Alfa, São Paulo
28(supl.):1-26, 1984

A CIÊNCIA DA LEXICOGRAFIA

Maria Tereza Camargo BIDERMAN*

RESUMO: Três tópicos são examinados: 1) História sucinta da Lexicografia de duas línguas latinas (espanhol e francês) e do português. São avaliados os principais dicionários dessas línguas do século XVI ao século XX. 2) Tipologia de obras lexicográficas. São indicados e comentados os principais tipos de dicionário existentes nas línguas latinas e no inglês. 3) O uso do computador na Lexicografia contemporânea. Essa máquina revolucionou a Lexicografia, podendo executar tarefas básicas e enfadonhas como: compilar, classificar e ordenar dados léxicos e contextuais para a confecção de dicionários e depois recuperá-los facilmente e com rapidez.

UNITERMOS: Lexicografia; thesaurus; *dicionário, dicionário histórico; dicionário etimológico; dicionário ideológico; dicionário técnico; dicionário científico; dicionário de freqüência; banco de dados léxicos; corpus;* indices verborum*; concordância; dicionário de máquina.*

I. HISTÓRIA SUCINTA DA LEXICOGRAFIA DE ALGUMAS LÍNGUAS LATINAS E DO PORTUGUÊS

A antigüidade não produziu obras lexicográficas no sentido que hoje damos a esse termo. Os únicos trabalhos de cunho vagamente lexicográfico daquelas eras são os glossários, sobretudo os produzidos pela escola grega de Alexandria e, entre os latinos, o *Appendix Probi*. Esses precursores do moderno lexicógrafo eram, na verdade, filólogos ou gramáticos, preocupados com a compreensão de textos literários anteriores, ou com a correção de "erros" lingüísticos. Os filólogos alexandrinos, p.ex., buscaram elaborar léxicos e glossários sobre os textos homéricos para a sua melhor compreensão. O gramático romano Varrão (I séc. A.C.) em *De lingua latina* tentou fornecer dados de natureza semântica e etimológica sobre algumas palavras latinas.

Na Idade Média valeria a pena lembrar apenas as *Etimologias* de Santo Isidoro de Sevilha (570-636) e alguns glossários. As *Etimologias* têm escasso ou nulo valor científico e lingüístico, sendo muito fantasistas. Baseiam-se numa concepção mística do mundo e da linguagem que tende a referir a língua e as palavras a um sistema de significação que se reporta a Deus, adulterando-lhes, pois, a natureza. Na verdade, documentam o mundo cultural da Idade Média e sua concepção de universo.

Entre os glossários podemos citar o *Glossário de Reichenau* (séc. VIII D.C.) e o *Glossário de Cassel* (séc. IX D.C.) em terras do império carolíngio. Os dois *Glossários de Reichenau* contêm pouco mais de 2.000 vocábulos. São listas de palavras tiradas da *Vulgata* (versão latina da bíblia) de difícil compreensão para a época do autor, traduzi-

* Departamento de Lingüística - Instituto de Letras, Ciências Sociais e Educação - UNESP - 14800 - Araraquara - SP.

das no vernáculo românico da região. *O Glossário de Cassel* (265 palavras) é similar; trata-se de tradução do latim para o vernáculo germânico da região.

Também em terras hispânicas foram elaborados alguns glossários: as *Glosas Emilianenses e Silenses* (séc. X ou XI).

A verdadeira lexicografia, porém, só se vai iniciar nos tempos modernos. Os primeiros dicionários espanhóis foram: o *Universal Vocabulario* de Alonso de Palencia (1490) e os vocabulários *Latino Español* (1492) e *Español Latino* (1495) de Antônio de Nebrija, autor também da primeira gramática espanhola. Aliás, no século XVI na Europa, a lexicografia que se estava formando e desenvolvendo compreendia apenas os dicionários bilingües como esses de Nebrija. Quando o homem renascentista começou a ampliar os seus horizontes culturais abandonando de vez a sua reclusão medieval dentro de sua própria cultura, descobriu a necessidade de aprender línguas, evidentemente as línguas européias mais faladas na época (século XVI). Além da consciência adquirida da distância entre o latim e as línguas vernáculas do seu tempo, o homem renascentista precisava de outros instrumentos de intercâmbio lingüístico num mundo que se abria para um novo diálogo e trocas entre as jovens nações européias. Assim, multiplicam-se os dicionários bilingües na Espanha, na França, na Itália, em Portugal, bem como as gramáticas de cada uma das línguas que se tornaram oficiais para as nações-estado da Europa no século XVI.

Os dicionários seicentistas eram cheios de lacunas e os dicionaristas da época copiavam-se uns aos outros.

A lexicografia monolingüe surge e se desenvolve ao longo do século XVII, aperfeiçoando, aos poucos, as suas técnicas. O *Tesoro de la Lengua Castellana* de Covarrubias é de 1611. Tem muito aspectos positivos até hoje. O dicionário da Academia Espanhola - *Diccionario de Autoridades* — iniciou sua publicação em 1739. Terá sucessivas edições nos séculos XVIII, XIX e XX. A última é de 1983. Foi-se aprimorando através dos séculos. É de tipo seletivo e normativo. Uma sigla que o designa geralmente D.R.A.E. (Diccionario de la Real Academia Española).

No século XVII, o "grand siècle" da civilização francesa, vários são os dicionários monolingües do francês de boa qualidade para a época: Richelet (1680), Furetière (1690), o dicionário da Academia Francesa (1694). No século XVIII além da nova versão do Dicionário da Academia (1718) e do excelente *Dictionnaire de Trévoux,* um dos mais importantes feitos lexicográficos da época foi a obra dos enciclopedistas franceses. Com eles nasceu o modelo de enciclopédia que hoje adotamos, ou seja, um repertório geral dos conhecimentos humanos. A despeito de ter ficado aquém das aspirações de Diderot devido à desigualdade entre os seus colaboradores, é um trabalho notável para o seu tempo. No verbete *dicionário* eis a definição escrita por Diderot: "Num dicionário da língua francesa, há principalmente três coisas a considerar: a significação das palavras, o seu uso e o tipo de palavras que devemos incluir neste dicionário. A significação das palavras se estabelece por boas definições; seu uso, por uma excelente sintaxe; seu tipo, enfim, pelo próprio objetivo do dicionário. A esses três objetivos principais, pode-se acrescentar três outros subordinados a esses: a quantidade ou a pronúncia das palavras, a ortografia e a etimologia" (5, p. 102). Essa definição é válida ainda para os nossos dias. No século XIX amplia-se o número de obras lexicográficas francesas e apura-se a sua qualidade. Alguns dicionários da época: Laveaux, Raymond, Landais, Academia (1835), Littré (1872), Larousse (1866-1876), o *Dictionnaire Général* de Hatzfeld e o dicionário medieval de Godefroy. Dentre esses o *Littré* pode ser considerado uma obra-prima da lexicografia francesa, mesmo para os modernos critérios lexicográficos. Littré dedicou-se monacalmente à confecção do seu dicionário

durante 30 anos. Foi um inovador para o seu tempo; embora o seu exemplário só inclua autores anteriores a 1830 (os clássicos para Littré), constitui um modelo de repertório léxico e de escolha de citações como ilustração das palavras-entrada. O dicionário de Pierre Larousse teve dimensão considerável: 17 volumes. O seu *Grand Dictionnaire Universel du XXème Siécle* mostra uma vocação mais de enciclopedista do que de dicionarista — caso de Littré — de quem se distingüia também por ser menos purista e mais liberal. Esse "dicionário universal" constitui um notável repositório de informações sobre a sua época.

No século XX, na lexicografia francesa, existe uma grande abundância e variedade de dicionários e enciclopédias, fenômeno que se registra em várias das grandes culturas e civilizações contemporâneas. Assim ocorre na italiana, na alemã, na espanhola, na luso-brasileira, na anglo-americana, etc. Atualmente a lexicografia se expande e assume modalidades várias em função do vasto público, das grandes massas sequiosas de informações sobre a sua língua, sobre as línguas estrangeiras e sobre o universo. O dicionário se tornou um objeto de consumo de primeira necessidade.

Quanto à lexicografia francesa contemporânea, convém assinalar os vários dicionários Larousse, Robert, etc. A série dos Larousse vai desde obras elementares como *Mon premier Larousse, Nouveau Larousse des débutants* (1977), *Dictionnaire du vocabulaire essentiel* (1963) até o *Grande Larousse Encyclopédique* de 10 volumes. Como dicionários da lingua os *Robert (Grand Robert, Petit Robert, Micro Robert)* são certamente os melhores e modelo exemplar de trabalho lexicográfico. Contêm abundantes sinônimos e excelente exemplificação. Os significados da palavra no interior do verbete são classificados partindo-se do sentido mais antigo. Seu exemplário (moderno e contemporâneo sobretudo) é excelente. O *Grand Robert* procurou ser uma espécie de tesouro do francês contemporâneo e o *Micro Robert* (30.000 verbetes) é um ótimo instrumento de uso escolar e ideal para utilização no ensino do francês como segunda língua.

No domínio das enciclopédias deve-se apontar também algumas obras magistrais: *Encyclopédie Française, Encyclopédie de la Pléiade, Encyclopédie du Savoir Moderne,* etc.

