A FACE QUANTITATIVA DA LINGUAGEM: UM DICIONÁRIO DE FREQÜÊNCIAS DO PORTUGUÊS

Maria Tereza Camargo BIDERMAN[1]

■ RESUMO: O *Dicionário de Freqüências* do léxico do português brasileiro contemporâneo baseou-se num *corpus* de língua escrita, variedade brasileira de 5 milhões de palavras (1950 a 1990). Alguns resultados quantitativos: apenas 42.212 unidades léxicas diferentes totalizaram os 5 milhões de ocorrências do *corpus*, excluídos topônimos e antropônimos. Os dados estatísticos do dicionário registram altíssima freqüência das palavras instrumentais (artigos, preposições, pronomes, conjunções etc.) bem como de verbos auxiliares e modalizadores. O mesmo ocorre com palavras de significação muito geral, arquilexemas, altamente polissêmicos. Na vertente oposta estão as palavras de baixa freqüência sobretudo os *hapax legomena*, que contribuem maciçamente para o total de 42.212 lexias registradas neste *corpus*. De fato, as palavras de baixa freqüência totalizam grande parte desse *index verborum*; caso contrário, o repertório vocabular seria muito menor. A categoria *substantivo* contribui com a maioria de vocábulos que ocorreram apenas uma vez no *corpus*, assim como os tecnicismos da linguagem científica. O vocabulário jornalístico é o mais neutro e o menos temático, constituindo uma espécie de média entre os outros gêneros de linguagem.

■ PALAVRAS-CHAVE: Lexicoestatística; dicionário de freqüências; hapax legomena; index verborum; vocabulário multiuso.

1 Professora aposentada do Departamento de Lingüística – Faculdade de Ciências e Letras – UNESP-14800-901 – Araraquara – SP – Brasil – mtbider@ibm.net.

1 Muitas teorias foram elaboradas para tratar o fenômeno da linguagem. Uma delas, a Estatística Lingüística, considera a face quantitativa da linguagem. Não resta dúvida que a elevadíssima freqüência dos fenômenos lingüísticos justifica tal abordagem. Essa ciência interdisciplinar produziu resultados eficazes ao analisar dados das mais diversas línguas, chegando a algumas conclusões expressivas de natureza universal sobretudo nas décadas de 1960 e 1970. Constatou-se assim a estabilidade dos símbolos lingüísticos – letras, fonemas, palavras, categorias gramaticais manifestam uma recorrência tão regular que tornam possível a sua previsibilidade. Na verdade a freqüência de letras, fonemas, número de sílabas e comprimento da palavra são independentes do estilo individual e constituem um condicionamento lingüístico. Assim é possível prever os fonemas, grafemas, vocábulos e unidades gramaticais que poderão ocorrer nos discursos oral e escrito dos falantes e escritores.

Com base nos resultados da Estatística Léxica ou Lexicoestatística podemos afirmar também que a freqüência é uma característica típica da palavra. Aliás, a norma lingüística se baseia na freqüência dos usos lingüísticos. Assim, a norma lexical nada mais é que a média dos usos freqüentes das palavras que são aceitas pela comunidade dos falantes.

E não só isso. Também as mudanças lingüísticas que, no decorrer da história, levam de um estado de língua a outro, advêm da freqüência de certos usos em detrimento de outros. Quando estudamos a evolução do latim até as línguas românicas verificamos que mudanças estruturais que acarretaram o surgimento de novos sistemas lingüísticos (as línguas latinas) se deram em virtude da freqüência com que certos fenômenos lingüísticos ocorriam. Um exemplo típico é o da derrocada do sistema de paradigmas nominais do latim clássico. No latim havia cinco modelos de declinação de substantivos que, no período românico e nas modernas línguas românicas, desapareceram totalmente, em virtude da recorrência da obliteração das desinências características de cada caso em cada um dos paradigmas de declinação.

Na aplicação das teorias estatísticas ao estudo da linguagem, muitos parâmetros da Estatística Clássica não se mostraram adequados, em virtude do grande número de graus de liberdade existentes nos sistemas lingüísticos. Todavia, na análise comparativa de línguas, esse modelo teórico revelou-se útil para o estudo das genealogias lingüísticas. Maurice Swadesh criou a Glotocronologia a fim de tentar descobrir a época de separação de línguas oriundas de uma mesma

família lingüística e não documentadas. É sabido que as línguas indígenas da América, entre muitas outras línguas pouco conhecidas do globo, constituem um enorme desafio para os lingüistas no que concerne sua origem, bem como sua filiação lingüística. Para tentar solver esse mistério, Swadesh criou a teoria glotocronológica, que aplica um tratamento estatístico ao léxico das línguas comparadas. Usando como referência uma lista de 100 palavras básicas para qualquer língua e que deve existir em qualquer cultura, Swadesh confrontava os vocábulos equivalentes nas línguas cuja filiação estava buscando. Através de uma fórmula elaborada por ele, Swadesh pode descobrir o tempo de fragmentação e separação de algumas línguas indígenas da América.

Além disso, notou-se uma certa constância na distribuição do vocabulário em qualquer tipo de texto para várias línguas, a saber: francês, espanhol, italiano, romeno e português. Em todas essas línguas as altas freqüências são habitadas pelos vocábulos gramaticais e por um número reduzido de palavras lexicais de significado muito geral. Refiro-me aos estudos de Estatística Léxica aplicados às cinco principais línguas românicas, realizados em Stanford University por Alphonse Juilland e uma equipe de pesquisadores que elaboraram os "dicionários de freqüência" do espanhol, francês, romeno, italiano e português. Essas pesquisas feitas com grande rigor evidenciaram que o topo das listas de freqüência é constituído por palavras gramaticais ou itens vocabulares de grande instrumentalidade semântico-gramatical. O único senão é relativo ao tamanho dos *corpora* que foram então analisados: para cada uma das línguas o *corpus* totalizava 500 mil palavras embora fosse constituído com uma grande heterogeneidade. Para a época em que esses estudos foram feitos (década de 1960 e início dos anos 70) não se podia fazer mais, pois os computadores desse tempo eram ainda muito ineficientes e sua capacidade de armazenamento e de tratamento de grandes volumes de dados ainda deixava a desejar. E não se esqueça que estamos falando dos Estados Unidos e do mais importante centro de computação desse país. Estávamos na época dos grandes "mainframes". O computador pessoal ainda não existia e menos ainda os pentium e toda a magnífica parafernália de hardware e de software dos anos 80 e sobretudo 90. O aparecimento dessa tecnologia tornou possível a criação de gigantescas bases de dados lingüísticos para tratamento computacional, estatístico, lexicográfico e muitos outros mais. E assim chegamos a uma era em que já é possível analisar de modo mais pertinente o léxico de uma língua, a despeito das imensas dificuldades postas pelo gigantesco volume de dados que constitui o vocabu-

lário de qualquer língua de civilização, bem como a enorme complexidade desse sistema de signos lingüísticos.

