LÉXICO E VOCABULÁRIO FUNDAMENTAL

Maria Tereza Camargo BIDERMAN[1]

■ RESUMO: O papel do léxico na estrutura e funcionamento da língua; seu lugar nos estudos da linguagem. Descrição de pesquisa realizada pela Universidade de Lisboa para obtenção de um vocabulário fundamental do português. Discussão dos conceitos e da terminologia técnica em Lexicologia. Procedimentos utilizados na identificação das unidades lexicais; será o dicionário um parâmetro confiável para a identificação dos lexemas? Comentários críticos sobre os resultados da pesquisa de Lisboa, particularmente dos dados do Inquérito da Disponibilidade. Considerações sobre a especificidade e o generalismo no vocabulário. Um vocabulário fundamental para o português do Brasil.

■ PALAVRAS-CHAVE: Léxico; vocabulário fundamental; lexicoestatística; freqüência de palavras; vocabulário de valor multiuso; terminologia lexicológica.

Introdução

O léxico: seu papel na arquitetura e funcionamento da língua. Lugar do léxico nos estudos da linguagem. O aprendizado da língua e o vocabulário.

Nas últimas décadas, os lingüistas não têm dado muita atenção a problemas de grande relevância relativos ao léxico. Contudo, 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. Sabemos, também, que a referência à realidade extralingüística nos discursos humanos faz-se pelos 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.

Por outro lado, o léxico está associado ao conhecimento, e o processo de nomeação em qualquer língua resulta de uma operação perceptiva e cognitiva. As-

1 Curso de Pós-Graduação em Letras - Faculdade de Ciências e Letras - UNESP - 14800-901 - Araraquara - SP.

sim, no aparato lingüístico da memória humana, o léxico é o lugar do conhecimento, sob o rótulo sintético de palavras – os signos lingüísticos.

Um importante problema relacionado ao léxico é o do aprendizado tanto do vocabulário de uma primeira, como do vocabulário de uma segunda língua.

Infelizmente, a aquisição do vocabulário tem sido negligenciada pela pesquisa lingüística, como bem acentuou Meara (1980). Além disso, essa pesquisa tem sido assistemática e sem continuidade, não permitindo que cheguemos a conclusões claras.

Desde a década de 1940, a justificação para a escolha dos *indices verborum* no ensino/aprendizagem do vocabulário tem-se baseado na freqüência de seu uso na língua. Essa tem sido também a técnica e a prática no ensino de uma segunda língua.

Pouca pesquisa tem sido feita sobre essa complexa matéria. Pouco sabemos sobre como o léxico é aprendido e como é estocado na memória. Esta última questão, porém, tem merecido a atenção de psicólogos que se preocupam com o problema da memória. Sob um determinado prisma, o léxico pode ser considerado um problema da memória. As entradas lexicais são, de fato, entradas da memória. Os problemas de registro, armazenamento e recuperação das palavras na codificação e decodificação da mensagem lingüística constituem uma das questões mais intrigantes da memória. Tudo leva a crer que o léxico se estrutura de tal forma que permita a recuperação muito rápida, instantânea mesmo, das palavras que o integram. Com certeza, uma das propriedades constitutivas da unidade lexical, e que possibilitam a sua recuperação no acervo da memória, é a freqüência da palavra.

Desde o início dos anos 60, pesquisas baseadas em métodos estatísticos evidenciaram a existência de um núcleo lexical no interior do léxico de um idioma, que ocorre em qualquer tipo de discurso formulado na língua em questão. Os dicionários de freqüência das línguas românicas, elaborados por Juilland et al. (1964; 1965; 1971; 1973), mostraram que, nas cinco línguas (espanhol, português, francês, italiano e romeno), cerca de 80% de qualquer texto são constituídos pelas 500 palavras mais freqüentes da língua, incluindo-se aí um conjunto de palavras de valor semântico muito geral e a totalidade das palavras gramaticais dessas línguas. Outras pesquisas foram realizadas sobre as línguas românicas com o objetivo de estabelecer vocabulários básicos para o ensino dessas línguas a estrangeiros. A ciência da Estatística Léxica ou Lexicoestatística desenvolveu-se muito em razão desse fim pragmático.

Dada a enorme extensão do léxico, uma seleção lexical criteriosa e baseada em princípios lexicoestatísticos apresentou-se como a melhor alternativa para estabelecer os *indices verborum* das palavras mais freqüentes e usuais dentre as centenas de milhares que constituem o léxico de uma língua de civilização moderna. Dessa forma, podem-se evitar o empirismo e uma seleção vocabular com base apenas na intuição.

Por conseguinte, o ensino de línguas a estrangeiros propiciou a delimitação de um vocabulário mínimo, indispensável à comunicação. Pretende-se, assim, atender às exigências de comunicação rápida do mundo moderno, visando a objetivos essencialmente práticos.

A seguir serão analisados a pesquisa e os resultados de estudos feitos em Portugal para obtenção de um vocabulário básico do português.

1 Descrição de pesquisa realizada pela Universidade de Lisboa para obtenção de um vocabulário fundamental do português

1.1 A metodologia adotada na recolha e análise dos dados do *Português Fundamental* (PF) baseou-se em pesquisas anteriores realizadas sobre o Francês Fundamental e o Espanhol Fundamental. Essas pesquisas sobre os vocabulários básicos das variedades européias das três línguas latinas – francês, espanhol e português – tinham como escopo obter um vocabulário fundamental, que pudesse ser utilizado no ensino dessas línguas a estrangeiros. Para se chegar a tal vocabulário, foram utilizados modelos de análise de estatística léxica, para se evitar a arbitrariedade na seleção das palavras que seriam utilizadas na pedagogia lingüística.

O *corpus* utilizado continha dois tipos de bancos de dados: a) *corpus da freqüência*; b) *corpus da disponibilidade*.

O *corpus da freqüência* compõe-se de textos de entrevistas gravadas, realizadas com 1.400 informantes, de ambos os sexos, na faixa etária de 15 a 65 anos e representantes da população portuguesa de todos os distritos do país, inclusive das ilhas de Açores e da Madeira. Tais entrevistas ocorreram na residência ou no local de trabalho dos entrevistados. A entrevista não seguia um modelo rígido. Deixava-se o entrevistado falar espontaneamente e estimulava-se a sua manifestação sobre um tópico que o interessasse. Os temas tratados foram variadíssimos, constituindo o *corpus* global um acervo bem heterogêneo e diversificado da língua portuguesa, na sua variante lusitana. Logo depois, o entrevistador transcrevia a entrevista ortograficamente, para não se dar o caso de ele esquecer detalhes da situação e da conversa que o ajudariam a reproduzir a entrevista o mais fielmente possível.

Das entrevistas transcritas (1.400 textos), foram extraídos excertos de 500 palavras gráficas (em seqüência), processadas no computador. O processamento computacional desses fragmentos das entrevistas forneceu as listas de freqüência de palavras, ou seja, o *corpus da freqüência*. Essas listas foram ordenadas de duas maneiras: 1. em ordem decrescente de freqüência; 2. em ordem alfabética. (Deve-se esclarecer que as unidades de texto identificadas pelo computador, posteriormente quantificadas, eram seqüências de caracteres entre dois espaços brancos. Foram assim obtidas 25.107 formas diferentes.)