Atualmente o "Institut de la Langue Française" está executando a gigantesca tarefa de compilar o maior arquivo de dados léxicos da língua francesa de todos os tempos, utilizando recursos computacionais e documentais muito modernos, bem como uma grande equipe de lexicógrafos, analistas de sistema, documentalistas e outros. O *Trésor Général des Langues et des Parlers Français* deve ser o maior acervo documental jamais compilado sobre uma língua. Só os fundos documentais sobre o francês dos séculos XIX e XX totalizam cem milhões de palavras (ocorrências). O *Dictionnaire de la Langue Française du 19ème. et du 20ème. Siècle* (nove volumes publicados e mais seis a publicar) constitui certamente o mais monumental trabalho lexicográfico já empreendido e executado sobre uma língua natural. Contém todo tipo de informação lingüística sobre a palavra-entrada, a partir de dados documentais reais que são citados para abonar cada significado, cada uso referido. O verbete inclui construções sintáticas e seus valores semânticos acompanhados de abonações totalmente identificadas: autor, nome da obra, data da publicação, página. Além disso, o dicionário fornece uma seqüência cronológica dessas abonações. Também indica os usos da palavra conforme o registro lingüístico: discurso oral ou escrito, discurso geralmente literário, especificamente literário, literário e raro, antiquado, raro. Informa ainda em que tipo de linguagem especial o termo é usado e qual o seu significado nesse registro, que pode ser: música, liturgia católica, fotografia, etc. Contém também informações sobre a pronúncia, a ortografia, a etimologia e a história da palavra. Finalmente, aponta dados quantitativos sobre

a palavra: a freqüência absoluta literária e a freqüência relativa literária, dados estatísticos esses que se distribuem em quatro segmentos cronológicos: primeira metade do século XIX; segunda metade do século XIX; primeira metade do século XX; segunda metade do século XX.

A lexicografia portuguesa tem uma história mais pobre que a francesa. O melhor dentre os mais antigos dicionários do português é o bilingüe de Rafael Bluteau — *Vocabulario Portuguez e Latino* em 8 volumes, Coimbra 1712-1721. O dicionário do Padre Bluteau é obra muito rara. Trata-se de um dicionário bilingüe português-latim que contém muita informação e bastante variada sobre essas duas línguas. Foi escrito para um falante do português. Tem características enciclopédicas com numerosos detalhes sobre a realidade e o mundo, evidenciando a vasta cultura do Padre Bluteau. Um dos méritos desse dicionário é o de alistar todos os autores portugueses que compuseram o corpus que forneceu o exemplário das abonações dos verbetes. O dicionarista indica o autor, a(s) obra(s), o local e data da impressão. Não é apenas um dicionário bilingüe cujo objetivo seria fornecer a palavra ou expressão latina que traduzisse um termo português; na verdade, Bluteau elaborou um trabalho misto, pois a parte relativa à língua portuguesa constitui praticamente um dicionário da língua portuguesa.

Um dos mais antigos dicionários unilingües do nosso idioma é o *Elucidário de palavras e frases que em Portugal antigamente se usarão* (sic) *e que hoje regularmente se ignorão* (sic): *obra indispensável para entender sem erro os documentos mais raros, e preciosos, que entre nós se conservão* (sic) de Frei Joaquim de Santa Rosa de Viterbo, Lisboa, 1798. Como Frei Viterbo se preocupou com facilitar ao leitor de textos antigos ou arcaizantes, a compreensão dessa linguagem, sua lista de verbetes registra sobretudo vocábulos caídos em desuso, ou com valores semânticos alterados. Por exemplo: *badulaque* = guisado de carne cortado em miúdos; *barro* = lugar pequeno, quinta, casa de campo; *cahimento* = diminuição, falta, queda, desfalecimentos. Raramente indica a classe gramatical da palavra, ou qualquer outra informaçâo gramatical no corpo do verbete. Às vezes o faz como no caso dos advérbios: *cha* adv... *chus* adv... Também há poucos verbos incluídos; a maior parte da nomenclatura é constituída de substantivos. Alguns verbetes mereceram exaustivo tratamento, certamente por identificarem referentes da cultura medieval, já pouco conhecidos ao tempo de Viterbo: *cavallaria, cavalleiro. Charidade,* p.ex., tem onze entradas diferentes, numeradas com algarismos romanos. Quanto à palavra *cruz,* talvez por causa da condição de religioso do autor, mereceu carinho especial: 35 colunas e meia, o mais amplo texto do dicionário; além disso, inclui 23 desenhos ilustrativos, identificando os diferentes símbolos e tipos de cruz. A rigor o *Elucidário* de Viterbo é mais um glossário do que um dicionário.

Entre os mais abalizados dicionários do passado temos o Morais, que leva o título de "*Dicionário da Língua Portuguesa* recopilado dos vocabulários impressos até agora e nesta segunda edição novamente emendado e muito acrescentado" por Antônio de Morais e Silva. A primeira edição de 1789 é obra rara, quase inincontrolável. Morais considerou esta primeira edição como obra do Padre Rafael Bluteau, visto que escreveu no frontespício desse dicionário: "... composto pelo Padre Rafael Bluteau, reformado e acrescentado por Antônio de Morais Silva, natural do Rio de Janeiro." A segunda edição de 1813, porém, Morais já atribui a si próprio, antepondo-lhe o seu nome. Essa edição é rara. Felizmente Laudelino Freire fez dela uma reprodução facsimile em 1922 no Rio de Janeiro. É a versão mais encontrada do velho Morais em bibliotecas públicas e particulares, embora também seja obra rara. Seguiram-se numerosas edições do Morais, todas elas, porém, inferiores à edição de 1813 (2.ª ed.).

O dicionário de Morais (2.ª.ed.,1813) constitui um marco na lexicografia de língua portuguesa. É o primeiro dicionário de uso da língua, muito avançado para os padrões lexicográficos da época. Apesar de ter-se baseado na obra do Padre Bluteau, sobretudo na primeira edição, na segunda edição Morais libertou-se de seu modelo, ampliou consideravelmente a obra com respeito ao número de verbetes, incluídos, e mais que isso, apurou o seu trabalho lexicográfico. Omitiu informações de tipo enciclopédico incluídas no Bluteau, revelando consciência de que um dicionário da língua não é uma enciclopédia. No prólogo Morais informa o leitor como executou o seu trabalho, de quais critérios se serviu, repassando problemas como: o corpus usado na abonação dos verbetes, a escolha das entradas, a elaboração do verbete, a ortografia. Depois de render um preito de gratidão a seu protetor, o Senhor de Balsemão, nos tempos do exílio da Inglaterra, conta quão dedicadamente se aplicou à leitura dos clássicos portugueses que enriqueciam a copiosa biblioteca do conde. Lendo e relendo os bons autores durante seis anos, foi apurando o seu domínio da língua materna que, "como muita gente, presumia saber arrazoadamente."

Veja-se a crítica feita ao Bluteau na passagem que segue:

"Acompanhei este estudo dos livros clássicos com os auxílios de Bluteau, que achei muitas vezes em falta de vocábulos e frases, e mui freqüentemente sobejo em dissertações desapropositadas, e estranhas ao assunto, que fazem avolumar a sua obra." (7, p. IX).

Continua dizendo que escolheu no Bluteau o que era propriamente português, deixando de lado muitos termos da cultura antiga, seguindo os passos dos "melhores dicionaristas das línguas vivas". Não omitiu mais termos antigos como gostaria, para não ser acusado de omisso. Por fim, vem esta graciosa publicidade do seu dicionário:

"Do que recolhi de minhas leituras fui suprindo as faltas, e diminuições, que nele achava; e quem tiver lido o Bluteau, e conferir com o seu este meu trabalho, achará que não foi pouco o que ajuntei; e mais pudera acrescentar, se as minhas circunstâncias me não levassem forçado a outras aplicações mais frutuosas. Todavia não venderei ao público por grande o serviço que lhe fiz; basta que conheça, que lhe poupei a despesa de dez volumes raros; que lhe dou o bom que neles há, muito melhorado, por uma décima parte, ou pouco mais do seu custo, com a comodidade de não andar revolvendo tantos tomos; e isto é alguma coisa, enquanto não aparece uma outra melhor." (7, p. X).

É um excelente dicionário para a sua época. A abonação dos verbetes é geralmente recolhida nos melhores escritores dos séculos XVI e XVII. Assim fazem parte do corpus de autoridades que Morais cita como modelos de boa linguagem: Luis de Camões, Gil Vicente, Damião de Góis, Diogo de Couto, Duarte Nunes Leão, Fernão Mendes Pinto, Francisco de Sá de Miranda, Francisco Rodrigues Lobo, Garcia de Resende, Gomes Eanes Zurara, Frei Heitor Pinto, Jerônimo Corte Real, João de Barros, Padre Antônio Vieira, D. Francisco Manuel de Melo, Padre Manuel Bernardes, etc. Não foram só os escritores literários aqueles de quem Morais recolheu citações para o seu arquivo lexicográfico. Utilizou também autores de obras técnicas e científicas dos seguintes domínios do conhecimento: filosofia, moral, religião, ciências sociais, política, filologia e linguagem, matemática, física, química, astronomia, botânica, geologia, medicina, engenharia, agricultura, artes, história. Preocupou-se ainda com registrar termos de linguagens especiais de uso na linguagem comum. Um dos méritos do seu dicionário è exatamente de indicar o registro lingüístico da palavra-entrada.

Assinala, muitas vezes, o fato de o termo ser usado em determinada linguagem especial. Cf. nos exemplos seguintes: *hepático, narcótico* (med.), *ácido, solução* (quím.), *fluido, foco* (fís.), *nebuloso, órbita* (astron.), *calmaria, quarto, zarpar* (náut.),

apóstrofe, hipérbole (ret.), *decisório, usufruto* (dir.), etc. bem como conotações estilísticas de certas palavras típicas de diferentes registros de linguagem falada e/ou escrita: *bajular, barganha* (fam.), *fanico, gana, pespegar* (vulg.), *bandulho, lambada, pança* (chulo), *geringonça, gomarra* (gír.) *gaança, homízio* (ant.), *catapereiro, gelhos* (rúst.).

O Morais teve várias edições nos séculos XIX e XX; 3.ª.ed.-1823; 4.ª.ed.-1831; ...; 7.ª. ed. - 1877. De 1944-1957 a Editora Confluência publicou a 10.ª edição do dicionário de Morais em 12 volumes, versão revista e ampliada por José Pedro Machado. A página de rosto avisa o leitor que se trata de uma edição *corrigida, muito aumentada e atualizada,* pois os editores pretendiam corrigir as muitas inexatidões do Morais em matéria de ortografia (é a do acordo luso-brasileiro de 1945), de etimologia, a forma de definição da palavra e atualizar as informações científicas. Contudo, essa versão desfigurou a obra de Morais. É um outro dicionário, baseado no Morais, que contém várias gralhas, a despeito dos seus méritos.