Como se sabe o léxico constitui um sistema aberto de demarcação praticamente impossível. De fato, o crescimento do léxico faz-se numa progressão geométrica, em virtude da criação contínua de palavras novas. O processo das mudanças sociais é um moto-contínuo: a criação humana é incessante e a invenção de novos referentes nunca cessa. Portanto, está posto um desafio enorme para a nossa capacidade analítica: qual será a extensão do léxico de uma língua? Rey-Debove (1970, p.4) levantou a hipótese de que o léxico de uma língua de civilização como o francês ou o inglês ultrapasse 200 mil unidades podendo atingir 500 mil palavras, caso se registrem todos os vocabulários das línguas de especialidades, ou seja, as terminologias técnico-científicas. Como o léxico categoriza o conhecimento humano na forma de palavras a possibilidade de sua ampliação é praticamente infinita.

Não resta dúvida que é extremamente difícil dar conta de todo o acervo de palavras da língua portuguesa para um tratamento lexicoestatístico e computacional dentro dos parâmetros desejáveis. De longa data, desde os anos 70 ao colaborar com Juilland em Stanford University, almejava fazer uma grande pesquisa lexicoestatística sobre o léxico do português brasileiro contemporâneo com o fito de obter resultados que pudessem ser referência básica para o ensino da língua portuguesa, assim como para a elaboração de dicionários mais bem concebidos para a comunidade dos falantes do Brasil. Quando da criação da base de dados textuais em forma digital na Faculdade de Ciências e Letras da UNESP de Araraquara sob o comando do Prof. F. da Silva Borba, meu antigo desideratum tornou-se possível, sobretudo depois que o CNPq passou a apoiar minhas pesquisas. Ainda não cheguei aos resultados definitivos pelos quais almejava, porém, julgo que as pesquisas que passarei a discutir e comentar a seguir já constituem resultados altamente relevantes sobre a estrutura quantitativa do léxico do português brasileiro contemporâneo. Vou discutir resultados de uma primeira versão desta pesquisa por já possuir todos os efetivos prontos. Lembro que o objetivo destas pesquisas era, antes de mais nada, elaborar um *dicionário de freqüências* do português brasileiro contemporâneo.

Na segunda versão do *dicionário* reorganizei o *corpus* de língua escrita e acrescentei um *subcorpus* de língua falada totalizando 6 milhões de ocorrências. Eliminei alguns textos idiossincráticos como, por exemplo, obras de Guimarães Rosa, ou então no domínio técnico-científico, obras muito pontuais. Também resolvi reunir num só *subcorpus*

romances e contos e peças dramáticas para formar o subcorpus LL: linguagem literária. No que tange a linguagem jornalística (LJ) eliminei uns tantos textos e acrescentei a esse *subcorpus* um volume considerável de materiais da revista *Veja* dos anos de 1992, 1993, 1994 e 1995. Tratando-se de um semanário de temática muito abrangente sobre a contemporaneidade, selecionei os textos que seriam incluídos para que eles pudessem representar todos os domínios possíveis. Reformulei também todo o *subcorpus* de LT (linguagem técnico-científica) excluindo algumas obras e acrescentando outras, para que ele fosse mais representativo dessa subárea (linguagem técnico-científica). Nessa versão excluí o *subcorpus* LO (linguagem oratória). Foram os seguintes os projetos que forneceram materiais de língua falada (*subcorpus* LF): 1. os vários projetos NURC: São Paulo, Recife, Salvador, Porto Alegre e Rio de Janeiro; 2. o projeto PEUL do Rio de Janeiro: 3. o projeto de Maceió; 4. o projeto de linguagem rural (bóias-frias) da UNESP de Assis; 5. dados do banco de dados da PUC, São Paulo; 6. descrição da fala de universitários de Curitiba (dados de Berlinck[2]). Entretanto, não vou aqui comentar resultados dessa segunda pesquisa; pretendo deixar apenas consignado que foi feito também o confronto entre o vocabulário da língua escrita e da língua falada.

Vejamos agora alguns dados e resultados da primeira versão do *dicionário de freqüências* que abrange tão-somente a variedade escrita da língua.

2 A base textual por mim utilizada estava constituída por um *corpus* de 5 milhões de palavras do português do Brasil, língua escrita (dados de 1950 a 1995). Eis a composição do *corpus*:

- LR – literatura romanesca (romances e contos): 1.394.855 palavras;
- LD – literatura dramática: 620.386 palavras;
- LT – literatura técnico-científica: 1.223.605 palavras;
- LJ – literatura jornalística: 1.458.174 palavras;
- LO – literatura oratória: 442.172 palavras.

O *subcorpus* LR compõe-se de romances e contos. O *subcorpus* LD compõe-se de peças dramáticas, alguns roteiros de filmes e textos de telenovelas.

---

2 Dados manuscritos de pesquisa realizada por Rosane Berlinck (UNESP-Araraquara) com universitários de Curitiba, PR.

A linguagem jornalística (LJ) compreende qualquer tipo de texto jornalístico (noticiário em geral e editorial) dos principais jornais brasileiros, bem como de semanários de grande circulação no Brasil. Por exemplo: *O Estado de S. Paulo*, *Folha de S. Paulo*, *Jornal do Brasil*, *O Globo*, *Correio Brasiliense*, *Zero Hora*. Algumas revistas de grande circulação que integram esse corpus: *Veja*, *Isto É*, *Visão*, *Exame*, *Placar*. Um volume considerável de materiais da revista *Veja* constam desse acervo.