1.2 A seguir, os pesquisadores procuraram identificar os *lemas* (as unidades lexicais canônicas). Sob cada *lema* foram indexadas as formas diversas que ocorreram no discurso quando se tratava de palavras flexionáveis, como os substantivos, os adjetivos e os verbos. Se o quantitativo de um lema atingisse o limiar de quarenta, esse vocábulo era listado no rol das palavras que se consideravam fundamentais. Se o total das ocorrências de um vocábulo não totalizasse quarenta, o lexema era rejeitado, isto é, não era considerado parte do vocabulário fundamental da língua.

Ao fim da tarefa de identificação dos lemas com freqüência igual ou superior a quarenta, constatou-se que muitas palavras de uso freqüente na fala cotidiana não haviam ocorrido. Isso já era esperado, em virtude das experiências anteriores do Francês Fundamental e do Espanhol Fundamental. Por isso a equipe lançou-se à recolha do *corpus da disponibilidade*. De fato, em entrevistas não dirigidas, em que se dá o máximo de liberdade ao locutor para garantir a espontaneidade da fala, os tópicos da conversação passam a ser aleatórios, acarretando um léxico também aleatório, uma vez que certos vocábulos só ocorrem quando se fala de determinados assuntos e em certas situações. Sucede ainda que a situação e o contexto suprem a necessidade de mencionar verbalmente uma ou outra palavra, substituída por gestos ou por dêiticos. Contudo, esses vocábulos são muitas vezes indispensáveis para a comunicação. Por isso, a segunda etapa do trabalho foi a recolha do *corpus da disponibilidade*. Organizou-se um inquérito com base em 27 *Centros de Interesse* (CI), a saber: 1. corpo humano; 2. vestuário; 3. estabelecimentos de ensino; 4. saúde e doença; 5. higiene pessoal; 6. desportos; 7. refeições, alimentos e bebidas; 8. cozinha e objetos que vão à mesa; 9. meios de transporte; 10. viagens; 11. a cidade; 12. aldeia e trabalhos do campo; 13. casa; 14. família; 15. vida sentimental; 16. correio; 17. meios de informação; 18. casas comerciais; 19. profissões e ofícios; 20. arte; 21. tempo (condições atmosféricas); 22. religião; 23. café; 24. animais; 25. plantas, árvores e flores; 26. divertimentos e passatempos; 27. verbos referentes à vida mental.

Os inquéritos foram dirigidos a 800 pessoas entre 17 e 18 anos, em virtude da maior disponibilidade dos jovens, já que se tratava de pesquisa demorada. Os resultados vocabulares desses inquéritos totalizaram 465 mil palavras que se distribuíram por campos semânticos relacionados com a vida diária e a realidade sociocultural de Portugal. Entrementes, sobreveio a revolução política portuguesa de 25 de abril de 1974. Conseqüentemente, ocorreram mudanças sociais e políticas profundas, depois de quarenta anos de ditadura em que Portugal quase se isolara do mundo europeu. Ora, o léxico reflete diretamente a realidade sociopolítica e cultural. Por isso os responsáveis pelo PF concluíram que seria indispensável completar o inquérito sobre a disponibilidade para investigar a possível incorporação à língua comum de novos vocábulos surgidos após o 25 de abril. Fez-se, pois, um *inquérito complementar* em 1980 sobre os seguintes tópicos: 1. vida política; 2. relações de trabalho; 3. problemas econômicos de caráter coletivo.

1.3 A pesquisa da freqüência tinha fornecido de 500 a 600 palavras de sentido muito geral e de valor polissêmico (substantivos, adjetivos, verbos e advérbios) que podem ocorrer em situações de comunicação muito diversas, exprimindo conceitos classificatórios, tais como: *tempo, trabalho, homem, indivíduo, família, problema, parte, tipo, cidade, pensar, conhecer, comprar, vender, deixar*. Por conseguinte, trata-se de um vocabulário restrito, podendo servir a um nível de competência modesto.

Por outro lado, os *inquéritos de freqüência vocabular* destinam-se a selecionar um *vocabulário de valor multiuso*, ou seja, que possa ser utilizado num número grande de situações, independentemente da época, do lugar, dos interlocutores em presença e dos conteúdos tratados. Dessa forma, nesta pesquisa, assim como em outras similares, os substantivos e verbos mais freqüentes no *corpus* recolhido eram palavras de carga semântica muito geral; de fato, tais vocábulos se caracterizam por funcionarem como pro-formas no discurso. Confira-se por exemplo: *coisa, pessoa, gente, maneira, problema, caso, fato, momento, fazer, dar, gostar, ficar*. Com respeito aos adjetivos, constatou-se que os mais freqüentes exprimem qualidades (positivas ou negativas), ou julgamentos gerais e vagos, "constituindo uma primeira aproximação pouco precisa e sem gradações", tais como: *grande, pequeno, pior, bom, melhor, novo, difícil, caro*.

A recolha do *corpus da disponibilidade* baseou-se no critério de *vocabulário disponível*, conceito estabelecido pela equipe do Francês Fundamental. Um *vocabulário disponível* é constituído de palavras de baixa freqüência e pouco estáveis, mas usuais e úteis. Esse vocabulário está associado aos interesses dos falantes, a suas motivações e intenções comunicativas, aos conteúdos escolhidos para a comunicação lingüística. Em suma, o *corpus da freqüência* baseia-se na *performance*, na situação de interação verbal, e o *corpus da disponibilidade* resulta do nível de competência do falante, já que a recolha de dados tem como fundamento a memória verbal dos informantes.

2 Questões lexicológicas e terminologia técnica: o sistema lexical e a unidade léxica

2.1 Em se tratando de trabalho lexicológico, na selva terminológica que nos cerca, é preciso estabelecer com clareza os limites de um conceito, bem como defini-lo com precisão, não ignorando a tradição, com seus acertos e erros. Por outro lado, é perigoso misturar as teorias, pois como bem afirmou Saussure, o ponto de vista do cientista cria o objeto. Ora, a pesquisa do PF era um trabalho de estatística lexical; assim, era natural que os pesquisadores se reportassem a um especialista na área, a saber: Charles Muller. Embora Muller tenha examinado com rigor a questão da unidade léxica, não foi feliz na cunhagem de seus termos, assim como

na referencialidade que lhes atribui. A crítica principal que lhe faço é a de utilizar palavras da linguagem comum para cunhar conceitos lexicoestatísticos novos, a saber: "*mot*" et "*vocable*". A oposição que Muller estabelece entre *mot* e *vocable* é pertinente; foi inadequada, porém, a etiquetagem dos termos. Se *mot* é uma palavra inadequada por causa de sua imprecisão e de seu desgaste lingüístico, o mesmo se pode dizer de *vocable*. O dicionário etimológico de Bloch & Wartburg (1950) informa que *vocable*, registrado por volta de 1400, é usual no século XVI. Mais tarde tornou-se raro, voltando à circulação no século XIX. O *Littré* (1964-1965) define *vocable* : "*1º Terme de grammaire. Mot, partie integrante du langage*". A seguir dá abonações do século XVI (Rabelais e Montesquieu). Por conseguinte, a despeito de não ser usada durante um certo tempo, *vocable* é palavra antiga no francês, podendo ser considerada sinônima de *mot* nesse idioma. Donde se deduz que melhor fora que Muller evitasse o termo *vocable*, já que ele será sempre co-referido a *mot*. Na literatura técnica em língua inglesa, estatísticos lingüísticos de renome, como Herdan, opuseram os termos *type* e *token*, que tampouco são adequados. Contudo, não oferecem a desvantagem adicional de procederem do repertório da Lexicologia.