Um bom dicionário português do século XIX é o de Frei Domingos Vieira: *Grande Dicionário Português* ou *Tesouro da Língua Portuguesa,* 1871-1874. Vieira deixou pronto o plano da obra e um arquivo de verbetes, anotações e abonações. A redação final desse dicionário foi executada por uma equipe que, após a morte de Vieira, aproveitou o trabalho por ele realizado. É um dicionário bastante completo e informativo para o século XIX. Via de regra os significados e usos lingüísticos são ilustrados com citações de bons autores. São indicados: o étimo da palavra-entrada, expressões idiomáticas e sintagmas freqüentes em que ocorra esse vocábulo lema. Tome-se como exemplo a palavra *pena.* Há três entradas homônimas: 1.ª) *pena,* s.f. (do lat. *poena*); 2.ª) *pena,* s.f. ant. por *Penha* e 3.ª) *pena,* s.f. (do lat. *penna).* O primeiro verbete, bastante extenso, discorre sobre as várias acepções de *pena*; 'castigo'; 'cuidado, sofrimento'; 'dor, moléstia'; 'dificuldade, trabalho'; sempre acompanhando as explicações de tipo sinonímico com abonações. Seguem-se numerosas expressões em que esta palavra ocorre: 'alma em pena', 'sem pena', 'com duras penas', etc. E ainda os usos especiais dessa palavra enquanto termo forense - 'punição, castigo', acompanhados de sintagmas que o encorporam: *pena capital, pena corporal, pena de talião, pena judicial, pena legal, pena pecuniária,* etc.

Não se vai discutir aqui as outras duas formas homonímicas, pois o propósito dessa exemplificação é apenas dar uma rápida informação sobre o Vieira. Geralmente esse dicionário define bem a palavra-entrada. Às vezes, a extensão do verbete é um pouco exagerada do ponto de vista do léxico em geral: o verbete *paço,* p. ex., cobre três colunas inteiras, incluindo vários sentidos, inúmeras citações e expressões formadas com essa palavra. De uma perspectiva histórica da língua talvez se justificasse; não, porém, para os contemporâneos dos dicionaristas.

O Dicionário de *Aulete, Dicionário Contemporâneo da Língua Portuguesa;* de 1881 só foi planejado e iniciado por Caldas Aulete, pois esse dicionarista morreu antes de concluir sua obra. Foi completado por Santo Valente e colaboradores. Eis a opinião abalizada de Gladstone Chaves de Melo:

"... lugar de destaque cabe ao *Aulete* entre os dicionários portugueses, não só porque ele preenche inteiramente a sua finalidade, como, principalmente, porque traça novos rumos à lexicografia portuguesa, criando o tipo do dicionário moderno. "(6, p. 40). Chaves de Melo considera a segunda edição do *Aulete* "o mais útil e o mais satisfatório de todos os nossos léxicos (sic)". Dizia isso em 1947. Elogia sobretudo a qualidade das definições, embora lamente a carência e os defeitos das informações etimológicas, bem como a insuficiente atualização do léxico técnico e científico. Em 1958 a Edi-

tora Delta publicou uma versão brasileira do *Aulete*, elaborada por Hamílcar de Garcia: *Dicionário Contemporâneo da Língua Portuguesa* (5 volumes). Essa versão tem vários defeitos, embora pretenda ser mais completa que o primeiro *Aulete*. Foram aí incluídos muitos brasileirismos. Além da abonação de autores portugueses, essa edição foi acrescida da abonação de escritores brasileiros. Entretanto, tal documentação é bastante imprecisa e incompleta, indicando-se geralmente apenas o nome do autor citado e, por vezes, de maneira truncada. A revisão gráfica é defeituosa.

A primeira edição do Cândido de Figueiredo *(Novo Dicionário da Língua Portuguesa)* é de 1899. Esse dicionário pretendia ser o repositório mais completo do léxico português de todos os tempos bem como de regionalismos portugueses, brasileirismos, e de territórios onde se falava e fala o português. Cândido de Figueiredo publicou cinco edições do seu famoso dicionário que tem qualidades mas tem também terríveis defeitos. É um dicionário rico sobretudo com respeito ao número de palavras incluídas no seu acervo léxico. Há nele muitas palavras raras. A forma do verbete é muito simples. As definições são curtas, às vezes erradas e tolas, especialmente as relativas a termos técnicos ou de procedência técnico-científica. Muitos significados registrados são obsoletos há muito tempo. Cf. as entradas: *blusa, chingo, cômodo, comunicação, excursão*. Relativamente à nomenclatura que deu entrada neste dicionário veja-se, por ex., a seguinte seqüência de verbetes na letra S: *sueste, suebérgios, sudro$_1$, sudro$_2$, sulvento, sum, sumaca, sumagral, sumagrar, sumagre, sumagreiro, sumalar, sumalário, sumanais, sumanta, sumatra, sumaúna* etc. Foram também incluídos neologismos de aceitação duvidosa como: *sucessível, sucessoral, sucessorial*, etc. O Volume do léxico arrolado parece constituir uma grande riqueza; na verdade, como o Cândido de Figueiredo não tem a configuração de um *thesaurus*, mas de um dicionário de uso comum da língua, toda essa riqueza vocabular é quase inútil. Em 1949 (?) a Livraria Bertrand publicou uma 14.ª edição (?) do Cândido de Figueiredo com base na 5.ª ed. e prefácio de Júlio Dantas, na época presidente da Academia de Ciências de Lisboa. O acadêmico Júlio Dantas afirmava nesse prefácio que o Cândido de Figueiredo era o melhor dicionário da língua no seu tempo, "o mais opulento, o mais vivo e tecnicamente, o mais perfeito." Tal afirmação é certamente muito discutível.

Nos tempos contemporâneos é preciso considerar, pelo menos, o mais popular dentre todos os dicionários da língua portuguesa: *Novo Dicionário da Língua Portuguesa* de Aurélio Buarque de Holanda Ferreira, Rio de Janeiro, Nova Fronteira, 1.ª ed., 1975. Baseado numa versão anterior publicada sob o nome de *Pequeno Dicionário Brasileiro da Língua Portuguesa* que tivera sucessivas reedições, Aurélio aumentou substancialmente a sua obra lexicográfica com o auxílio de assistentes e colaboradores. No prefácio o autor nos informa que pretendeu fazer um dicionário de "tipo médio ou inframédio". Na verdade, o *Aurélio* se aproxima do tipo *thesaurus* no que diz respeito ao número de entradas do dicionário: "bem mais de cem mil verbetes e subverbetes", segundo o próprio autor. Um dicionário médio seria um dicionário de 40.000 a 50.000 verbetes aproximadamente e um inframédio seria um tipo *Micro Robert* contendo uns 30.000 verbetes. O *Aurélio* chega a lembrar a configuração do Cândido de Figueiredo (5.ª ed.) com relação ao número de entradas e ao tipo de palavras do léxico que ele abriga. Evidentemente distancia-se do Cândido de Figueiredo, relativamente à seriedade e probidade no tratamento de muitos verbetes, como aqueles que descrevem a terminologia técnico-científica. O *Aurélio* acolheu muitas palavras raras, um grande número

de regionalismos, de vocábulos desusados ou obsoletos, de termos exclusivamente literários, um vasto acervo de termos técnicos e científicos, o que inchou demais a sua nomenclatura. Veja-se, por exemplo, uma seqüência qualquer de entradas encontradas na letra A: *aliazar* (var. de *aljazar), alicali* (reg.), *alicantina* (espanholismo?), *alicome* (var. de *alincorne), alidade, alismatácea* (bot.), *alissóide* (mat.), *aljazar* (reg.?), *almácega* (reg. port.?), *almargio, alotriógnatos* (zool.). E vai por aí afora: *almonal, antomídeos, aragonita, arapuar, ardeiformes, ardômetro, ardosieira, arecal, arecíneo, areiabranquense, areia-engolideira, areia-gulosa, areia-manteiga, areia-preta, areias-gordas.* Resolvi fazer um teste com a letra R, a qual letra recobre um conjunto médio de palavras do léxico global, considerando-se o número de palavras-entrada de cada letra no dicionário. Num total aproximado de 2.628 palavras (págs. 1189-1225) eu não conhecia 989, ou seja, uns 27% desse total. Claro está que para uma pessoa que tem a competência vocabular que tenho, sendo além disso um especialista em Lexicologia e Lexicografia há vários anos, representa um índice elevado de desconhecimento do léxico total do português. É verdade que nenhum falante por mais competente que seja em matéria vocabular, jamais conseguirá incluir no seu léxico ativo e passivo grandes parcelas do léxico geral da língua. Isso se verifica sobretudo quando o léxico em questão se identifica com o *thesaurus* da língua. Assim concluo que a recolha vocabular feita por mestre Aurélio para povoar o seu dicionário, levou-o a compor uma espécie de *thesaurus* do português, embora não exaustivo, pois não cobriu todas as épocas da história do português (considerando o século XVI como ponto de partida), nem todas as variantes lingüísticas, nem tão pouco as terminologias técnicas e científicas na sua totalidade. Pode-se até especular qual seria o valor quantitativo desse *thesaurus.* Faço uma estimativa de umas 500.000 palavras-tipo (sem considerar os valores polissêmicos), ou cinco vezes o acervo total do *Aurélio,* excluídos os nomes próprios, os quais somariam números muito elevados de tipos vocabulares distintos. Lembro, porém, que Aurélio já incluiu muitos gentílicos e antropônimos, do tipo: *aliancense* (natural de Aliança, PE, Br.), *andaraiense* (natural de Andaraí, BA, Br.), *areia-branquense* (natural de Areia Branca, RN, Br.), *raul-soarense* (natural de Raul Soares, MG, Br.), *reboucense* (natural de Rebouças, PR, Br.), *recreense* (natural de Recreio, MG, Br.).