O *subcorpus* de LT (linguagem técnico-científica) inclui obras variadas sobre os vários domínios do conhecimento humano, a saber: *agronomia, antropologia, arqueologia, artes, astronomia, biologia, clínica médica, culinária, direito, farmacologia, física, geologia, hidrologia, história, lingüística, marcenaria, medicina, mineralogia, patologia, psicologia, psiquiatria, zoologia*, etc. Os textos da LT foram extraídos de manuais de divulgação e obras didáticas destinadas ao público em geral bem como a estudantes universitários. Queria-se garantir a presença do vocabulário dos diversos domínios do conhecimento humano sem saturar esse *subcorpus* com uma terminologia muito especializada que só interessa aos especialistas.

Procedi à lematização das formas da base textual e à tabulação dos dados lexicoestatísticos.

Com base neste universo lexical com que trabalhei, cheguei a algumas conclusões sobre o léxico do idioma português, que serão apresentadas a seguir. Creio ser possível fazer algumas projeções sobre a estrutura geral do léxico português contemporâneo.

3 Considere-se primeiramente as palavras de *altíssima freqüência*.

3.1 Cerca de 42% do total das ocorrências do corpus é constituído por pouco mais de mil palavras, as mais freqüentes da língua (ver Gráfico 1).

Este gráfico atesta o seguinte: 80% de qualquer texto do português é constituído por estas mil palavras, que são reiteradas continuamente. Esse resultado confirma as conclusões da pesquisa realizada por Duncan (1972) sobre o português (*A Frequency Dictionary of Portuguese Words*). É verdade que Duncan utilizou um *corpus* relativamente pequeno (500 mil palavras) como já disse. Suas conclusões, porém, são idênticas às minhas. As mil palavras mais freqüentes do seu *corpus* (de arquitetura semelhante ao do meu *corpus*) constituíam 84,57% do total. Ora, as pesquisas lexicoestatísticas feitas paralelamente sobre

as outras línguas românicas por Juilland e a equipe de pesquisadores de Stanford University forneceram resultados quase idênticos. No espanhol, no francês, no italiano e no romeno, as mil palavras mais freqüentes compunham mais de 80% de qualquer texto escrito em cada uma dessas línguas. Julgo que podemos extrapolar estes resultados e formular a hipótese de que o mesmo deve ocorrer em qualquer língua.

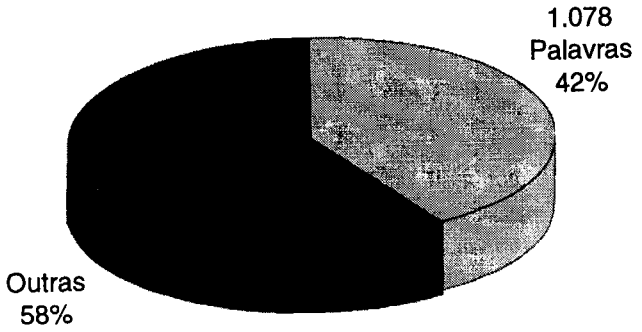

GRÁFICO 1 - Distribuição das palavras com freqüência maior ou igual a 500.

No português brasileiro contemporâneo, que palavras são estas? São elas: todas as ***palavras instrumentais*** como *artigos, pronomes, preposições, conjunções, advérbios, numerais*, e algumas palavras lexicais ou plenas das classes *substantivo, adjetivo* e *verbo*. Eis um exemplário das palavras com freqüência maior ou igual a 500, aquelas que integram o referido total de 1.078, ou seja, as palavras de altíssima freqüência no português brasileiro:

artigos: definidos: *o, as, os, as* [total: 383.116]; indefinidos: *um, uma, uns, umas* [total: 98.797];

pronomes: *ele, eles* [total: 18.964], *ela, elas* [total: 9.666]; *eu; você; nada, ninguém; que; qual, qualquer, tal, tudo; este(s), esses(s), esta(s), essas(s), meu(s), minha(s). etc.;*

preposições: *de* [total: 180.228], *em* [total: 55.794], *para* [total: 50.848], *por* [total: 32.241];

contrações de preposições: *do(s), da(s), no(s), na(s);* locuções prepositivas: *acima de, abaixo de, antes de, atrás de, depois de, embaixo de, em cima de, ao lado de, em vez de;*

advérbios: *agora, já, ainda, depois, depressa, cedo, hoje, ontem, muito, pouco, bastante, mais, quase, mal, nunca, sempre, não, logo, também, lá etc.;*

conjunções: *e* [total: 55.794], *em, mas, como, que* [total incluindo homônimos: 137.617];

locuções conjuncionais: *depois que, logo que, para que* etc.

3.2 Exemplos de *palavras plenas*:

substantivos: *ação, água, amor, área, alma, arte, ato, banco, base, cabeça, cabelo, carro, cidade, coisa, começo, cor, domingo, dor, dúvida, economia, espírito, família, forma, história, hora, homem, mãe, maneira, mão, mês, mulher, pai, palavra, pessoa, qualidade, rapaz, realidade, rio, rua, sala, sangue, tempo, terra, uso, vez, vida, voz.*

adjetivos: *alto, baixo, bom, bonito, difícil, duro, fácil, geral, humano, largo, maior, mau, novo, primeiro, santo, são, social, velho.*

verbo: *acabar, bastar, começar, comer, dar, descer, dormir, entrar, estar, falar, fazer, ficar, ir, passar, poder, querer, receber, responder, saber, ser, ter, tirar, trazer, ver, vir, viver.*

Estas palavras constituem o núcleo do vocabulário do português e podem ser consideradas como essenciais para a comunicação neste idioma. Na segunda pesquisa com um *corpus* de 6 milhões de ocorrências que incluía um *subcorpus* de língua falada, ficou também evidente que esse repertório básico de palavras se comportava da mesma maneira do ponto de vista estatístico.