A despeito da conceituação registrada no *Dictionnaire de Linguistique*, de Dubois et al. (1973), julgo que não se deva reproduzir Muller como aí consta. Em português como em francês, *vocábulo* (*vocable*) é palavra antiga da língua, sendo igualmente sinônima de *palavra* (*mot*). Bluteau (1712-1721) e Morais Silva (1.ed., 1789; 2.ed., 1813) registram esses termos com esses valores semânticos. E lembremos ainda que a fonte de dados de ambos são textos e autores dos séculos XVI e XVII. Em suma, para um novo conceito de *Estatística Lexical*, deve-se forjar um termo novo, para evitar imprecisões, ambigüidades.

Julgo que não é adequada a definição atribuída ao termo técnico *vocábulo* pela equipe do PF – "vocábulo é a unidade paradigmática que, no caso de ser flexionável, pode revestir no discurso várias formas" ("Métodos e técnicas", 1987, v.1, p.317). Ou seja: em Lexicologia, deve-se evitar *vocábulo*, referindo-se à unidade do sistema lexical. Por outro lado, ainda na esteira de Muller, os pesquisadores do PF utilizaram o termo *palavra* para designar a forma ocorrida no discurso. Assim, estabelecem equivalência entre: *palavra* = ocorrência = forma, por oposição a *vocábulo* (unidade abstrata da língua). Melhor seria usar indiferenciadamente as palavras *vocábulo* e *palavra* para as realizações discursivas, continuando a longa tradição do português, lembrando a imprecisão desses termos, e respeitando a sinonímia implícita na mente dos falantes do idioma no que respeita a essas palavras. No caso da unidade lexical abstrata, será melhor utilizar o termo *lexema* e chamar de *lema* sua representação canônica no dicionário. Por conseguinte, estabeleceríamos as seguintes oposições e correlações: *léxico* é o conjunto abstrato das unidades lexicais da língua; *vocabulário* é o conjunto das realizações discursivas dessas mesmas unidades. No plano das realizações discursivas qualquer seqüência significativa será chamada indiferente e imprecisamente de *palavra* ou *vocábulo*. A unidade denominativa para um conjunto de formas flexionadas que compõem um paradigma será

denominada *lexema/lema*. *Lema* é também a entrada canônica nos dicionários da língua em questão. O uso desses termos técnicos eliminaria as ambigüidades, indesejáveis em ciência. O termo *palavra* é operacional como elemento da linguagem comum. Num uso não específico é a designação pertinente, já que qualquer falante do idioma identifica o seu *designatum* sem problemas. Também o termo *forma* não é ambíguo para a designação referida. O termo *palavra* é inadequado, porém, quando se trata de identificar as *unidades léxicas* da língua (nível do sistema), sobretudo numa práxis contábil como a da *Estatística Léxica*, em que é necessário distinguir bem aquilo que se conta.

O termo *monema*, proposto por Martinet, não me parece funcional em Lexicologia. Inversamente, o termo *lexia*, proposto por Pottier, é bastante útil, sobretudo por ser um termo técnico, e não correr o risco de ser maculado com as conotações discursivas, que podem gerar a ambigüidade encontrada em *palavra* e/ou *vocábulo*. Assim, no plano da língua, o termo *lexema* refere a unidade abstrata do léxico. As manifestações discursivas dos *lexemas* devem ser referidas tecnicamente como *lexias*.

Consideremos outro problema teórico que se põe de imediato: a *identificação* das unidades léxicas no texto, em virtude das imprecisões e inadequações do sistema ortográfico e da tradição gráfica. Registram-se dois tipos de unidades: *lexias simples* e *lexias complexas*. Exemplos de *lexias simples*: *escola, meio, hora, esperar, fazer, esse, ali, alguém etc.* Exemplos de *lexias complexas*: *fim de semana, sala de jantar, dona de casa, além de, de repente, pouco a pouco, de pé, para com, fora de mão*. Portanto, *lexias complexas* são aquelas unidades lexicais que, no plano da escrita, são grafadas como uma seqüência de unidades, embora correspondam a um único referente no plano da língua.

Ainda com respeito aos conceitos teóricos básicos da Lexicologia parece-me importante clarificar mais um ponto. De que unidades se compõe o léxico? Convém insistir nessa questão, já que se constata que alguns lingüistas parecem entender diversamente a questão. Para nós, o léxico é constituído por todos os elementos lexicais da língua, vale dizer: os *lexemas* de valor lexical (as palavras plenas) e os *lexemas de valor gramatical* (as palavras gramaticais, vocábulos-morfema), que alguns lingüistas chamam *gramemas*, adotando a terminologia pottieriana. Aliás, Pottier inclui nessa classe também os afixos, somando os elementos de valor meramente mórfico às unidades de nível superior, a saber: as palavras gramaticais. Não me parece operacional essa categorização porque um tanto ambígua, a despeito de esses dois tipos de elementos guardarem semelhanças em seu uso e valor lingüísticos. Quanto a incluir no léxico tanto as palavras plenas como as palavras gramaticais, convém lembrar que essa é uma velha tradição nas línguas ocidentais. Desde o século XVI os dicionários das línguas ocidentais registram essas duas categorias de *lexemas*. Essa prática lexicográfica não pode ser ignorada, pois os dicionários são as únicas descrições globais dos léxicos das línguas.

2.2 Os pesquisadores de Lisboa não deram tratamento uniforme a cada *corpus* – o da freqüência e o da disponibilidade – relativamente à identificação das unidades lexicais. No *corpus* da freqüência, o computador segmentou os textos de maneira arbitrária, como sói acontecer no tratamento automático de textos. Assim, a máquina dividiu as seqüências gráficas, separando-as conforme os espaços brancos indicavam. Isto é: foram reconhecidas como unidades quaisquer seqüências de caracteres situadas entre dois brancos. Assim, na fase de lematização, com exceção das locuções (adverbiais, prepositivas, conjuncionais etc.), em que se recuperou a unidade lexical, grafada sob a forma de *lexia complexa*, e outros poucos casos (fogo de artifício, após-guerra, casa de jantar, sala de jantar, mestre-de-obras etc.), creio que muitas *lexias complexas* terão sido indevidamente segmentadas pelo computador, visto que a pré-codificação não parece ter sido exaustiva. No *corpus* da disponibilidade, este problema não existiu, já que os informantes tinham registrado as *lexias complexas* como tal nos boletins de inquérito. Portanto, na entrada de dados, as unidades complexas foram assim registradas.

Para identificar a unidade lexical, a equipe do PF utilizou como árbitro o *Vocabulário da Língua Portuguesa* (VLP), de Rebelo Gonçalves (1966). Embora a prática usual em trabalhos de lexicoestatística fosse escolher um dicionário como base de referência, essa decisão não foi a ideal. A metodologia de atribuir ao dicionarista a arbitragem na identificação e categorização das unidades lexicais cria vários escolhos para o lexicólogo. Não existe em língua portuguesa um dicionário que tenha operado com critérios aceitáveis pelo atual estágio dos conhecimentos em Lexicologia. Ademais, como o léxico está em perpétua mutação e movimento, acompanhando as mudanças socioculturais, nenhum dicionário conseguirá registrar fidedignamente esse acervo, pois as unidades complexas encontram-se em estágios diferentes de cristalização. A rigor, nenhum dicionário pode ser considerado árbitro. Os estatísticos léxicos têm adotado tal critério por uma questão de comodidade, sabendo contudo da sua precariedade. Trata-se sempre de um obra incompleta, inacabada, dada a natureza *in fieri* do léxico. De fato, todo dicionário precisaria ser atualizado, no mínimo, a cada dez anos.