Aurélio afirma que utilizou um vasto corpus para extrair daí as abonações dos verbetes: 770 autores e 1610 obras. Tais fontes de abonação incluem escritores portugueses e brasileiros desde o século XVI até os primeiros anos da década de 1970. Vou citar alguns dentre os mais importantes para que se possa formar um juízo sobre as fontes do *Aurélio*:

1. Ficcionistas:

José de Alencar (bras. séc. XIX)
Alexandre Herculano (port. séc. XIX)
José Américo de Almeida (bras. séc. XX)
Mário de Andrade (bras. séc. XX)
Ciro dos Anjos (bras. séc. XX)
Aluízio de Azevedo (bras. séc. XIX)
Artur Azevedo (bras. séc. XIX)
Raul Brandão (port. séc. XX)
Camilo Castelo Branco (port. séc. XIX)
Euclides da Cunha (bras. séc. XX)
Osman Lins (bras. séc. XX)

Machado de Assis (bras. séc. XIX)
Fernando Namora (port. séc. XX)
Pedro Nava (bras. séc. XX)
Mário Palmério (bras. séc. XX)
Eça de Queirós (port. séc. XIX)
Graciliano Ramos (bras. séc. XX)
Ricardo Ramos (bras. séc. XX)
José Lins do Rego (bras. séc. XX)
Aquilino Ribeiro (port. séc. XX)
João Guimarães Rosa (bras. séc. XX)
Urbano Tavares Rodrigues (port. séc. XX)

2. Poetas:
Mário de Andrade (bras. séc. XX)
Oswald de Andrade (bras. séc. XX)
Luís de Camões (port. séc. XVI)
Castro Alves (bras. séc. XIX)
Gonçalves Dias (bras. séc. XIX)
Carlos Drummond de Andrade (bras. séc. XX)
Antônio Ferreira (port. séc. XVI)
Alphonsus de Guimarães (bras. séc. XIX)
Jorge de Lima (bras. séc. XX)
Fernando Pessoa (bras. séc. XX)
Antero de Quental (port. séc. XIX)
José Régio (port. séc. XX)
Fagundes Varela (bras. séc. XIX)

3. Ensaístas, historiadores, humanistas:
Alceu Amoroso Lima (bras. séc. XX)
Oscar Araripe (bras. séc. XX)
Frei Amador Arrais (port. séc. XVI)
Rui Barbosa (bras. séc. XIX)
Cleonica Berardinelli (bras. séc. XX)
Sérgio Buarque de Holanda (bras. séc. XX)
Hernani Cidade (port. séc. XX)
D. Francisco Manoel de Melo (port. séc. XVII)
Paulo Mercadante (bras. séc. XX)
João Ribeiro (bras. séc. XX)
João do Rio (bras. séc. XX)
Frei Luis de Sousa (port. séc. XVII)
Padre Antônio Vieira (port. séc. XVII)

Da extensa lista citada extraem-se algumas conclusões. Os escritores compõem um corpus relativamente homogêneo com predomínio do literário e, portanto, não se lhes pode apor a etiqueta "autores dos mais desvairados gêneros" como assevera mestre Aurélio. Além de umas tantas obras na área das ciências humanas (moral, crítica literária, história, sociologia etc.) não fazem parte do corpus publicações no domínio das ciências biológicas e exatas. Por conseguinte, os termos técnicos e científicos não devem ter sido colhidos em obras técnicas e científicas; aliás, esses termos não são abonados. Logo, o *Aurélio* é um dicionário com tendência a constituir um *thesaurus* mas não se fundamenta equitativamente em fontes escritas de todos os domínios do conheci-

mento humano. As abonações são pouco freqüentes, a despeito da riqueza das fontes, arroladas no final do dicionário. Cito ao açaso alguns exemplos de verbetes que foram abonados com citações de autores: *aparelhar* (C. Castelo Branco, Camões), *arrancar* (C. Castelo Branco, Machado de Assis, Manoel Bandeira, Júlio Ribeiro), *atenazar* (Alexandre Herculano), *atender* (Ciro dos Anjos), *bicha* (Raul Brandão), *biqueira* (Carlos Drummond de Andrade), *bisaco* (Câmara Cascudo, Osman Lins), *borla* (Machado de Assis), *botar* (José Lins do Rego), *bramar* (Camões), *carecer* (José de Alencar, Machado de Assis), *cataclismo* (João Ribeiro), *coar* (Raul Brandão), *coxia* (Eça de Queirós), *derramar* (Gonçalves Dias), *desempeno* (Euclides da Cunha), *esbaforido* (Eça de Queirós), *esboço* (Eça de Queirós), *esmolar* (Eça de Queirós), *eternizar* (Jorge de Lima), *exarcebar* (Hernani Cidade), *exprimir* (Alceu de Amoroso Lima), *exprobrar* (Rui Barbosa), *ferroar* (José de Alencar). Predominam os chamados modelos de linguagem, os clássicos, como Machado de Assis, Eça de Queirós.

Aurélio registra uma vasta fraseologia (sintagmas lexicalizados ou em vias de lexicalização), dando-lhes entrada na palavra mais importante da expressão: o substantivo em primeiro lugar, seguindo-se em ordem de importância: o verbo, o adjetivo, o pronome, o advérbio. Nesse particular seguiu o critério do *Diccionario de la Real Academia Española*, segundo ele próprio nos informa no prefácio. O autor afirma ainda que a sinonímia nos vocábulos-entrada é abundante, bem como outras informações sobre o campo léxico do verbete em epígrafe: antônimos, parônimos, o que é verdade.

Vou tentar fazer, a seguir, uma rápida crítica dos modelos de verbete do dicionário *Aurélio*. A pedra de toque de um dicionário é a definição da palavra-entrada. Muitas vezes usando os modelos clássicos de definição, Aurélio não foi muito feliz. A melhor definição é aquela que define e/ou descreve a palavra através de uma paráfrase. Tome-se, por exemplo, a definição de *amparo* = "ação ou efeito de amparar". Essa definição seria boa se ao consultar o verbete *amparar* não encontrássemos lá como definição: "1. Dar, ou servir de amparo a, estear, escorar". Aurélio incidiu na circularidade, remetendo do substantivo *(amparo)* ao verbo *(amparar)* e vice-versa, o que deixa as coisas no mesmo lugar. Ainda que tenha acrescentado os sinônimos *estear, escorar*, de fato não definiu *amparar* com uma paráfrase autônoma, de tal modo que em *amparo* pudesse remeter a *amparar*. Mencionou-se aqui apenas um exemplo, porém, esse fato se repete numerosas vezes. Por outro lado, o recurso de definir por sinôminos, muito utilizado por Aurélio, não é um bom método. Além do exemplo acima citado, cf. *amparar*: 2. proteger, patrocinar, favorecer. A forma do verbete no *Aurélio* é criticável em numerosíssimos casos. São deficientes muitas definições e há abuso do recurso à sinonimia para "definir". Muitas vezes também é contestável a ordem hierárquica das acepções de uma palavra polissêmica que possua inúmeros significados; nesses casos o dicionário dispõe inadequadamente a seqüência não ficando muito claro o critério de distribuição dos sentidos registrados e, não raro, alguns números abaixo aponta como significação diversa, um sentido que já fora incluído anteriormente (cf. *amparar, construção, estrela, linha, montar, pensar)*. Parece que algumas vezes o critério de distribuição é prioritariamente sintático e não semântico. Aliás, no prefácio, Aurélio refere que deu grande atenção à regência verbal. Por vezes não dá entradas distintas a lexemas que poderiam ser considerados homônimos como: $montar_1$ = "estar ou pôr-se sobre um animal, geralmente cavalo" X $montar_2$ = "colocar em estado de funcionamento". Também registra termos muito raros, às vezes especificamente literários, sem informar o consulente dessa conotação específica da palavra. Ex.: *aljava, aljofre, ara, bisaco, borzeguim, brial, epístola, epitáfio, ergástulo, ermita, escarmento,*

faula, fenecer, fiacre, flamante, flébil, mácula, madeixa, mansuetude, melena, nácar, nefando, noctâmbulo, novel, osedante, obséquias, odorífero, ominoso, onírico, outoniço, rebuço, recalcitrar, recôndito, sáfico, salaz, sege etc.

II. TIPOLOGIA DE OBRAS LEXICOGRÁFICAS

1. O tipo mais comum de dicionário é o "dicionário padrão da língua", ou "dicionário de uso da língua", de que seriam exemplos vários dicionários da língua portuguesa comentados na última parte do item anterior: o *Morais*, o *Aulete*, o *Cândido de Figueiredo*, o *Aurélio*. Raros são os dicionários que preenchem idealmente o tipo do dicionário padrão da língua. Esse assunto será discutido miudamente no capítulo seguinte: *O dicionário padrão da língua.*

2. O "dicionário ideológico" ou "analógico" organiza os conceitos em campos semânticos, ao invés de ordenar as palavras em ordem alfabética como os dicionários comuns. Essa tradição também é antiga; no século XVII Comenius elaborou o primeiro dicionário desse tipo. O *Thesaurus* de Roget, feito para o inglês, foi publicado pela primeira vez em 1852 pela Longman em Londres.

Um dos melhores dicionários desse tipo é o *Diccionario Ideológico de la Lengua Española* de Julio Casares (Madrid, 1942). Vejamos o que diz o próprio Casares a respeito do dicionário ideológico:

"Os dicionários ordenados com este critério têm duas partes: a primeira é a propriamente ideológica, a segunda é a alfabética, ordenada exatamente como um dicionário semasiológico. Na parte ideológica as palavras se estruturam segundo seu enquadramento em colunas básicas que correspondem à divisão do universo em categorias fundamentais.

Na parte sinótica se encontra o plano geral da classificação; no caso do *Diccionario Ideológico de la Lengua Española* a divisão do universo lexical foi estabelecida em trinta e oito classes, das quais *Deus* compõe uma classe e o *universo*, trinta e sete classes." (4, p. 439)

Pode-se argüir de arbitrária essa divisão em 38 classes conceptuais, elaborada por Casares. Talvez o progresso dos estudos de semântica e o conhecimento maior do universo poderão proporcionar melhores dados e recursos para a confecção deste tipo de dicionário no futuro.