4 Vejamos uma amostra de como os dados vocabulares estão lematizados no *dicionário de freqüências* e os respectivos resultados quantitativos. No processo de lematização eliminei os nomes próprios, antropônimos e topônimos. De fato, os nomes próprios identificam um referente único com identidade própria distinta de todos os demais. Aliás, na afasia e em outras moléstias com perda de memória por esclerose ou por envelhecimento, é típico o esquecimento e até a perda completa do registro desses nomes. O nome próprio usado na apelação vocativa do interlocutor e na sua interpelação no diálogo não integra o vocabulário referencial e instrumental lingüístico que estou quantificando. Aliás, a nomeação, o "batismo" de uma pessoa tem a função de integrar o indivíduo na sociedade. E a designação de lugares tem a função de estabelecer marcos de localização espacial para os sujeitos do discurso. Em suma, o nome próprio nos remete ao campo dêitico e nos envia às três dimensões da dêixis: a pessoa, o espaço e o tempo (Molino et al., 1982, p.19). Seguindo a teoria de Molino e de Bühler (1961), podemos afirmar que o campo dêitico se distingue do campo

referencial (ou da representação), aquele onde se enquadra o léxico que estamos considerando e quantificando.

4.1 Exame ilustrativo de algumas *palavras plenas* de alta freqüência.

As palavras gramaticais e certo grupo de verbos são as classes mais estáveis da língua. São palavras multiuso que aparecem em qualquer texto, independentemente de seu conteúdo temático. Donde a distribuição uniforme dessas palavras entre os vários gêneros escrutinados. De fato, pode-se constatar que elas são igualmente freqüentes em todos os cinco tipos de literatura analisados, seja qual for o gênero ou o tema tratado. O mesmo pode ser dito, com algumas ressalvas, de um grupo de adjetivos e substantivos de significado muito geral.

- exemplo de substantivos: **mulher** e **tempo**

No paradigma dos substantivos coloquei sob o lema de entrada todas as flexões de número e grau bem como variantes populares.

	LR	LD	LT	LJ	LO	TOTAL
mulher	sub					4.882
MULHER	1.467	983	186	860	118	3.614
MULHERES	438	165	195	331	50	1.179
MULHÉ (pop)	0	20	0	0	0	20
MUIÉ (pop)	8	3	2	0	0	13
MULHERAÇO	0	0	1	0	0	1
MULHERÃO	1	0	0	0	0	1
MULHERINHA	1	0	0	0	0	1
MULHERÕES	0	1	0	0	0	1
MULHERZINHA	16	2	0	4	0	22
MULHERZINHAS	0	0	0	1	0	1
MULHER-DA-VIDA	1	0	0	0	0	1
MULHER-DAMA	11	4	0	0	0	15
MULHER-ESPOSA	1	0	0	0	0	1
MULHER-FÊMEA	1	0	0	0	0	1
MULHER-HOMEM	1	0	0	0	0	1
MULHER-MACHO	1	0	0	0	0	1
MULHER-SOLDADO	0	0	0	1	0	1
MULHERES-DAMA	1	0	0	0	0	1
MULHERES-DAMAS	6	0	0	0	0	6
MULHERES-MARAVILHAS	0	0	0	1	0	1

		LR	LD	LT	LJ	LO	TOTAL
tempo	sub						6.670
TEMPO		1.913	716	1.199	1.507	481	5.816
TEMPOS		227	28	201	277	91	824
TEMPÃO		12	5	0	1	0	18
TEMPINHO		9	0	0	2	0	11
TEMPINHOS		0	0	0	1	0	1

- exemplo de adjetivos: **grande** e **santo**

Quando se trata de adjetivos que podem ser também substantivos (caso de homonímia) as formas foram incluídas no paradigma e quantificadas conjuntamente; contudo, registrou-se a ambigüidade das duas categorias possíveis. No primeiro exemplo **grande**, esse vocábulo quase sempre funciona como adjetivo; no grau superlativo e mesmo diminutivo é sempre adjetivo. Quanto a **santo**, a freqüência do substantivo é quase tão grande quanto a do adjetivo.

		LR	LD	LT	LJ	LO	TOTAL
grande	adj./sub.						7.343
GRANDE		1.118	392	1.484	1.795	695	5.484
GRANDES		291	63	643	589	244	1.830
GRANDALHÃO	ho	7	0	1	1	0	9
GRANDALHONA	ho	0	0	0	1	0	1
GRANDÃO	ho	1	0	0	1	0	2
GRANDESSÍSSIMO		2	0	0	0	0	2
GRANDINHA		1	0	0	0	0	1
GRANDINHO		1	0	0	0	0	1
GRANDÍSSIMA		1	0	0	1	0	2
GRANDÍSSIMO		0	0	1	0	0	1
GRANDÍSSIMOS		1	0	0	0	0	1
GRADISSÍSSIMA		3	0	0	0	0	3
GRANDISSÍSSIMAS		0	1	0	0	0	1
GRANDISSÍSSIMO		1	0	0	0	0	1
GRANDÕES	ho	0	0	2	0	0	2
GRANDONA	ho	0	0	0	1	0	1
GRANDONAS	ho	1	0	0	0	0	1

		LR	LD	LT	LJ	LO	TOTAL
santo	adj./sub.						2.520
SANTA	ho	149	378	102	245	79	953
SANTAS	ho	1	0	1	7	6	15
SANTO	ho	323	135	148	148	79	833

SANTOS	ho	162	43	106	268	69	648
SANTARRÕES	ho	1	0	0	0	0	1
SANTINHA	ho	3	10	0	0	0	13
SANTINHO	ho	9	0	2	0	0	11
SANTINHOS	ho	3	0	0	3	0	6
SANTÍSSIMA		4	6	7	2	13	32
SANTÍSSIMO	ho	3	0	3	0	2	8

Observe-se que substantivos e adjetivos de alta freqüência têm uma distribuição mais ou menos homogênea, proporcional ao tamanho do *subcorpus* de que procedem. Sinal evidente de quanto essas palavras independem do gênero, do estilo e do conteúdo temático.