A própria obra de Rebelo Gonçalves já era superada ao tempo da elaboração do PF, se considerarmos a relativa rapidez com que ocorrem as mudanças lexicais. Uma seqüência que, em 1966, seria um sintagma, poderia perfeitamente ter-se lexicalizado dez anos depois, o que sucede, aliás, com freqüência. Assim, por exemplo, parece que o dicionário de Rebelo Gonçalves não considera como lexicalizadas as seqüências *dona de casa* e *sala de jantar*, visto como as integra nos verbetes *dona* e *sala*, respectivamente. Não lhes dando entrada autônoma, não os considera *lexemas* do léxico português. A despeito de a referência semântica desses vocábulos justificar a correlação com as palavras de base, o fato é que os referentes que esses signos lexicais designam são diferentes dos referentes de *dona* e *sala*. É verdade que é uma antiga prática lexicográfica incluir numa entrada, como subentrada, *lexias complexas*, geradas da base (= a entrada); assim, *dona*, subentrada: *dona de casa*; *sala*, suben-

trada: *sala de jantar*. O *Aurélio* (1986) assim faz com *dona/dona de casa*; no caso de *sala*, não consta o verbete *sala de jantar*, nem mesmo como subentrada de *sala*, embora no verbete *sala* figure *sala de estar* como subentrada. O *Morais* (2.ed., 1813) não registra nenhuma delas, nem mesmo como subentradas, o que era de esperar, visto que tais realidades e/ou referentes não existiam ao seu tempo, com certeza. O *Aulete* (1.ed., 1881) não registra nenhuma delas e a razão deve ser a mesma. No caso de *sala de jantar*, não seria uma denominação freqüente naquele tempo. Contudo, neste século XX, depois dos anos 50, os referentes dos significantes *dona de casa* e *sala de jantar* já adquiriram um conceito específico, distinto de *dona* e de *sala*, justificando a inclusão de uma entrada individual para cada um deles no dicionário da língua. Creio que o semanticismo próprio dessas *lexias* é razão necessária e suficiente para individualizá-las lexicograficamente. Alguns lexicólogos preocupados com uma representação orgânica do léxico preferirão a subordinação de *dona de casa* a *dona* e de *sala de jantar* a *sala*, tendo em vista a estruturação e a ordenação dos elementos do léxico. No primeiro caso, porém, o significado de *dona de casa* já se distanciou bastante de *dona*. Valeria a pena lembrar aqui também que a linguagem humana tende a uma especialização contínua, recortando sempre mais e mais detalhadamente o mundo, criando novos *constructa*, individualizando conceitos de novas percepções e atribuindo, a uns e outros, signos lingüísticos unívocos. De um lado, porque o homem vai se apercebendo da univocidade de cada elemento do universo. Daí a heterogeneidade típica do léxico quando comparado com outros níveis da língua, já que é a instância da linguagem que dá conta, por excelência, da função referencial. De outro, os signos que nomeiam os *designata* se alteram, em virtude das mudanças ocorridas. Assim o léxico flui e reflui num moto-contínuo. Donde se conclui que, sendo o dicionário uma foto congelada de um estado do léxico, não pode jamais reproduzir esse processo incessante. Conclusão que nos levaria a desesperar de produzir dicionários. Não é, contudo, esse o ponto aqui focalizado. O que queremos ressaltar é que não vale a pena usar um dicionário como árbitro numa tarefa de estatística léxica, pois, de fato, ele não satisfaz as necessidades teóricas e práticas do lexicólogo. Melhor será o estabelecimento de uma lista de *normas e critérios* para a *identificação da unidade léxica* naquele dado momento em que se faz a estatística.

A título de exemplo, veja-se a lista seguinte de vocábulos não registrados no VLP e apontados por M. Luisa Segura no *corpus* do PF: "contracapa, hipermercado, pré-matrimonial, reclassificação, reestruturação, subdesenvolvimento, superpotência, aeroclube, eletrodoméstico, fotocomposição, fotonovela, telejornal" (PF – "Métodos e Documentos", 1987, v.1, p.329-38). Essas palavras, dentre outras indicadas pela autora, constituem compostos já lexicalizados no presente estado da língua, justificando sua inclusão no rol dos lexemas do português. Ora, não estamos cobrando do VLP o que ele não podia registrar, ou identificar como unidade lexical, na época da sua elaboração (antes de 1966); estamos apenas demonstrando o que se acabou de afirmar acima.

Outro senão advindo da eleição do VLP como árbitro foi a de terem sido consideradas apenas as *lexias complexas* de natureza nominal, visto esse dicionário não proceder uniformemente com respeito a sintagmas lexicalizados de tipo verbal, preposicional, adverbial, conjuncional etc. Os autores do PF foram levados a uma incoerência metodológica por adotar o VLP como autoridade lexical. Conferir o que diz M. Luisa Segura da Cruz à página 333: "Tendo em conta, todavia, que o emprego do hífen é simples convenção ortográfica e verificando a flutuação que existe não só no seu emprego, como na forma de indexação desses compostos em dicionários ..., consideramos ainda como unidades de texto alguns outros grupos de palavras que não se apresentam ligados por hífen no VLP, mas que nos pareceram ter já sofrido um processo de cristalização". Ou seja, os autores do PF foram obrigados a se afastar de sua autoridade lexicográfica, quando constatavam que ela não satisfazia as necessidades da identificação das unidades lexicais.

Com relação à magna questão dos *nomes próprios*, creio que a melhor solução teria sido excluí-los *in limine*. Rigorosamente, como bem assinala M. Luisa Segura da Cruz, a função do nome próprio é a identificação de um referente único. Muito embora a equipe do PF tenha estabelecido critérios razoáveis para o aproveitamento de unidades lexicais provenientes de nomes próprios, tais como as apontadas às páginas 352 e 353 (PF – "Métodos e Documentos", 1987, v.1), acho que melhor fora abandonar de vez seqüências como: Liceu Rainha Dona Leonor, Oficinas de São José, Convento de Santa Clara, a Velha, O Senhor Jesus da Piedade, Nosso Senhor Jesus Cristo, Sagrado Coração de Maria, Rio Douro, Avenida de Roma, Praça da Alegria, Cabo Branco, Península Ibérica, Português do Atlântico, Livraria Sá da Costa etc. Em inventários de lexicoestatística com finalidades como a do PF, julgo pertinente ignorar os nomes próprios.

Inversamente, parece-me excelente o tratamento dado às *locuções*. De fato, esse complexo problema lexical, tão descurado por lexicólogos, lexicógrafos e gramáticos, merecia o apuro e a precisão das trabalhosas análises realizadas pelos autores do PF. Fizeram eles um cuidadoso trabalho de garimpagem, recolhendo dados esparsos e incompletos de dicionários e gramáticas, para elaborar a sua lista provisória de 1.818 locuções prepositivas, adverbiais, conjuncionais e pronominais, como: *acima de, a frio, ainda que, além de que, ainda assim, com gosto, daqui a pouco, de acordo com, depois de amanhã, ele mesmo, eu mesmo, eu próprio, graças a Deus, seja quem for, sem mais nem menos, tanto quanto, umas vezes, várias vezes* etc. A maioria dessas locuções são unidades lexicais para as quais a língua não dispõe de elementos simples, constituindo elas, porém, unidades léxicas do português. Ademais, particularmente as locuções prepositivas e adverbiais constituem classes abertas. Em seguida, a equipe do PF elaborou testes adequados para poder decidir sobre o grau de lexicalização dessas locuções.