Dentro deste modelo um dicionário contemporâneo do português muito útil, embora menos refinado, é o *Dicionário Analógico da Língua Portuguesa* (Idéias afins) de Francisco Ferreira dos Santos Azevedo, Brasília, Editora Coordenada, 1974 (?). Estrutura os conceitos e signos léxicos de acordo com o seguinte esquema classificatório: I: Relações abstratas / II: Espaço / III. Matéria. / IV: Intelecto / V: Vontade / VI: Afeições /. Esse modelo de classificação foi certamente emprestado ao *Roget's Thesaurus of English Words and Phrases Newed* prepared by Susan Mc Lloyd. London, Longman, 1982, cujo índice de temas é o seguinte: 1. "Abstract relations" (Relações abstratas) / 2. "Space" (Espaço) / 3. "Matter" (Matéria) / 4. "Intellect: the exercise of the mind" (Intelecto: o exercício da mente) / 5. "Volition" (Volição) / 6. "Emotion, religion and morality" (Emoção, religião e moralidade).

Hallig et Wartburg também elaboraram um dicionário similar para o alemão: R. Hallig & W. von Wartburg. Begriffssystem als Grundlage für die Lexicographie. Versuch eines Ordnungsschemas. Berlim, 1953. Em 1963 os autores publicaram uma nova versão reformulada e ampliada. Apresentam aí um esquema de classificação de conceitos.

Em 1981 a Longman de Londres publicou um novo dicionário ideológico para o inglês, cujo editor-chefe foi Tom Mc Arthur: *Longman Lexicon of Contemporary English*, baseado no *Thesaurus* de Roget. Os campos semânticos, ou grandes áreas de significação em que foram organizados os conceitos são os seguintes:

A. Vida e coisas vivas.
B. O corpo: suas funções e seu bem-estar.
C. Os seres humanos e a família.
D. Construções, casas, a casa, roupas, pertences pessoais, cuidados pessoais.
E. Alimentação, bebida e agricultura.
F. Sentimentos, emoções, atitudes e sensações.
G. Pensamento e comunicação, linguagem e gramática.
H. Substâncias, materiais, objetos e equipamento.
I. Arte e artesanato, ciência e tecnologia, indústria e educação.
J. Números, medidas, dinheiro e comércio.
K. Entretenimento, esporte e jogos.
L. Espaço e tempo.
M. Movimento, localização, viagem e transporte.
N. Termos gerais e abstratos.

Este tipo de dicionário que estrutura os conceitos em redes de significação pode ser questionável. É possível que o modelo de Casares só se aplique ao espanhol, o modelo do *Lexicon* só se aplique ao inglês e assim por diante. Vale dizer: cada sistema só seria válido para a língua natural para que foi elaborado. Se aceitarmos a teoria de Sapir Whorf e de outros lingüistas sobre o relativismo lingüístico, teremos que admitir que a conceptualização da realidade é típica de cada língua. Isso significaria que cada sistema lingüístico classifica os dados da realidade e da experiência de uma maneira própria, ou seja: o léxico de cada idioma categoriza o mundo e a realidade social e cultural de acordo com o seu próprio modelo classificatório. Para os partidários dos universais lingüísticos tal argumentação seria inaceitável. Talvez no meio termo se situe a verdade. Haveria áreas do conhecimento humano empírico, nomeadas no léxico de cada língua, que seriam exclusivas dessa língua e da cultura que ela expressa. Contudo, no universo cultural em expansão em que hoje vivem os homens, estaria ocorrendo uma convergência dos sistemas classificatórios, expressos por denominações lexicais. E mais ainda: na aldeia global dos meios de comunicação em que está vivendo o homem contemporâneo, intensifica-se a tendência à universalização de conceitos, sobretudo no domínio técnico-científico, fenômeno esse que pode ser bem representado por um organismo internacional para a padronização de termos: o *Infoterm* de Viena.

3. O dicionário histórico constitui uma outra modalidade. Existem vários tipos de dicionários históricos. Há um que se baseia no vocabulário e na língua de determinada época histórica. São exemplos desse tipo dos vários dicionários sobre a Idade Média que possuem algumas línguas européias. Cf. F. Godefroy — *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe. au XVe. siècle*, 1881; Boggs, Kasten, Kemiston & Richardson — *Tentative Dictionary of Medieval Spanish* (1946) e outros que descrevem o francês, o inglês, o alemão medievais. Também o *Dictionnaire de la langue française du seizième siècle* de Edmont Huguet (1946) constitui um exemplo de dicionário dedicado e um estágio anterior da língua francesa, nesse caso o século XVI. O *Dictionnaire du français classique* de Dubois, Lagaro & Lerord (1971) trata da língua do século XVII. Esse tipo de dicionário é muito útil na leitura de obras datadas das épocas históricas a que eles se consagram. Assim o dicionário de Godefroy constitui

um instrumento indispensável para ajudar a compreender e interpretar os autores franceses dos séculos IX ao XV.

Outro tipo de dicionário histórico é o pancrônico, muitas vezes rotulado de etimológico. Sendo elaborado a partir da perspectiva da língua contemporânea, ele se ocupa dos estágios anteriores do idioma, remontando à origem das palavras; tenta acompanhar a evolução histórica dos vocábulos, assinalando os diferentes valores semânticos por eles assumidos no decorrer do tempo, indicando *pari passu* as datações de cada um deles.

Um excelente exemplo é representado pelo *Diccionario crítico-etimológico de la lengua castellana* de J. Corominas (1954). Esse dicionário registra os étimos das palavras da língua espanhola contemporânea e procura seguir a evolução dos seus significados e usos no âmbito das línguas e dialetos da Península Ibérica. Assim não é apenas um dicionário histórico do espanhol, mas também de outras línguas e dialetos hispânicos tais como: catalão, português, galego, aragonês, navarro, andaluz etc.

Existem numerosos dicionários enquadrados dentro desta modalidade, tanto de línguas modernas como antigas. Alguns exemplos: A. Meillet & A. Ernout — *Dictionnaire étymologique de la langue latine* (1939); W. Meyer-Lübke — *Romanisches Etymologisches Wörterbuch* (1935); Wartbug & Baldinger — *Fränzosisches Etymologisches Wörterbuch*; J. P. Machado — *Dicionário Etimológico da Língua Portuguesa* (1953).

Um dicionário modelo dentre os históricos é o *Oxford English Dictionary* (OED). Esse dicionário cuja elaboração foi iniciada em 1857, teve sua primeira edição publicada entre 1884-1928. Constitui uma revolução na lexicografia, só superado hoje pelo *Trésor de la langue française*. O OED contém 1.800.000 citações da literatura inglesa desde as primeiras datações de cada palavra. O verbete do OED tem como principal característica a documentação rigorosa das ocorrências da palavra-entrada. As ocorrências das palavras dicionarizadas foram coletadas em 5 milhões de passagens da literatura inglesa desde as suas origens até o começo do século XX. O setor de Lexicografia da Oxford University Press, sob a direção do Dr. Burchfield, trabalha continuamente na atualização deste monumento da língua inglesa. No OED a documentação de cada uma das acepções está ordenada cronologicamente a partir do primeiro registro conhecido. Os autores do OED procuraram fornecer, no mínimo, uma documentação por século; às vezes, várias. Quando a palavra se tornou obsoleta, o OED registra o fato e documenta a última ocorrência datada.

Remeto o leitor ao item relativo à história da lexicografia francesa, feito anteriormente, onde há dados sobre o *Dictionnaire de la langue française du 19ème. et du 20ème. siècle*, que é um dicionário histórico de uma certa forma.

Quanto aos dicionários históricos da língua portuguesa, existem poucas obras a serem comentadas. Sobre o *Dicionário Etimológico da Língua Portuguesa* de Antenor Nascentes (1932), vou transcrever a abalizada opinião de Gladstone Chaves de Melo a seu respeito:

"... está baseado na edição mais antiga do *Romanisches Etymologisches Wörterbuch* de Meyer - Lübke, edição que ficou praticamente inutilizada pela terceira, ultimada em 1935; segundo: porque arrola hipóteses etimológicas, sem indicar quais as imprestáveis, quais as prováveis, quais as inteiramente aceitáveis." (5, p. 63)

José Pedro Machado publicou em 1951 um *Dicionário Etimológico da Língua Portuguesa* em dois volumes. O verbete do dicionário de Machado contém geralmente o seguinte tipo de informação: 1. origem do vocábulo/ 2. discussão do étimo e das variantes quando há controvérsia/ 3. documenação e/ou abonação das formas e sentido

dos valores semânticos anteriores à contemporaneidade/ 4. derivados, para os quais se acrescenta o mesmo tipo de informação; tais derivados aparecem como subentradas da entrada principal, sendo destacados com negritos para facilitar a sua identificação. Por exemplo: em *barro: barreiro, barrosa, barroso*; em *base: basear, basídio, basi*; em *-bata; acrobata, hidróbata.* Essa forma de agrupar os cognatos, reunindo formas derivadas da raiz portuguesa e formas oriundas da raiz grega de que deriva a portuguesa, dificulta sobremaneira o manuseio do dicionário por uma pessoa não versada em filologia e lingüística histórica. Por essa razão, Machado procurou sanar essa dificuldade incluindo no fim do segundo volume um "índice remissivo", o qual informa sobre a localização da palavra procurada, indicando a página em que ela se encontra. O dicionário de Machado contém uma série de defeitos dos quais devo assinalar dois, pelo menos: a) inclui muitos vocábulos obsoletos, desusados, raros e regionais, deixando de acrescentar palavras correntes nos tempos modernos (Cf.: *caçabe, caçapo, cacatu, cachagem, damões, darbar, datismo, emulgente, finícola, forda, frimário*); b) ao citar autores medievais e clássicos indica uma edição moderna de que se serviu como documentação (p. ex.: Fernão Lopes, *Crônica de D. João I*) sem indicar a época em que o texto foi escrito. Ora, o consulente médio não conhece as datas das obras de Fernão Lopes, nem da maioria dos autores usados como fontes documentais. Acresce que, na introdução do dicionário, não consta uma relação dos autores e edições utilizadas, nem a datação pertinente a ambos.