4.2 Os verbos têm um estatuto diferenciado no *corpus*. Em 1974 Müller analisou detidamente o comportamento dos vinte verbos mais freqüentes do francês em três *corpora* bem diferentes da língua francesa. Eis os *corpora* confrontados: 1. dados do *Francês Fundamental* de língua falada da década de 1950, totalizando 312.135 ocorrências; 2. dados de língua escrita do *Frequency Dictionary of French Words* (1920-1940) totalizando 500 mil palavras; 3. dados do *Trésor de la Langue Française*, que totalizavam na época da pesquisa 71 milhões de palavras e registravam textos de 1789 a 1964. Müller constatou que num discurso qualquer em francês, de cada cinco palavras, uma é verbo. Concluiu também que esses vinte verbos mais freqüentes do francês fornecem de 1/3 a 2/3 das formas verbais de um texto qualquer com a seguinte distribuição: 1/3 em textos literários mais elaborados e 2/3 na língua oral mais espontânea. Além disso, esses vinte verbos mais freqüentes situam-se na escala decrescente de freqüência em posições quase idênticas; isso confirma também que distribucionalmente eles operam de maneira muito similar na língua, não importando o tipo de variáveis lingüísticas consideradas, a saber: língua falada ou escrita, linguagem literária, técnico-científica, jornalística etc. E mais: note-se que esse três *corpora* cobrem períodos diacrônicos distintos da língua francesa. Os resultados demonstram, portanto, que o comportamento lingüístico desses verbos tem-se mantido quase imutável ao longo de duzentos anos. São, pois, verbos muito estáveis no idioma. Embora eu não tenha feito ainda o mesmo tipo de análise para os dados lexicoes-tatísticos dos vinte verbos mais freqüentes do português, creio que podemos formular a hipótese de que as conclusões de Müller são válidas também em nossa língua.

Vejamos quais são os vinte verbos mais freqüentes do *corpus* em ordem decrescente de freqüência. Note-se que eles coincidem com os

verbos mais freqüentes na pesquisa feita em Portugal sobre a língua falada e que redundou na elaboração do vocabulário do português fundamental. São eles:

1º	ser:	50.222
2º	ter:	34.586
3º	ir:	28.965
4º	estar:	27.746
5º	poder:	16.593
6º	dizer:	15.445
7º	haver:	15.004
8º	fazer:	14.279
9º	dar:	10.792
10º	ver:	10.391
11º	saber:	10.247
12º	querer:	9.986
13º	ficar:	8.605
14º	achar:	7.980
15º	dever:	7.758
16º	falar:	5.259
17º	chegar:	4.628
18º	precisar:	4.039
19º	começar:	3.596
20º	olhar:	3.383

- exemplo de paradigma verbal com todas as ocorrências flexionais do verbo **querer**.

querer	ver						9.983
QUEIRA		44	27	12	23	6	112
QUEIRA-		1	0	0	0	0	1
QUEIRAM		3	2	6	22	10	43
QUEIRAMOS		0	0	4	1	1	6
QUEIRAS		1	3	0	1	1	6
QUER		895	847	349	533	225	2.849
QUER-SE		0	0	1	0	0	1
QUERE	var	3	4	0	0	0	7
QUERÊ-		1	0	0	0	0	1
QUEREIS		1	0	0	5	8	14
QUEREM		92	127	33	137	29	418
QUEREMO-		0	0	0	0	1	1
QUEREMOS		40	24	34	49	53	200

QUERENDO			230	135	18	73	18	474
QUERENDO-			1	0	0	0	0	1
QUERER	ho	sub	234	98	51	`77	26	486
QUERER-			2	0	0	1	0	3
QUERERÁ			2	2	0	0	1	5
QUERERÃO			3	0	0	2	1	6
QUERERDES			2	0	0	0	0	2
QUEREREM			0	0	1	3	1	5
QUERERIA			11	0	0	3	2	16
QUERERIAM			0	0	0	1	0	1
QUERERÍAMOS			1	0	0	0	0	1
QUERERÍEIS			0	0	0	0	1	1
QUERES			40	122	1	1	4	168
QUERES-			0	2	0	0	0	2
QUERIA			721	267	36	250	21	1.295
QUERIA-			3	0	0	1	0	4
QUERIAM			81	19	21	50	8	179
QUERIAM-SE			0	0	0	1	0	1
QUERÍAMOS			9	2	4	6	1	22
QUERIAS			1	5	0	0	0	6
QUERO			618	894	44	266	225	2.047
QUERO-			2	5	0	4	5	16
QUIS		3a./1a.	380	100	18	116	32	646
QUISÉ (pop)			0	8	0	2	0	10
QUISEMOS			2	0	2	0	1	5
QUISER			146	161	14	80	16	417
QUISERA			33	2	0	4	6	45
QUISERAM			20	19	5	12	8	64
QUISERDES			0	3	1	1	3	8
QUISEREM			17	10	11	11	5	54
QUISERES			0	13	0	2	1	16
QUISERMOS			1	2	12	6	3	24
QUISESSE			141	17	10	57	3	228
QUISESSEM			13	2	10	4	4	33
QUISÉSSEMOS			5	0	4	4	1	14
QUISESTE			0	9	0	0	1	10
QUISESTES			0	1	0	0	8	9

Este quadro requer alguns comentários. Em primeiro lugar as formas do paradigma que concorreram para o elevado total de 9.983 ocorrências foram: *quer* (2.849), *querem* (418), *querendo* (474), *querer* (486), *queria* (1295), *quero* (2047), *quis* (646), *quiser* (417). As flexões mais registradas são as terceiras pessoas, infinitivo e gerúndio. Contudo, por

ser um verbo modalizador são também freqüentes formas do subjuntivo como *quiser* e *queira*, fenômeno raro para a maioria dos verbos.

A lista dos verbos mais freqüentes é encabeçada pelos auxiliares **ser**, **estar**, **ter**. Até o verbo **ir** registrou um elevado número de valores modais e aspectuais, razão para estar também nos primeiros lugares da hierarquia dos verbos usuais. Constam dessa lista ainda verbos modalizadores como **poder**, ou vicários, e/ou suportes como **fazer**, **dar**; entre os de significação plena apenas **dizer**, **falar**, **olhar** e **ver**.