A análise da *homografia* foi uma das etapas mais trabalhosas da pesquisa e também a mais demorada. Para identificar os homônimos, foram examinadas as concordâncias dos 65 mil contextos contendo esses homógrafos, identificando,

caso a caso, o lexema representado por cada uma das ocorrências registradas. Por conseguinte, os resultados são números exatos. Devo acrescentar ainda a respeito da homografia que só se identificaram as palavras plenas (substantivos, adjetivos e alguns advérbios). Os casos de homografia na faixa das altas frequências (os instrumentos gramaticais) não foram examinados e identificados em contexto, devido ao objetivo último do PF, que era fornecer a lista das palavras mais freqüentes do português para o ensino da língua a estrangeiros. Ora, os lexicólogos do Centro de Linguística da Universidade de Lisboa (CLUL) sabiam que todos os instrumentos gramaticais, os vocábulos-morfema, tinham que constar do vocabulário fundamental. Portanto, não havia interesse em multiplicar, em progressão geométrica, este trabalho brutal – para distinguir, por exemplo, *o* (artigo) de *o* (pronome pessoal), de *o* (pronome demonstrativo) e assim por diante – com formas como *a, os, as, que, se, como, onde, nos* etc., de altíssima freqüência na língua portuguesa.

2.3 A análise dos dados revelou resultados estatísticos interessantes. Em casos de homonímia *substantivo* x *adjetivo* – *amigo* (substantivo) x *amigo* (adjetivo); *jovem* (substantivo) x *jovem* (adjetivo); *ideal* (substantivo) x *ideal* (adjetivo) –, a apreciação empírica dos dados induziria à categorização como adjetivo, a categoria primeira. Ora, a análise dos contextos dessas e de outras formas homógrafas revelou que os substantivos são mais frequentes.

A lista das palavras lematizadas com suas respectivas frequências fornece muitas informações interessantes, particularmente sobre os vocábulos flexionáveis. Consideremos o caso dos verbos. Examinando verbos de alta freqüência (*achar, andar, chegar, começar, dar, dizer, fazer, gostar, poder, querer, saber, ver, vir*) e alguns de frequência entre média e alta (*acabar, chamar, falar, levar, olhar, passar, pôr, trazer*), constatamos que a pessoa mais frequente é a 3ª pessoa do singular; em escala menor, a 3ª pessoa do plural; a seguir, vem a 1ª pessoa do singular. Quanto aos modos: o indicativo é o mais frequente, seguido do infinitivo. O subjuntivo teve baixa freqüência – dentre os verbos citados, só *poder* e *querer* em usos modais. O gerúndio só ocorreu com *fazer*; provavelmente se o *corpus* fosse do português brasileiro, ele seria freqüente, diminuindo conseqüentemente a freqüência do infinitivo, já que construções registradas em entrevistas, tais como "estava a pensar", "tava a chover", "passava as aulas a brincar", corresponderiam, no português brasileiro, a: "estava pensando", "tava chovendo", "passava as aulas brincando". No indicativo, modo mais frequente, os tempos mais freqüentes foram, primeiro o presente e depois o perfeito; em menor escala ocorre o imperfeito; o futuro do presente não ocorreu nenhuma vez; o futuro do pretérito uma só vez (*poder*) com valor modal.

Vejamos outros dados numéricos. Verbos em que a primeira pessoa do singular do presente do indicativo foi a mais frequente: *achar* (1.189 ocorrências de um total de 1.472), *gostar* (909 ocorrências de um total de 1.962) e *saber* (2.181 ocorrências de um total de 3.707!). Verbos em que a terceira pessoa do singular do pre-

sente do indicativo foi a mais freqüente: *acontecer, dar, dever, poder, querer* (3.891 ocorrências em 5.008!). Verbos em que o infinitivo constituiu a maioria das freqüências: *buscar, dizer* (4.310 em 6.887), *fazer* (1.770 em 5.163), *passar*.

Para não alongar excessivamente esta análise, vou fazer uns poucos comentários a respeito de outras categorias. Veja-se a forma *claro* e suas variantes (*clara, claras, claros*). Compulsando os contextos nos arquivos, constatou-se algo curioso: o substantivo feminino *clara*(-s) só ocorre em situações em que o tópico da conversa é uma receita de cozinha como em: "bate-se muito bem a clara e depois deita-se o açúcar". Opor a usos adjetivos como: "no verão usa-se roupas claras", "quando a água tá clara, vê-se bem", "porque não usava uma linguagem mais clara". Infelizmente a equipe do PF não discriminou o uso adverbial de *claro* em contextos como os seguintes: "A gente, tá *claro*, recebe só o líquido", "*claro* aqui é um centro de pesca".

O caso das receitas de cozinha já referido remete a um outro igualmente curioso. Trata-se dos *homônimos e homógrafos forma*[1] [fórma] e *forma*[2] [fôrma]. Cf. alguns contextos de *forma*[1]:

Significando <<modo, maneira>>:

(1) "... alargar a cidade duma *forma* impressionante."
(2) "Estavam organizados de *forma* política capaz de vencer a repressão."
(3) "Tinham outras *formas* de ganhar o pão de cada dia."

Significando <<estrutura corpórea, aparência física, musculatura>>:

"Ele ainda não conseguiu adquirir uma verdadeira *forma* física."

forma[2]:
Significando <<molde>>:

(a) "Juntam-se seis claras batidas em castelo, mexe-se e deita-se dentro duma *forma* que foi molhada com água fria."

(b) "Depois para desenformar mete-se a *forma* dentro de água a ferver..."

Comparando os numerosos contextos em que ocorreu *forma*[1], (1),(2) e (3) entre outros, e os pouquíssimos em que ocorreu *forma*[2] (a) e (b), cheguei à conclusão que segue. *Forma*[1] pode ocorrer em qualquer tema de conversação ou tópico de discurso e pode ser precedida de qualquer determinante, ou seguida de adjetivo ou de sintagma com valor adjetivo; ocorre também em locuções. Os únicos casos de ocorrência de *forma*[2] nos textos do PF foram em receitas de cozinha. É muito provável que tais fatos reflitam os usos comuns na língua. Afora esse emprego no domínio semântico das receitas de bolos e comidas, *forma*[2] também pode ser usada quando o tema é a confecção de cerâmicas, calçados, com o significado de <<molde>>. Suspeito também que em atividades manuais, como fazer queijo, fazer rapadura [no Brasil] numa *forma*, seria igualmente um contexto semântico com possibi-

lidade de ocorrência desse lexema; lembro, porém, que essas realidades fatuais estão desaparecendo de nossa cultura luso-brasileira. No arquivo do PF só encontrei um caso em que se falava do ofício do oleiro no qual ocorre *forma*[2], além daqueles das receitas já mencionadas.

Esses dois únicos casos comentados na classe nominal – o de *claro/clara* e o de *forma*[1]*/forma*[2] – exemplificam características específicas do léxico, que venho constatando de longa data, quando analisei exaustivamente o vocabulário da obra de Fernando Pessoa [tese de doutoramento]. Mesmo quando se trata do léxico da língua geral e não de vocabulários técnico-científicos, o vocabulário tende a se especializar. A saber: certas palavras só ocorrem quando o tópico do discurso for um determinado assunto, ou for referido um conteúdo particular, ou situação específica. Aí estão esses dois exemplos de *clara* (substantivo feminino) e de *forma*[2] (= molde) para comprová-lo. Aliás, essa é a razão por que se fez um inquérito da disponibilidade, que será comentado mais adiante.