O mais recente dos etimológicos é o *Dicionário Etimológico Nova Fronteira da Língua Portuguesa* de Antônio Geraldo da Cunha, Rio de Janeiro, 1982. É um bom dicionário no gênero, mas sem grandes pretensões. Pautou-se pelo modelo do dicionário etimológico do francês de Bloch e Wartburg e na versão abreviada do Corominas para o espanhol. No dicionário de A. G. da Cunha o verbete indica o significado da palavra-entrada e registra as variantes históricas do vocábulo com a respectiva datação.

4. Dicionários de tipo especial.

Alguns exemplos de dicionários de tipo especial em língua portuguesa serão referidos a seguir.

Agenor Costa — *Dicionário de Sinônimos e Locuções da Língua Portuguesa* 2.ª edição, 1954, 2 volumes. Francisco Fernandes também é autor de um "dicionário de sinônimos", mas a sua melhor obra é, sem dúvida, o *Dicionário de Verbos e Regimes*, Porto Alegre, Editora Globo. Teve várias edições e reimpressões. A última que conheço é a 4.ª edição, 13.ª impressão, 1968. A primeira edição foi de 1940. Nessa quarta edição Fernandes nos fornece "mais de 11.000 verbos em suas diversas acepções e regências". Constitui um excelente trabalho para a época em que foi composto, mas contém defeitos que clamam por uma refacção da sua obra, ou a elaboração de outra similar. Muitos dos verbos incluídos na nomenclatura de Fernandes são desusados, outros, muito raros; julgo até que alguns são invenções suas. Cf. "*desfolegar*, o mesmo que resfolegar"; "*desamigar* o mesmo que inimizar"; "*insalivar, oscitar*, o mesmo que bocejar"; "*outar*, o mesmo que joeirar". Esses dois últimos verbos Fernandes provavelmente os copiou da versão que Machado fez do Morais (cf. item I.). Ora, no Morais-Machado não existe abonação dessas formas e a gente se pergunta se não seriam invenção de Machado, pois elas não existem no velho Morais (ed. de 1813). Quanto a *desfolegar*, os exemplos de Morais são do *Livro de Alveitaria* (séc. XIV) e Gil Vicente (séc. XVI) já obsoletos; deveriam ter sido, portanto, omitidos.

Está sendo elaborado um *Dicionário Gramatical do Português Contemporâneo*, sob a coordenação do Prof. Francisco da Silva Borba (UNESP, Campus de Araraqua-

ra), trabalho esse patrocinado pela FAPESP. A equipe de colaboradores do Prof. Borba são colegas seus do Departamento de Lingüística: Cacilda de Oliveira Camargo, Maria Helena de Moura Neves, Sebastião Expedito Ignácio, Odette G. L. A. Souza Campos, José Luiz Fiorin, Antônio Silveira Reis, Djalma Dezotti, Maria Celeste C. Dezotti, Elvira Wanda Vagones Mauro, Clóvis Barleta de Morais. São os seguintes os objetivos principais do dicionário: a) contribuir para a interpretação correta de uma oração e, conseqüentemente, chegar ao valor mais exato de um texto; b) contribuir para melhorar a atuação do usuário no manejo da língua escrita. São os seguintes seus objetivos suplementares: 1) mostrar como se entrosam, na dimensão pragmática da linguagem, a sintaxe e a semântica; 2) mostrar como a língua é um conjunto de possibilidades: a partir das realizadas, prever as de realização possível; 3) fornecer, pela descrição do sintagma verbal, material de consulta e de cotejo para especialistas que trabalham com fatos sintáticos do português contemporâneo do Brasil; 4) sugerir temas, enfoques e técnicas de análise sintático-semântica. Esse dicionário destina-se a todos aqueles que manejam a forma escrita da língua portuguesa como instrumento de trabalho. As abonações dos verbetes baseiam-se em textos em prosa (mais ou menos 30.000 páginas) dos últimos trinta anos, abrangendo a produção dos vários pontos do País (romances, jornais etc.). Deverá totalizar aproximadamente 7.000 verbetes quando estiver pronto. (Nota: essas informações foram fornecidas pelo Prof. J. L. Fiorin).

Dicionários especializados sobre aspectos particulares da língua portuguesa não são raros. F. da Silva Borba e Z. Jota publicaram cada um deles, dicionários de lingüística. Um clássico nessa área é do de J. Mattoso Câmara Jr. — *Dicionário de Lingüística e Gramática*, 8.ª ed. Petrópolis, Vozes, 1978. No domínio da linguagem vulgar e da gíria existem vários dicionários publicados recentemente. Cf.: Albino Lapa *Dicionário de calão*, 2.ª ed. Lisboa, Editorial Presença, 1974. Mário Souto Maior, *Dicionário do palavrão*. Recife, Guararapes, 1980. Eduardo Nobre, *O calão — Dicionário da gíria portuguesa*. Lisboa, Casa do Livro 1980. Euclides Carneiro da Silva, *Dicionário da gíria brasileira*. Rio de Janeiro, Bloch, 1973. Ariel Tacla — *Dicionário dos marginais*. Rio de Janeiro, Forense — Universitária, 1981. Manuel Viotti — *Novo dicionário da gíria brasileira*. 3.ª ed. Rio de Janeiro/São Paulo, Livraria Tupã Editora, s/d.

Finalmente vou considerar os dicionários especialmente dedicados a um domínio do conhecimento, que não a linguagem. São dicionários científicos e/ou técnicos. Nos tempos contemporâneos eles se multiplicam mais e mais, sobretudo por causa da especialização crescente que se verifica em cada ramo da ciência e da técnica.

A editora Longman de Londres está trabalhando num macroprojeto de Lexicografia, tendo como base o dicionário Webster, cujos direitos autorais foram por ela adquiridos. No banco de dados léxicos estocados na memória do computador, cada palavra está codificada com um código especial que a distingue: o código em questão pode identificá-la como integrante do léxico comum da língua, ou como termo especializado de uma determinada área do conhecimento. Com esse trabalho classificatório básico, os lexicógrafos da Longman pretendem elaborar muitos dicionários técnicos e científicos, depois da publicação da versão atualizada do Webster. Agiriam da seguinte forma: utilizando o código com que a palavra foi rotulada, pediriam ao computador todo o repertório vocabular daquela área científica ou técnica e já teriam, na saída, a listagem terminológica que comporia o cerne do dicionário, acoplada aos contextos em que as palavras dessa lista tinham ocorrido, dados esses básicos para a redação de cada um dos dicionários técnico-científicos que pretendem publicar.

A título de informação vou alistar, a seguir, uma série de dicionários especializados do tipo técnico-científico que existem em língua portuguesa. Alguns são glossários e não propriamente dicionários. Foram feitos em épocas diferentes, geralmente por especialistas da área, os quais não possuíam formação lexicográfica. Padecem por isso de muitos defeitos.

Blakiston. *Dicionário Médico Ilustrado*. S. Paulo, Organização Andrei Editora S/A, 2.ª ed., s/d.

Farmacopéia Brasileira. 3.ª ed. S. Paulo, Organização Andrei Editora S/A, 1977.

Pedro A. Pinto. *Dicionário de Termos Médicos*. Rio de Janeiro, Editora Científica, 7.ª ed., 1958 (1.ª ed. 1926 ?).

F. E. Rabello. *Nomenclatura Dermatológica*. Rio de Janeiro, Impressora Brasileira Ltda., 1974 (1.ª ed. 1950).

Carmelino Scartezzini. *Dicionário Farmacêutico*. Rio de Janeiro, Editora Científica, 1956.

Dicionário de Informática Inglês-Português. 3.ª ed. Revista, aumentada e atualizada. Rio de Janeiro. SUCESU, 1982 (?).

Dicionário Geográfico Brasileiro. 2.ª ed. Porto Alegre, Ed. Globo, 1972. Gilberto Luiz da Cruz. *Livro Verde das Plantas Medicinais e Industriais do Brasil*. (2 vols.). Belo Horizonte, Gráficas Velloso, 1965.

Rodolpho Von Hering. *Dicionário dos Animais do Brasil*. S. Paulo, Editora da Universidade de Brasília, 1968.

Existem outros tipos de dicionário ou repertórios lexicográficos. O dicionário inverso (ou grafêmico) é muito útil para o estudo dos processos de sufixação e da produtividade léxica de determinados sufixos. Por conseguinte, particularmente útil no caso de línguas de tipologia flexional bastante desenvolvida como os idiomas românicos, cuja principal fonte de criação léxica é a derivação sufixal e prefixal. Existe um bom dicionário deste tipo para o francês: A. Juilland - *Dictionnaire Inverse de la Langue Française*. (1965). Um dicionário desse tipo pode ser útil não apenas para lingüistas e especialistas em línguas. Recentemente o alto comando militar da França adquiriu muitas cópias desse "dicionário inverso" do francês de Juilland para utilizá-lo na descodificação de códigos secretos pelos serviços de inteligência. Tenho informação de que existem dois dicionários inversos do português publicados por estrangeiros e dos quais tenho apenas a referência: Dieter Messner. *Dictionnaire Inverse da la Langue Portugaise* e *Dicionário Inverso Português*, Moscou, 1973.

A cultura luso-brasileira precisa refazer muitas dessas obras para atualizá-las e aprimorá-las lexicograficamente; deve também elaborar outros dicionários relativos a áreas do conhecimento que não possuem nenhum dicionário especializado. O poder público deveria criar um órgão para ocupar-se de tão magna tarefa.