Estes vinte verbos registram altíssima freqüência por serem reiterados continuamente no texto. Dada a centralidade do verbo na articulação do discurso, é normal a enorme repetição dessas palavras. O que pretendo enfatizar aqui é que são apenas alguns verbos que assumem tal papel em detrimento dos cerca de 6 mil verbos registrados no *corpus*. Inversamente, porém, há um número relativamente grande de verbos (mais de 3 mil) que têm freqüência baixa, ou média apenas. E esses são a maioria. Alguns exemplos (freqüência à direita): *badalar*: 7, *beijar*: 32, *dançar*: 19, *debulhar*: 6, *endoidar*: 19, *galopar*: 23, *habilitar*: 11, *habitar*: 14, *induzir*: 81, *infeccionar*: 6, *jantar*: 62, *jejuar*: 2, *lesar*: 15, *miar*: 88, *motivar*: 10, *multar*: 8, *murchar*: 27, *ninar*: 72, *nivelar*: 15, *obstruir*: 19, *pacificar*: 11, *regredir*: 18, *relembrar*: 68, *saudar*: 84, *sepultar*: 10, *tocar*: 30, *tombar*: 59, *vetar*: 57, *violar*: 29. E também aqueles com freqüência relativamente alta mas não tanto assim: *analisar*: 381, *citar*: 401, *encostar*: 238, *liberar*: 116, *libertar*: 155, *manifestar*: 408, *reproduzir*: 176, *telefonar*: 292. O curioso a respeito desses verbos é que os totais de ocorrências são determinados apenas por umas tantas formas do verbo, como já constatamos em um verbo de altíssima freqüência como *querer*. Essas formas são sempre as mesmas flexões de tempo, modo e pessoa: o infinitivo, o gerúndio, as 3as pessoas do singular do presente e do pretérito perfeito e imperfeito; a seguir, são mais freqüentes: as 3as pessoas do plural dos mesmos tempos e na mesma seqüência. Em uns raros verbos a primeira pessoa do singular do presente e do pretérito perfeito ocorre muitas vezes. Todas as demais formas do paradigma verbal têm freqüência muito baixa (1, 2) ou nula. Pode-se concluir que a virtual possibilidade de existência de 74 formas para os 6 mil verbos da língua portuguesa registrados neste *corpus* não passa de virtualidade. Essa potencialidade não ocorre jamais nem mesmo com aqueles vinte verbos de altíssima freqüência. Essa constatação permite asseverar que é preciso rever integralmente a questão do ensino das conjugações verbais nas escolas primárias e secundárias para falantes nativos e também o ensino do verbo para estrangeiros.

5 As baixas freqüências.

No domínio dos vocábulos de baixa freqüência encontram-se as *palavras raras* as quais ocorrem principalmente nos *subcorpora* LT e LR. No primeiro (LT) estão os termos técnicos por excelência, as terminologias das linguagens especializadas. Contudo, há palavras de freqüência baixa que, além da LT, ocorrem apenas na LJ, mostrando que a mídia escrita é o veículo que faz circular vocábulos das áreas técnicas, tecnológicas e científicas dentro da comunidade geral dos falantes [tipo: *fac-simile* e *fabrico*]. Estamos falando de palavras que têm baixa freqüência no *corpus* total mas não são propriamente raras. Seu numerário exíguo se deve ao fato de serem usadas apenas quando o(s) usuário(s) aciona(m) uma determinada área do conhecimento.

Convém assinalar ainda outro fato importante relativamente à estrutura lexicoestatística geral do *corpus*. As palavras concretas têm emprego restrito e vão-se tornando mais e mais específicas à medida que descemos na escala das freqüências decrescentes até atingirmos a freqüência 1.

Temos aqui outro dado espantoso: estas palavras compõem 25% do corpus. Explicitando: num total de 42.212 palavras diferentes (lemas) 10.452 palavras ocorreram apenas uma vez. Esse resultado confirma o que havia sido constatado por Richman et al. (1971) sobre a língua inglesa na pesquisa intitulada *American Heritage Word Frequency Book*. Aí também os *hapax legomena* constituíram mais de 40% do *corpus*. Creio que no AHWFB não foram eliminados os nomes próprios como em minha pesquisa; essa é provavelmente a razão para se ter atingido uma porcentagem muito maior no inglês.

5.1 Amostragem de algumas palavras plenas de baixa freqüência: os hapax legomena. Exemplos de palavras de freqüência 1 no *corpus*:

- LR: acontecente (adj), acorcondudado (adj), ajoujo (sub), babujar (ver), conhecença (sub), conselhagem (sub), desalongar (ver), despautério (sub), glossário (sub), lábaro (sub), langoroso (adj), lanhar (ver), quaresmal (adj), regrar (ver), tulha (sub), turbilhonar (ver), xilocaína (sub), zangão (sub);
- LD: bacará (sub), lassitude (sub);
- LT: barbaresco (adj), giardíase (sub), globuloso, laceração (sub), lactase (sub), lactoflavina (sub), pacifismo (sub), palafítico (adj), palafrem (sub), paleogeno (sub), quaternização (sub), quiasmo (sub), regulamentador (adj), regurgitação (sub), relampejante (adj), tucumã (sub), tucuna (sub), tucupi (sub), xelita (sub), xenobiótico (adj), zircônio (sub);

- LJ: lanugem (sub), laparotomia (sub), padronagem (sub), paleografia (sub), quadriga (sub), queixa-crime (sub), reidratar (ver), tugúrio (sub), tumefazer (ver), turboélice (sub), zen-budista (adj).

A esmagadora maioria das palavras raras, *hapax legomena*, são substantivos. Eventualmente ocorrem alguns adjetivos e muito raramente um verbo, como demonstra o pequeno exemplário acima. Na LT a freqüência 1 é representada por palavras muito especializadas, que geralmente só são usadas neste gênero. A linguagem literária também registra elevado número de palavras raras. Às vezes são criações idiossincráticas resultantes de uma característica típica da arte: a busca da inovação. O artista viola a norma por razões estéticas, aproveitando as virtualidades de criação que o sistema lexical lhe permite e propicia. O criador literário deseja exatamente não escrever como o vulgo e evita o vocábulo banal, usual.

A seguir, o Gráfico 2, demonstrando este resultado das palavras de freqüência 1.