A equipe do PF estabelecera o *limiar* de 40 ocorrências para a seleção das palavras, com base em fórmula elaborada por Paul Rivenc [FL = F1/D x N x K], baseada, por sua vez, na freqüência e na dispersão das palavras.

Uma pesquisa lexicoestatística realizada para a língua inglesa (textos escritos) – *The American Heritage Word Frequency Book* – estabeleceu o limiar de 20 ocorrências (SFI = *Standard Frequency Index*) para a língua inglesa. Analisando os valores reais do *corpus* pesquisado (mais de 5 milhões de ocorrências, composto de 17 gêneros diferentes), John B. Carroll estabeleceu esse patamar para uma distribuição normal (cf. Statistical Analysis of the *Corpus*, Richman, 1971, p.XXXIII). Segundo Carroll, para freqüências abaixo de 20, o *corpus* deveria ser extraordinariamente grande (500 milhões de palavras). A partir de 20, porém, as freqüências reais refletem as probabilidades reais de uma distribuição lognormal em um *corpus*. Em virtude desse trabalho fidedigno de Carroll, creio que poderíamos adotar o limiar de 20 ocorrências também para a língua portuguesa.

3 Comentários sobre o inquérito da disponibilidade (id). Especificidade e generalidade no vocabulário

3.1 A lista dos CIs é subjetiva e reflete uma certa visão de mundo, como já alertaram os próprios autores do PF. Eu discordaria da inclusão de um CI 16 ("Correio"). Parece-me um pouco exíguo como tema. Poderia ter sido incluído num CI que englobasse os meios de comunicação e informação – melhor do que meios de informação (CI 17). Para nós, no Brasil, também o CI 12 deveria ter uma outra denominação: em vez de "Aldeia e os trabalhos do campo", talvez "Agricultura e a vida rural". No Brasil seria preciso incluir a "pecuária"; portanto, um CI sobre esse temário poderia denominar-se: "Agropecuária e vida rural". Claro está que a pesquisa foi feita em Portugal, sobre a realidade portuguesa. Contudo, estou levantando esses problemas ao especular sobre a possibilidade de utilização dos resultados da

pesquisa do PF para a constituição de um vocabulário fundamental para o português brasileiro.

3.2 Os informantes preencheram seus boletins informativos com palavras de significação muito geral, mostrando talvez sua incompetência lingüística, jovens estudantes que eram. Assim, o verbo *fazer* foi indicado em vários CIs, às vezes, apropriadamente, formando uma locução verbal, essa sim pertencente ao campo semântico focalizado. É o caso de: CI 10 ("Viagens"): *fazer um desvio, fazer escala, fazer o itinerário* etc., todas com freqüência 1; CI 21 ("Tempo"): *fazer calor, fazer mau tempo, fazer vento*; ou ainda com *estar*: *estar calor, estar a chover, estar encoberto, estar frio, estar frio de rachar, estar quente* etc.

Relativamente à não-especificidade e caráter geral de muitos adjetivos e verbos indicados pelos informantes, vejamos alguns exemplos. CI "Meios de transporte" – adjetivos atípicos: *rápido, lento, veloz, confortável, cômodo, perigoso, incômodo*. Assim também em quase todos os outros CIs: "Viagens", "A casa e os móveis da casa", "A família e a vida familiar", "Meios de informação", "Profissões e ofícios" etc. O mesmo generalismo se verificou com os verbos. Cf. "A cidade": *andar, atravessar, entrar, ir, ver, viver* etc. "Meios de informação": *ouvir, ler, ver, falar, comunicar, transmitir* etc. "O café": *chamar, ler, tomar* etc. "Divertimentos e passatempos": *fazer, ir, ver*. "Vida política": *enganar, esclarecer, falar, ganhar, prender, unir* etc. Para muitos dos casos citados, poder-se-á afirmar que os adjetivos e verbos referidos não são despropositados no CI em epígrafe. Sem dúvida. Contudo, não são específicos desses domínios semânticos. Inversamente, às vezes, os dados recolhidos são pertinentes ao CI em apreço. São específicos os exemplos que seguem. Adjetivos – CI "Refeições, alimentos e bebidas": *saboroso, salgado, apetitoso, insosso, amargo, cozido, azedo, delicioso, picante, cru*. CI "Aldeia e os trabalhos do campo": *rural, semeado, lavrado, cultivado, fértil, ceifado, colhido, agrícola, campestre, verdejante, regado*. CI "A religião": *religioso, católico, protestante, crente, cristão, ateu, beato, sagrado, santo, budista, pagão, santificado, batizado, divino, ortodoxo, devoto, maometano, pecador, piedoso, dogmático, fervoroso, espiritual*. Infelizmente, devido à meta de se obter um número relativamente reduzido de palavras (repertório em torno de 2 mil) e do método de seleção adotado, a maioria dos termos específicos foi eliminada e não entrou para o vocabulário fundamental. Idem para o CI "Refeições, alimentos e bebidas": *comer, beber, almoçar, jantar, cozinhar, saborear, lanchar, mastigar, engolir, cozer, alimentar, fritar, cear, assar*; CI "Aldeia e os trabalhos do campo": *semear, colher, lavrar, ceifar, cavar, cultivar, regar, plantar, vindimar, mondar, arar, sachar, podar*.

Os mais específicos são os substantivos, mesmo os mais freqüentes. No *index verborum* da freqüência, mais de 60% dos substantivos são concretos, ao passo que entre os verbos contam-se apenas 35% concretos, evidenciando a referencialidade típica da categoria do substantivo e a que ponto dele depende a configuração verbal do universo.

Os dados colhidos no ID podem suscitar, pois, uma questão teórica relativamente ao léxico. Os substantivos constituem a categoria que melhor exprime a especificidade dos referentes, e não as demais classes de palavras lexicais, ou seja, adjetivos e verbos. O fato de os informantes do Inquérito de Disponibilidade terem fornecido poucos adjetivos descritivos e muitos avaliativos, opinativos bastante genéricos (cf. *bom, mau, belo, bonito, feio, extraordinário, lindo, grande, fácil, difícil, perfeito, horrível, fantástico* etc.) poderia induzir-nos a essa conclusão. Até mesmo alguns adjetivos que apareceram no CI "A arte": *clássico, colorido, abstrato, pintado, artístico, harmonioso*, podem aplicar-se a muitos referentes, embora sejam mais descritivos que os anteriormente mencionados. Por outro lado, poder-se-á argumentar que a avaliação positiva ou negativa, exprimindo juízos de valor por parte do informante, deriva do fato de se tratar de uma situação de comunicação em que o falante estaria exercitando a função expressiva ou emotiva da linguagem. Contudo, a função que deveria sobrepor-se no caso desta pesquisa seria a função referencial.

4 Um vocabulário fundamental para o português do Brasil

4.1 Extrapolando o domínio da pesquisa do PF, vou levantar algumas questões relativas à constituição de um *vocabulário fundamental* para o português brasileiro. Se quisermos partir do *index verborum* elaborado pelos pesquisadores do Centro de Lingüística da Universidade de Lisboa, devemos atentar para alguns problemas.

