paulo: Humanitas Publicações - FFLCH/USP. pp.151-173. (Projetos Paralelos, v.2).

Zapparoli, Zilda Maria (2009). *Sistema CorPor - versão desktop*. Disponível em: <http://www.corpor.fflch.usp.br>.

Zapparoli, Zilda Maria (2002). Um pouco da história da análise informatizada do léxico no Brasil.Nunes, José Horta; Petter, Margarida, orgs., *História do Saber Lexical e Constituição de um Léxico Brasileiro.* São Paulo / Campinas: Humanitas / Pontes, pp. 223-253.

Zapparoli, Zilda Maria; Camlong, André (2002). *Do Léxico ao Discurso pela Informática*. São Paulo: EDUSP/FAPESP, 256 p. + CD-ROM.