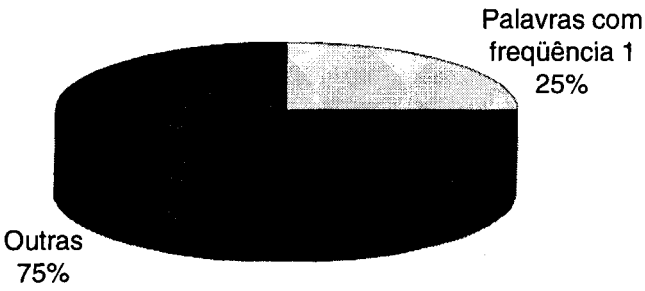

GRÁFICO 2 - Participação das palavras de freqüência 1 no *corpus*.

6 Vejamos algumas outras características lingüísticas que o *corpus* do *dicionário de freqüências* evidenciou.

A presença pouco expressiva dos *estrangeirismos* registrada na massa geral dos dados. Antes de fazer esta pesquisa eu julgava que o português brasileiro estava sendo invadido por um número muito elevado de anglicismos a ponto de eclipsar o vocabulário vernáculo. Não foi o que constatei. De fato, eles ocorrem bastante na linguagem jornalística por razões óbvias. Apenas uma área específica abunda em estrangeirismos, anglicismos em particular: a informática.[3] Estão quase ausen-

3 Ver neste volume artigo sobre anglicismos no vocabulário da informática.

tes da linguagem literária, dramática e oratória. Um problema que se repete com esses empréstimos estrangeiros é o da insegurança relativamente à sua grafia. Sobretudo em textos jornalísticos há grande oscilação na grafia de estrangeirismos, sobretudo se seu uso for muito recente na língua. Ora, a língua inglesa, principal fonte de estrangeirismos na contemporaneidade, tem fonética e fonologia bem como ortografia muito distinta do idioma português. Pode-se afirmar que a maioria dos brasileiros não sabe como grafar os anglicismos e os estrangeirismos em geral.

Registro aqui alguns exemplos de estrangeirismos que ocorreram no corpus, indicando entre parênteses outras formas variantes que ocorreram.

> *baby, baby-doll, bacará (bacarat), bacon, baguette, balé (ballet), ban-lon, bandoneón, bangue-bangue (bang-bang), baseball, bas-fond (basfond), dancing, dândi (dandy), decor, e-mail, fã (fan), fã-clube (fan-club), fade-in, fade-on, gafe (gaffe), gag, gaijim, handicap, hangover, happy-hour, hard-core, hare-krishna, head-hunters, hippie, hit, hit-parade, know-how, hobby, holding, ien (iene), iídiche, maitre, panache, pane, papabili, quetzal, rack, raconto, raffinée, rai-ban (ray-ban), rangers, ranking, sanduíche (sandwich), waffles, xogum.*

Como se vê a maioria são anglicismos mas também há palavras de origem francesa, espanhola, italiana, japonesa.

7 Com base nos dados, pode-se afirmar que o *corpus* denuncia e revela uma relativa confusão reinando na *grafia da língua*. As inseguranças registradas nos textos não se restringem apenas aos estrangeirismos, onde é compreensível a hesitação. Há também uma proliferação de grafias conflitantes no caso de empréstimos ou de nomes derivados de línguas indígenas do Brasil.

É típico, por exemplo, o caso dos vocábulos nomeadores de referentes do universo físico e cultural brasileiro. Talvez porque os primeiros colonizadores e aqueles que ensinaram os nativos a falar e depois a escrever português tiveram grande dificuldade em apreender a nomenclatura brasílica, designadora de uma realidade até então totalmente ignorada pelos portugueses. Mais ainda: no período de formação do português brasileiro esses empréstimos lingüísticos originavam-se de línguas muito diversas e desconhecidas. Certamente os colonizadores apreendiam mal sua fonética e tiveram dificuldade para grafar esses signos lingüísticos exóticos, muitas vezes compostos de uma cadeia

sonora, que incluía fonemas inexistentes no português. O fato é que existe uma variação gráfica acentuada sobretudo em alguns domínios: nomes designadores de plantas e árvores, nomes designadores de pássaros e animais; nomes designadores de tribos indígenas e de elementos das culturas aborígenes. Por vezes, ocorre uma nomeação diversificada conforme a região para o mesmo referente (ou uma variedade mínima do mesmo), coexistindo vocábulos diferentes para o mesmo *designatum* conforme a região do Brasil. Tratando-se de elementos da flora brasileira, isso é típico. Há espécies vegetais que ocorrem no Brasil inteiro, com variações pequenas dadas as dimensões continentais do país, que possui regiões climáticas e ecológicas ligeiramente distintas. Vejamos alguns exemplos:

> braúna: *baraúna, garaúna, graúna, guaraúna, guiraúna, muiraúna, ybirá-una.*

Essa mesma árvore (madeira-de-lei) é ainda chamada em algumas regiões de *canela amarela, maria preta, parovaúna, rabo-de-macaco*. Outros exemplos:

> guabiroba: *guaviroba, uvaia, uvalha.*
> bacupari: *bacupary, vacapari, vacaparilha.*
> mamona: *mamono, mamoneira, rícino, carrapateira, bafureira.*
> guapuruvu: *bacuruvu* [yba-curú-iú = pau áspero mole em tupi]
> bacurau: *curiango.*

Em relação aos nomes das tribos indígenas e de suas línguas, a variação é muito grande. Cf: [estão separados por *travessão* (–) os grupos diferentes] *ajajeni, adzanêni, izaseni, tatutapuio – arauco, arauá, araua, aruan, araó – aruaque, aruak, aroaqui, aroaco, arauac, araguac, araguaco – ariti, pareci, paressi – baé, baré, barré – baniva, baniba, baniua, maniba, vaniva, baniwa, poignare – caingang, caingan, cainguan, Kaingang, Kaingygn – capaná, cupaná, coló, capinamau – cângite, cangiti, cangutu, cankete, cankiti – craó, Kraó, arahu, caraou – chicriabá, chicriobá, shicriabá – jamamandi, iamamandi, jammandi, yamamandi – manáo, manaó, manahó, manavo, managne, manahua – maxacalí, mashacali – uarequena, varekena, variquena, ariquena.*

À guisa de conclusão, vejamos alguns pontos a serem retidos:

1. Por enorme que seja o léxico de uma língua, é reduzido o repertório desse acervo efetivamente utilizado pelos falantes do idioma. Até

mesmo na língua escrita, que é a variante da língua que se serve de um vocabulário mais rico e mais variado. E isso apesar de os recursos léxicos do idioma serem grandes e a expansão do léxico ocorrer numa progressão geométrica. De fato, o uso desse tesouro lexical por parte dos usuários da língua é bem modesto. Julgo que a razão seria a enorme limitação da memória humana. Como já afirmou Rey-Debove (1970, p.4) o usuário médio domina uns 20 mil vocábulos do idioma, incluindo-se nesse total o vocabulário ativo e passivo. Enfim, gostaria de registrar aqui conclusão da segunda pesquisa lexicoestatística que incluía um *subcorpus* de língua falada. Como seria de esperar a língua oral registrou um vocabulário infinitamente mais modesto do que todos os *subcorpora* de língua escrita.