Alguns vocábulos constantes do PF não se usam, ou então são pouco comuns ou raros no Brasil, porque os costumes são diferentes. Por exemplo: relativamente aos hábitos alimentares (CI 7, "Refeições, alimentos e bebidas") constatamos que *pequeno-almoço* e *ceia* não se usam entre nós, e *sopa* não é tão freqüente no Brasil. *Vinho* ocupa o primeiro lugar na lista de freqüências do PF, o que não ocorreria no Brasil. Tampouco o *peixe* estaria entre as palavras mais freqüentes, precedendo a *carne*, que seria das palavras mais freqüentes no Brasil nesse campo semântico. Por outro lado, faltam nesta lista *farinha* (de mandioca) e *carne-seca* (carne-de-sol), pratos comuns e quase cotidianos em algumas regiões do Brasil, além de vários outros.

A parte do *vocabulário fundamental*, selecionada a partir do Inquérito de Disponibilidade, inclui as diferenças mais significativas entre o vocabulário usual na variedade européia por oposição à variedade brasileira. Os itens lexicais usuais em Portugal e não no Brasil podem ser de dois tipos: 1. os significantes usados no Brasil são diferentes, ou seja, usamos um termo diverso; 2. o signo total não se usa no Brasil (ou é raro entre nós) em virtude das peculiaridades do universo físico e cultural português por oposição ao Brasil. No total, as diferenças situam-se em torno de

umas cem palavras, o que não é muito num total de 2.217. Além daquelas que comento a seguir, não deveriam fazer parte de um vocabulário básico para o Brasil vocábulos como: *adega, chaminé, criado, faneca, freguesia* (= divisão administrativa), *gajo, linguado, marisco, mondar, muçulmano, nêspera, ovelha, pimento, postal* (substantivo), *posta-restante, província, taberna, tacho, vindima, vindimar, vinha.*

A seguir, confronto um pequeno rol de palavras típicas de cada uma das culturas dos dois lados do Atlântico, elencando em cada campo semântico os vocábulos selecionados para o PF e apontando seu equivalente no Brasil.

Em alguns CIs, a identidade entre o português europeu e o brasileiro é praticamente total (indico as raras diferenças entre parênteses): 1. "o corpo humano"; 3. "estabelecimentos de ensino" (*liceu* = colégio; *cábula* = vagabundo que falta às aulas; *bestial*); 4. "saúde e doença"; 6. "desportos" (= esportes); 13. "a casa e os móveis da casa" (*casa de banho* = banheiro); 14. "a família e a vida familiar" (*ralhar* = zangar); 15. "a vida sentimental"; 10. "viagens" (*bestial, giro*); 16. "o correio" (a escolha mostra a importância desta atividade para os portugueses); 15. "a vida sentimental"; 17. "meios de informação"; 20. "a arte" (*giro*); 19. "profissões e ofícios"; 24. "animais" (*cão* = cachorro); 22. "a religião"; 26. "divertimentos e passatempos" (*bestial, giro, porreiro*); 27. "verbos referentes à vida mental"; e no Inquérito Complementar, I. "vida política"; II. "relações de trabalho" (*rendimento* = renda); III. "problemas econômicos de caráter coletivo" (*rendimento* = renda; *cabaz de compras* = cesta básica).

Outras vezes, porém, existem diferenças maiores indicadas, a saber:

2. "o vestuário" (*camisola* = camiseta, malha, *cueca*(s) = calcinha(s) feminina(s), *calções* = bermuda , *fato* = terno, *fato de banho* = maiô, *gabardine* = capa-de-chuva, *peúga* = meia de homem ou soquete, *giro*); 5. "higiene pessoal" (*casa de banho* = banheiro, *desodorizante* = desodorante, *pasta dentrífica*, ou pasta de dentes, *duche* = chuveiro); 8. "cozinha e os objetos que vão à mesa" (*chávena* = xícara, *terrina* = sopeira, *jarro* = jarra, *frigorífico* = geladeira); 9. "meios de transporte" (*comboio* = trem, *autocarro* = ônibus, *eléctrico* = bonde, *metropolitano* = metrô, mota = moto); 11. "a cidade" (*café* = bar, *eléctrico* = bonde, *fumo* = fumaça, *metropolitano* = metrô, *montra* = vitrine, *sinaleiro* = sinaleiro, sinal, semáforo); 12. "a aldeia e os trabalhos do campo" (*aldeão* = roceiro, homem do campo, da roça, *apanha* = colheita, *ceifa* = colheita, roçado, charrua = arado ?, *eira* = terreiro ?, *fonte* = mina, minadeira, *monda* = poda, *quinta* = sítio, chácara, *lareira* = pedra do fogão de lenha, *sacho* = enxada); 18. "casas comerciais" (*talho* = açougue; não-usados no Brasil, pois seus referentes não existem em nossa realidade brasileira: *chapelaria, charcutaria, leitaria, pronto a vestir, retrosaria, tabacaria);* 21. "o tempo" (*arrefecer* = esfriar, *neve, nevar* são muito raros por razões climáticas); 23. "o café" (*bica* = cafezinho, *chávena* = xícara, *fumo* = fumaça, *galão* = copo alto de leite com café (média), *sande*(s) = sanduíche, *tabaco* = fumo); 25. "plantas, árvores e flores" (palavras e referentes raros

no Brasil por razões climáticas: *malmequer, tulipa, oliveira, sobreiro, castanheiro, carvalho, cerejeira, amendoeira*).

Gostaria de lembrar, ainda, que o domínio em que o vocabulário mais difere entre os dois países é o relativo a animais, plantas, árvores e flores, dada a diversidade da natureza e do clima. Portanto, não são freqüentes no Brasil as palavras *carvalho, castanheiro* (a castanheira da Amazônia só lá é comum), *cerejeira, macieira, malmequer, oliveira, sobreiro*, como já foi assinalado. O *pinheiro* que aqui só era freqüente no Paraná, em razão da *Araucaria brasiliensis*, agora se generalizou em muitas regiões por causa do reflorestamento, que tem sido incentivado pelo governo há quase duas décadas.

Inversamente, para o português do Brasil, faltam nesta lista de 2.217 vocábulos muitas palavras que designam plantas, árvores, flores, frutos e animais daqui. A título de ilustração cito algumas: *abacateiro* (*abacate*), *abacaxi, acácia, bananeira* (*banana*), *bromélia, cana, canavial, cafeeiro, cafezal, capim, caqui, goiaba, goiabeira, coco, coqueiro, ipê, jabuticaba, jacarandá, manga, mangueira, mandioca, maracujá, milho, palmeira, palmito, paineira, primavera, quaresmeira, orquídea, quiabo, samambaia, xaxim* etc. De fato, sendo a natureza do Brasil tão exuberante, a lista de palavras nesse campo semântico é grande. Claro está que não se incluiria num vocabulário fundamental um rol das numerosíssimas plantas, flores e frutas brasileiras; tão-somente aquelas que se podem encontrar corriqueiramente nas feiras, mercados, supermercados, floriculturas, jardins e parques públicos. Algo de semelhante pode ser dito a respeito da fauna brasileira e dos vocábulos que designam os referentes desse domínio semântico.

Também em outros domínios culturais, como a culinária e a música, por exemplo, encontraremos vocabulário ligeiramente diferenciado em virtude das especificidades de cada uma das duas culturas de língua portuguesa.

Gostaria de questionar, ainda, o total de palavras atribuído pela equipe de pesquisadores do PF ao conceito de *vocabulário fundamental*, seguindo as pegadas do Francês Fundamental e do Espanhol Fundamental. Não creio que se deva considerar o *vocabulário fundamental* como um repertório lexical mínimo, numericamente igual a 2 mil ou pouco mais de 2 mil palavras. Se considerarmos a heterogeneidade do universo e a complexidade da sociedade contemporânea, não se pode postular um tal repertório para as necessidades de comunicação no mundo contemporâneo. Assim, proponho que seja adotado como *vocabulário fundamental* um montante de 3 mil palavras, aproximadamente.