As enciclopédias são obras de referência que buscam reunir o máximo de informação sobre os mais variados domínios do conhecimento para consumo do público em geral, e não de especialistas. Podem ser do tipo alfabético ou por área do conhecimento. Em língua portuguesa existem poucas. A mais antiga é: *Dicionário Enciclopédico Salvat*. Salvat Editores S.A., Barcelona, 1955. A mais recente: *Enciclopédia Mirador Internacional*. S. Paulo/Rio. Enciclopédia Britânica do Brasil Publicações Ltda. 1976. A Editora Delta também publicou uma enciclopédia em português, baseada na Delta-Larousse e organizada por assunto.

III. O USO DO COMPUTADOR NA LEXICOGRAFIA CONTEMPORÂNEA

1. O advento do computador constituiu uma verdadeira revolução dentro da ciência da informática e da lexicografia em particular. A história dos grandes monumentos lexicográficos do passado como o dicionário francês de Littré e o *Oxford English Dictionary* testemunham o penoso sacrifício feito por lexicógrafos heróicos que compilaram, classificaram, ordenaram, organizaram e redigiram um enorme volume de dados léxicos com suas citações documentais, colhidas em centenas de milhares de passagens literárias. Ambas foram obras de várias décadas, a que os dicionaristas dedicaram parte de suas vidas, trabalhando, às vezes, até vinte horas por dia como no caso de Littré. Hoje os dicionaristas não precisam mais dar o seu sangue para elaborarem um tesouro lexicográfico porque uma grande parte do trabalho manual, monótono e estafante, pode ser feito pelo computador. Além disso, os contemporâneos contam com a vantagem de produzirem uma obra mais completa e de melhor qualidade, pois economizam sua energia com a parte repetitiva do trabalho, proporcionalmente muito volumosa no conjunto das tarefas; dessa forma poderão utilizar essa energia para a seleção do material compilado pela máquina e para a redação do texto final, que constitui a etapa mais importante de qualquer obra lexicográfica. O computador é particularmente útil na confecção de dicionários históricos, ou de quaisquer dicionários que pretendam documentar a informação fornecida em cada entrada. O testemunho de um lexicógrafo (A. J. Aitken, Edinburgh) que está trabalhando com um monumento da antiga língua escocesa (*Dictionary of the Older Scottish Tongue*, corpus total: 1 bilhão de palavras; corpus selecionado para o dicionário: 200 milhões de palavras) parece-me vir muito a propósito:

"Este artigo pretende desenvolver dois argumentos básicos do ponto de vista daquele que pratica lexicografia histórica. Em primeiro lugar, como a capacidade do computador para contar ocorrências de palavras em extensas passagens de texto pode possibilitar ao planejador de um dicionário histórico um controle muito mais preciso do que sucedeu até hoje, sobre o tamanho e a forma de sua coleção de citações documentais, possibilitando, portanto, o controle da magnitude da tarefa que ele está executando. Em segundo lugar, a manipulação posterior do arquivo do seu dicionário pode beneficiar-se das facilidades oferecidas pelo computador para armazenamento de dados, possibilitando o confronto desse material e a sua seleção; e mais ainda: o acesso flexível aos dados e a sua imediata recuperação bem como a facilidade e a conveniência de inserir revisões e suplementações." (1, p. 29)

2. O léxico constitui um conjunto aberto em qualquer sistema lingüístico e, por conseguinte, não apenas vastíssimo quando comparado com outros setores e níveis da língua (fonológico, morfológico, sintático) mas também em permanente expansão sobretudo numa língua de civilização. Por essa razão, o quantitativo é uma das propriedades fundamentais do vocabulário. Como a confecção de dicionários é tarefa em que manipulam gigantescos volumes de dados, justifica-se sobremaneira o uso de computadores para a execução de um empreendimento lexicográfico. De fato, o computador é particularmente apropriado e eficaz quando se trabalha com grande quantidade de dados. Inclusive financeiramente o seu uso torna-se recomendável. Vamos dar um exemplo que ilustra as afirmações acima.

Em 1969-1971 B. Richman, J. Carrol & P. Davies elaboraram um trabalho de léxico-estatística — *American Heritage Word Frequency Book* (AHWFB) — a fim de gerar dados léxicos que seriam utilizados posteriormente na confecção de dicionários

pela American Heritage Publishing Company. A partir dos resultados do AHWFB e de outras obras lexicológicas foi criado o *Children's Dictionary* (1979).

O *American Heritage Word Frequency Book* (1971) manipulou um corpus de 5.088.721 palavras, retiradas de 1.075 livros e revistas, dos quais se selecionaram amostras de 500 palavras (tamanho da amostra) em cadeia. Esse grande arsenal de dados léxicos forneceu apenas 86.741 palavras diferentes; chamemo-las de "tipo" (wordtypes). Isso já mostra a fantástica proporção de redundância e repetição nos discursos lingüísticos.

Veja-se o quadro seguinte com alguns dados estatísticos reveladores:

AHWFB	CORPUS TOTAL: 5.088.721		
freqüência acumulada de: *the* + *of* + *and*	happax legomena (tipos que ocorreram uma só vez)	palavras de 2 a 5 ocorrências	palavras de freqüência 1 a 20
653.023	35.079 tipos diferentes	25.358 tipos diferentes	74.686 tipos diferentes

opor aos 86.741 tipos do AHWFB

É impressionante nesta como em qualquer computação das mais variadas línguas, como um número muito pequeno de palavras (instrumentais lingüísticos, palavras-gramaticais) têm uma altíssima freqüência, contra um enorme número de palavras plenas com baixíssima freqüência. No exemplo acima tal fato é típico: três lexemas somam 653.023 ocorrências, enquanto 35.079 (os happax legomena) apenas 35.079 ocorrências. Também é muito pequena, estatisticamente desprezível num corpus de 5.088.721 palavras, o concurso das palavras de 2 a 5 ocorrências com os seus 25.358 tipos diferentes (compare-se com os 86.741 tipos diferentes de todo o corpus!). É por isso que muitos trabalhos de orientação estatística, especialmente os dicionários de freqüência, desprezam as palavras com freqüência abaixo de 5. Quando o corpus é grande, como é o caso deste AHWFB, mesmo as freqüências abaixo de 20 passam a ser irrelevantes. Note-se que elas concorreram com 74.686 tipos diferentes de palavras. Sobram do total de 86.741: 12.055 palavras diferentes, que seriam as palavras dignas de consideração. Esses vocábulos são relevantes tanto para a elaboração de livros pedagógicos como para dicionários de uso geral na língua como o *Children's Dictionary* anteriormente citado. Mais um dado estatístico importante: os 5.000 vocábulos mais freqüentes do AHWFB ocorreram 77 vezes (ou mais) no corpus num total de 4.547.336 ocorrências (de 5.088.721!). Com base nos resultados deste trabalho poderíamos afirmar com Howard H. Keller:

"A disparidade percentual é surpreendente — se um aluno possui um vocabulário de 5.000 palavras, ele conhecerá apenas 5,77% do total de lexemas da língua, mas essas 5.000 palavras constituirão 89% do vocabulário de qualquer texto!" (3, p. 177)

Os autores do AHWFB propuseram um novo parâmetro: FPM (frequency-per-million), ou freqüência-por milhão, para compensar as distorções criadas pela dispersão. (Nota: a dispersão é uma resultante da variedade dos gêneros literários, dos temas tratados, dos níveis de linguagem, dos estilos.). O índice FPM fornece, pois, um fator de reajuste da freqüência à dispersão.

"Os autores acreditam que essa FPM reflete melhor a freqüência verdadeira de uma forma que seria encontrada em um corpus de qualquer tamanho, dos pequenos aos infinitamente grandes." (3, p. 177)

A FPM constitui uma informação importante de um dicionário de freqüência. Ela acrescenta uma terceira dimensão à palavra, mostrando o valor ou a importância relativa dessa palavra no conjunto total do léxico, o que falta em um dicionário comum.

Não só os autores do *Children's Dictionary* (Houghton Mifflin, 1979) se serviram dos dados estatístico-vocabulares produzidos pelo AHWFB. O lexicógrafo A. J. Augarde, da Oxford University Press, elaborou o *Oxford Intermediate Dictionary* (12.000 verbetes) usando o AHWFB como referência básica. Selecionou as suas 12.000 palavras-entrada a partir da lista de palavras em ordem decrescente de freqüência do AHWFB. Estabeleceu o número de ordem 12.000 como parâmetro e como limiar inferior o vocábulo de ordem 10.000 e como limiar superior, o de número 20.000. Usando sua competência lingüística e seu bom senso, eliminou algumas palavras situadas entre a freqüência 10.000 e 12.000 e substitui-as por outras palavras situadas entre a freqüência 12.000 e 20.000, que julgou mais necessárias para um aluno da escola secundária.

No *Frequency Dictionary of Portuguese Words* (Duncan, Ph. D. Dissertation, 1972) o índice de dispersão representa a distribuição irregular de uma palavra nos cinco subcorpus que constituíram o corpus total: 1.° literatura dramática; 2.° romances e contos; 3.° ensaios; 4.° jornais e revistas; 5.° textos técnicos e científicos. Evidentemente a ocorrência de cada vocábulo (excetuados os gramaticais) é diferente conforme o tema tratado e o tipo de linguagem. Um exemplo típico: no corpus do FDPW a palavra *bonito* ocorreu 14 vezes em peças, 6 vezes em romance e contos, 6 vezes em ensaios e apenas 1 vez nos textos jornalísticos, técnicos e científicos. Esse fato mostra que essa palavra é característica de um tipo de linguagem, mas irrelevante em outros gêneros literários e registros lingüísticos. O seu índice de dispersão foi de 45,88%. Uma palavra com dispersão uniforme através de todos os gêneros seria, p. ex., o artigo *o*.

O FDPW propõe um parâmetro semelhante à FPM mencionada acima, isto é, o coeficiente de *uso*. Esse coeficiente reflete o uso de um lexema nas realizações lingüísticas. O coeficiente de uso é função da freqüência e da dispersão. A título de exemplo: segundo os dados do FDPW o coeficiente de uso de *bonito* é 19,73 — o do artigo *o* é 23.758,10. A enorme disparidade entre esses dois números evidencia claramente a importância desse parâmetro como propriedade fundamental das palavras dentro do léxico da língua.