2. É preciso ampliar as pesquisas relativas ao léxico. De fato, o vocabulário exerce um papel crucial na veiculação do significado, que é, afinal de contas, o objeto da comunicação lingüística. A informação veiculada pela mensagem faz-se sobretudo por meio do léxico, das palavras lexicais que integram os enunciados. Por outro lado, sabemos que a referência à realidade extralingüística nos discursos humanos faz-se através dos signos lingüísticos, ou unidades lexicais, que designam os elementos desse universo segundo o recorte feito pela língua e pela cultura correlatas. Assim, o léxico é o lugar da estocagem da significação e dos conteúdos significantes da linguagem humana.

Mais importante ainda: o léxico está associado ao conhecimento e o processo de nomeação em qualquer língua resulta de uma operação perceptiva e cognitiva. Assim sendo, no aparato lingüístico da memória humana, o léxico é o lugar do conhecimento sob o rótulo sintético de palavras – os signos lingüísticos. Eis por que precisamos começar a trabalhar com esta imensa galáxia de signos que devemos conhecer melhor. É preciso desvendar o mistério de como se estrutura o léxico da nossa língua.

3. Um importante problema relacionado ao léxico é o do aprendizado, tanto do vocabulário de uma primeira como de uma segunda língua. De fato, desde a década de 1940, as pesquisas de Lexicoestatística visavam chegar a um diagnóstico da estrutura quantitativa do léxico das línguas com o objetivo de elaborar listas de freqüência de palavras para selecionar adequadamente o vocabulário a ser utilizado no ensino/aprendizagem do léxico. Dada a enorme extensão do léxico, uma seleção lexical criteriosa e baseada em princípios lexicoestatísticos constituía a melhor alternativa para estabelecer os *index verborum* das palavras mais freqüentes e usuais dentre as centenas de milhares que consti-

tuem o léxico de uma língua de civilização moderna. Dessa forma pode-se evitar o empirismo na escolha do vocabulário para fins didáticos.

4. Enfim, um *dicionário de freqüências* pode oferecer também excelentes subsídios para outros usos pragmáticos baseados no léxico da língua, a saber: as telecomunicações e o reconhecimento da voz humana pelo computador e/ou equipamentos eletrônicos. Em suma: os dados estatísticos que acompanham cada um dos lemas (palavras-entrada) do nosso *dicionário de freqüências* podem prestar um grande serviço para numerosas atividades humanas, já que a palavra é o cerne da comunicação entre os homens.

Agradecimento

À Prof.ª Drª Guiomar F. Calçada, da Faculdade de Filosofia, Letras e Ciências Humanas da USP, pela ajuda na lematização do Dicionário de Freqüências.

BIDERMAN, M. T. C. The quantitative side of feature language: a Frequency Dictionary of Contemporary Brazilian Portuguese. *Alfa* (São Paulo), v.42, n.esp., p.161-181, 1998.

■ *ABSTRACT: The Frequency Dictionary of Contemporary Brazilian Portuguese lexicon was based in a corpus of the Brazilian variety of the written language, 5 million words (1950 to 1990). Some quantitative results: only 42,212 lexical units totalled 5 million word occurrences in the corpus, except for toponyms and anthroponyms. Statistical data register very high frequency of instrumental words (articles, prepositions, pronouns, conjunctions etc.) as well as auxiliary verbs and modalizers. On the opposite side are the words of very low frequency namely the* hapax legomena, *which contribute substantially to the total of 42,212 lexies* registered in this corpus. As a matter of fact, low frequency words totalize most of this* index verborum; *otherwise the vocabulary repertoire would be much smaller. The noun category contributes with the majority of words that occurred only once in the corpus, as well as the technical terms of scientific language. Journalistic vocabulary is the most neutral and the least thematical, representing an average between the other genres of language.*

■ *KEYWORDS: Lexicostatistics; frequency dicionary; hapax legomena; index verborum; multiuse vocabulary.*

Referências bibliográficas

BÜHLER, K. *Teoría del lenguaje.* 2.ed. Madrid: Revista de Occidente, 1961.

DUNCAN, J. A *Frequency Dictionary of Portuguese Words.* Stanford, 1972. Dissertation (Ph. D.) – Stanford University.

MOLINO, J. Le nom propre dans la langue. In: *Langages,* n.66, p. 5-20, juin 1982.

MÜLLER, C. Les verbs les plus fréquents du français. *Les français dans le monde,* n.103, p.14-7, mars, 1974.

REY-DEBOVE, J. Le domaine du dictionaire. In: La lexicographie. *Langages,* n.19, p. 3-34, sept. 1970.

RICHMAN, B. et al. *The American Heritage Word Frequency Book.* New York: Boston American Publishing, Houghton Mifflin, 1971.

Bibliografia consultada

BIDERMAN, M. T. C. *Teoria Lingüística*: lingüística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

JUILLAND, A., CHANG-RODRIGUEZ, E. *A Frequency Dictionary of Spanish Words.* Haia: Mouton,1964.

JUILLAND, A., Travessa, V. *A Frequency Dictionary of Italian Words.* Haia: Mouton, 1973.

JUILLAND, A., BRODIN, D., Davidovitch. *A Frequency Dictionary of French Words.* Haia: Mouton, 1971.

JUILLAND, A., EDWARDS, P. M. H., Juilland, I. *A Frequency Dictionary of Rumanian Words.* Haia: Mouton,1965.