Do *Dicionário de freqüências do português contemporâneo* (variedade brasileira), cuja primeira versão acabamos, extraí um *index verborum* de palavras com freqüência superior a 40 e confrontei esses dados vocabulares com a lista do PF. O referido dicionário tinha-se fundamentado numa grande base textual (*corpus* do Português Contemporâneo ou CP), coletada e estocada no Centro de Estudos Lexicográficos da FCL da UNESP, Campus de Araraquara, sob a direção do Prof. F. da S. Borba.

Nosso *corpus* brasileiro totalizou 5 milhões de palavras da língua escrita de 1950-1990, assim composto: 1 *literatura romanesca* (romances de contos), 2. *literatura dramática*; 3. *literatura técnico-científica*; 4. *literatura jornalística* (revistas e jornais de maior difusão no Brasil); 5. *literatura oratória* (discursos parlamentares e de presidentes, bem como sermões religiosos).

Examinei o rol desses lemas de frequência igual ou maior que 40 (variedade brasileira do português), confrontando-os com os resultados do PF. A grande maioria dos vocábulos é a mesma para as duas variedades do português. Contudo, muitas palavras que designam referentes da realidade física e do universo cultural português não ocorreram, ou então tiveram baixa frequência em nosso *corpus*, portanto, têm uso restrito na variedade brasileira, sendo utilizadas apenas em registros específicos, como o literário, por exemplo. Inversamente, um repertório não muito grande de palavras lexicais freqüentes no Brasil é raro no português europeu. Diga-se de passagem que a linguagem literária é tipicamente aquela em que ocorrem palavras de baixa frequência e *hapax legomena*, em razão dos estilos dos autores e de suas idiossincrasias. Também o vocabulário técnico-científico registrou um volume muito grande de palavras raras e de *hapax legomena* numa clara evidência da especialização dos vocabulários das linguagens técnicas e científicas. Claro está, contudo, que tais palavras não interessam quando se trata de vocabulário fundamental.

Convém lembrar também que nossa base textual constitui um *corpus* da língua escrita bastante grande (cinco vezes maior que o de Lisboa). Dada a abrangência do uso da linguagem escrita e a heterogeneidade intencional de nosso *corpus*, ele pode incluir também palavras de realidades que não a brasileira. Inversamente, o vocabuláno fundamental do PF baseou-se num *corpus* da língua oral. Assim, muitas das discrepâncias entre o nosso *index verborum* e o do PF pode-se dever ao abismo que suponho existir entre o vocabulário da língua falada e o da língua escrita, sendo o dessa última infinitamente mais rico e variado. Eis por que pretendo refazer o *dicionário de frequências*, incluindo um *subcorpus* da língua falada no *corpus* geral. Só assim poderemos chegar a conclusões realmente pertinentes sobre o vocabulário básico ideal para o português contemporâneo (variedade brasileira), o qual possa servir ao ensino do léxico da língua, tanto a falantes nativos como a aprendizes estrangeiros de nossa língua. Vocabulário fundamental esse que poderá servir ainda para elaborar produtos informáticos, especialmente no domínio das telecomunicações.

Para concluir, lembro que a herança cultural é passada às novas gerações através da linguagem. A língua é o veículo por excelência da transmissão da cultura. E o *léxico* da língua constitui um tesouro de signos linguísticos que, em forma de código semiótico, permite esse milagre. De um lado, ele pode ser transmitido verbalmente pela interação humana e social no processo da educação informal e formal, via aprendizagem. E, de outro, ele pode ser armazenado em forma codificada de engramas na memória do indivíduo, para que ele possa recuperar as palavras nesse tesouro vocabular, quando delas precisar para se expressar ou para se comunicar.

BIDERMAN, M. T. C. Lexicon and basic vocabulary. *Alfa (São Paulo)*, v.40, p.27-46, 1996.

■ *ABSTRACT: The role of lexicon in language structure and functioning; the place of lexicon in language studies. Description of research at the University of Lisbon to obtain a basic vocabulary of Portuguese. Discussion of concepts and technical terminology in Lexicology. Proceedings used in the identification of lexical units, answering the question: is a dictionary reliable in the process of identifying lexemes? Critical commentaries on the results of Lisbon research, mainly on the results of the "Inquérito de Disponibilidade" (Inquiry of Disponibility). Considerations on the specificity and generality of vocabulary. A basic vocabulary for Brazilian Portuguese.*

■ *KEYWORDS: Lexicon; basic vocabulary; lexicostatistics; word frequency; multiuse vocabulary; lexical terminology.*

Referências bibliográficas

BLOCH, O., WARTBURG, W. *Dictionnaire étymologique de la langue française*. 2.ed. Paris, 1950.

BLUTEAU, R. *Vocabulário portuguez e latino*. Coimbra: Colégio das Artes, 1712-1713. v.1-4; Lisboa: Pascoal da Sylva, 1716-1721. v.5-8.

CALDAS AULETE, F. J. *Diccionario contemporaneo da lingua portugueza*. 1.ed. Lisboa: Parceria Antonio Maria Pereira, 1881.

CARROLL, J. B., DAVIES, P., RICHMAN, B. *The American Heritage Word Frequency Book*. New York: Boston American Heritage Publ., 1971.

DUBOIS, J. et al. *Dictionnaire de linguistique*. Paris: Larousse, 1973.

HOLANDA FERREIRA, A. B. de. *Novo dicionário da língua portuguesa*. 2.ed. Rio de Janeiro: Nova Fronteira, 1986.

LITTRÉ, E. *Dictionnaire de la langue Française* (1863-1873). Paris: Gallimard/Hachette, 1964-1965.

MEARA, P. Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics Abstracts (New York)*, v.13, n.4, p.221-42, oct. 1980.

MORAIS SILVA, A. de. *Dicionário de língua portuguesa*. Fac-símile da 2.ed. [ed. modelo], 1813. Photographada pela "Revista de Língua Portuguesa" sob a direção de Laudelino Freire. Rio de Janeiro: Officinas da S. A. Litho-Typographia Fluminense, 1922.

STATISTICAL Analysis of the Corpus. In: RICHMAN, B. et al. *The American Heritage Word Frequency Book*. New York, Boston: America Heritage Publishing, Hougon Mifflin, 1971.

UNIVERSIDADE DE LISBOA. Centro de Lingüística. *Português fundamental*: vocabulário e gramática. Lisboa: Instituto de Investigação Científica, 1984. v.1.

_____. *Português fundamental*: métodos e documentos. Lisboa: Instituto de Investigação Científica, 1987. 2v.

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. *Frequency Dictionary of Spanish Words*. Haia: Mouton, 1964.

JUILLAND, A., EDWARDS, P. M. H., JUILLAND, I. *Frequency Dictionary of Rumanian Words*. Haia: Mouton, 1965.
JUILLAND, A., BRODIN, D., DAVIDOVITCH, C. *Frequency Dictionary of French Words*. Haia: Mouton, 1971.
JUILLAND, A., TRAVESSA, V. *Frequency Dictionary of Italian Words*. Haia: Mouton, 1973.
LAUDANNA, A., BURANI, C. (Coord.) *Il lessico*: processi e rappresentazioni. Roma: La Nuova Italia Scientifica, 1993.
MULLER, C. *Initiation à la statistique linguistique*. Paris: Larousse, 1968.