3. Evidenciada a dimensão quantitativa do léxico, vejamos quais os materiais de maior utilidade elaborados pelo computador a serem usados numa empresa lexicográfica. São eles: as listas de freqüência de palavras, acrescidas de alguns parâmetros de especial interesse (dispersão ou distribuição, uso, FPM) e as concordâncias.

As listas de freqüência de palavras ou *índices verborum* são basicamente de dois tipos: a) lista em ordem decrescente de freqüência; b) lista das palavras em ordem alfabética com a indicação de seus respectivos parâmetros. Na primeira lista as palavras são ordenadas hierarquicamente, partindo-se da palavra mais freqüente até aquelas que ocorreram uma só vez no corpus (os happax legomena). Os comentários acima feitos sobre o AHWFB evidenciam a existência de blocos com características típicas dentro dessas listas. Em primeiro lugar o bloco das palavras de alta freqüência, geralmente palavras instrumentais e umas poucas palavras lexicais de significação plena (alguns verbos, substantivos e adjetivos). No caso do português as 100 palavras mais freqüentes incluem:

4 artigos: a, o, um, uma;

12 adjetivos: algum, aquele, esse, este, grande, mesmo, meu, novo, outra, português, seu, todo;
17 advérbios: agora, ainda, assim, bem, depois, então, hoje, já, mais, menos, muito, não, onde, sempre, só, também, tão;
conjunções: como, mas, nem, ou, quando, se;
numerais: dois;
preposições: a, até, com, de, em, entre, para, per, por, sem, sobre;
pronomes: ela, elas, ele, eles, eu, isso, nos,*o que, qual, que (relativo), se (passivo), tudo;
substantivos: ano, casa, coisa, dia, estudo, homem, mulher, pai, parte, tempo, vez, vida;
verbos: chegar, dar, deixar, dever, dizer, encontrar, estar, falar, ficar, haver, haver (impessoal), ir, parecer, passar, poder, querer, saber, ser, ter, ver, vir.
Note-se que essas palavras constituíram 61,98% do corpus total!

As 10 palavras mais freqüentes do FDPW na ordem decrescente de freqüência foram:

1.° o (art.)	4.° que (pn. rel.)	8.° a (prep.)
2.° a (art.)	5.° ele (pn. pes.)	9.° que (conj.)
3.° de (prep.)	6.° ser (v)	10.° em (prep.)
	7.° ela (pn. pes.)	

No outro extremo da lista temos os happax legomena, na sua maioria substantivos. Aliás, todos os dicionaristas e estatísticos léxicos sabem que o substantivo é a classe de palavra mais importante. Em todas as línguas estudadas o substantivo ultrapassa de muito numericamente todas as demais categorias sendo, portanto, o tipo de vocábulo mais significativo na composição do léxico. Também do ponto de vista lexicográfico, o substantivo é a pedra de toque na definição de qualquer palavra. As baixas freqüências (2 a 5 ocorrências) são povoadas de substantivos e palavras de significação plena: *aperfeiçoamento, assentimento, absolver, atormentar, afetivo, aliviado.* Há muitos advérbios de modo entre essas palavras: *consideravalmente, positivamente, sensivelmente, sucessivamente* (dados do FDPW). Ora, o advérbio de modo em *-mente* tem um status gramatical curioso. Trata-se de uma palavra de significação lexical por ser formado de adjetivo qualificativo + sufixo *-mente*; e tem um valor gramatical bem distinto de outras palavras por não ter relação de dependência com outras formas léxicas, pois tem colocação praticamente livre na cadeia do discurso.

Os *indices verborum* têm interesse lexicográfico sobretudo quando acompanhados de parâmetros como a dispersão (ou distribuição), o uso, e a FPM. De fato, não basta saber que o verbo *poder* foi utilizado 1.282 vezes num total de 500.000 ocorrências; ou que a preposição *de* foi utilizada 25.313 vezes no mesmo número total de ocorrências. O índice de uso dessas duas palavras aponta claramente a sua importância na língua portuguesa. Para a preposição *de* o coeficiente de uso foi de 21.688,13 e para o verbo *poder*, 1.084,70, relativamente aos dados do corpus do FDPW. Compare-se com o mesmo índice para *bonito* — 19,73 — ou de substantivos como *pensamento* (coeficiente de uso: 54,55; freqüência: 85), *pesquisa* (uso: 3,61; freqüência: 3), *povo* (uso: 85,28; freqüência: 153).

ET ARMÉ DE LA MASSUE, IL SA ÉLANCE FIEREMENT M9986037 RA1104 UNIVERSELLE DU MONDE ! IL EST REVETU DE LA PEAU DU LION
ET PARTOUT L' EMPREINTE DE SA GRIFFE, O POUR K2324254 RA1105QEN DEPIT DE SA FORCE ET DE SA MASSE, CV EST LE PAS OU LION
ET L'ANTILOPE D' *HENRI *ROUSSEUAU, DEUX TO K7654674 RA1106 , DONT LES GRAVURES FONT UN EFFET GROTESQUE, LE LION
ET LES LOUPS VOIENT *MILON SAISI PAR SA VICTIM M8991159 RA1107SESDEUXFLANCSCOMMEDESTENAILLESINFLEXIBLES,-LESLIONS
ET LES A TUÉS DE SA MAIN, IL A VU UN GRAND CH M8991158 RA1108 . *MILON *DE *CROTONE, - MILON A JOUJE AVEC LES LIONS

Em 1984 o Instituto Nacional de Investigação Científica de Lisboa publicou o *Português Fundamental* (vol. I, Vocabulário), pesquisa sobre o vocabulário básico do português desenvolvida pelo Centro de Lingüística da Universidade de Lisboa desde 1970, inicialmente sob a direção do Prof. Lindley Cintra e depois do Prof. João Málaca Casteleiro. Esse projeto tinha por objetivo estabelecer o léxico básico para o ensino do português a estrangeiros. Utilizou-se o modelo de pesquisa do Francês Fundamental e do Espanhol Fundamental. O corpus que originou o vocabulário de *Português Fundamental* continha 700.000 ocorrências de palavras, coletadas em entrevistas realizadas com 1.400 informantes. A esse acervo de registros do discurso oral, somaram-se os inquéritos escritos, que complementaram os dados colhidos na primeira fase da pesquisa. Os dados lingüísticos foram tratados com um rigoroso modelo de análise de Estatística Léxica, gerando listas de freqüências de palavras. Dessas listas foram selecionadas 2.217 palavras de uso comum, consideradas como o vocabulário básico do português na comunicação oral. Esse trabalho léxico-estatístico constitui uma das tarefas lexicológicas que precisavam ser feitas sobre o nosso idioma. A análise da lista dessas 2.217 palavras mostra que não há coincidência total entre o vocabulário básico do português de Portugal e da variante brasileira. Várias palavras precisariam ser acrescentadas para servir à comunicação falada no Brasil, assim como várias outras poderiam ser eliminadas. Tais diferenças são facilmente explicáveis pela Sociolingüística, em virtude das variações sociais e geográficas da língua nos dois países — Portugal e Brasil.

As concordâncias de textos (KWIC ou Key-Word-In-Context) são listas dos contextos em que ocorreu uma determinada palavra (a palavra-chave) posta em evidência, geralmente à esquerda da saída (print-out) dos dados impressos pelo computador. Por exemplo: *corpo* Carlos é um rapaz forte: tem ★ ★ corpo ★ ★ de atleta.
O ★ ★ corpo ★ ★ do prefeito está sendo velado na prefeitura.
O ★ ★ corpo ★ ★ da narrativa é constituído por uma trama policial.
No ★ ★ corpo ★ ★ central da casa havia um átrio.
Normalmente antecedem (ou seguem) os dados em código, os quais permitem localizar a passagem, o texto, e a obra de onde foi extraída a citação.

Na elaboração do *Trésor de la langue française* foram produzidas dois tipos de concordância:

1.°) um microcontexto de uma linha (18 a 20 palavras) em que está inserida a palavra-chave:

M40138, "Et, Dérobant L' Éclair a L'Inconnu Sublime," Lier Ce Char d' Un Autre a des Chevaux à Toi? ou M2780325
M40139ET Ligne Directe Ascendante, La Première Est Celle Qui Lie Le Chef Avec Ceux Qui Descendent de Lui; La D M4870021
M40140 Sur La Peau, Puis, La Douleur Changeait: On Lui Liait La Cheville Avec Un Fil de Fer, On Lui Raidis L4f30171
M40141 Veines Qui Tremblent Sur La Main Comme Les Cordes Qui Lient Un Chevreau; — Père *Janet, Qui Est — Ce Qu Kf270001

2.°) um macrocontexto de 8 linhas onde ocorreu a palavra-chave:

L679 Maritain Humanisme Intégral 1936
1 Par l'économique. Et c'est bien plutôt la mise de
2 Côté, le rejet dédaigneux de toute idéologie
3 Métaphysique comme expression on reflet transitoire
4 D'un moment économique, qui, en un sens, et malgré
5 L'insistance avec laquelle le marxisme populaire

6 Exploite ce thème, est une apparence théorique
7 Illusoire, ou, comme les arguments des vieux
8 Sceptiques grecs, un thème drastique destiné à

Thomas M. Paikeday desenvolveu um programa para microcomputador que produz concordâncias de dois tipos:

1.°) microcontexto de 128 caracteres (média de 22 palavras):

The New York Times Everyday Dictionary, TEXT ANALYSIS BY MICROCOMPUTER, Page 2

had the flu, not toxic shock, and that no definitive link *between* Rely and the ailment had ever been established. Kidneigh arg s of automobile emissions. "There is a positive association *between* the removal of lead from gasoline and the lower blood-lead l swallowing air in a person with that habit. Hold an eraser *between* your teeth *between* meals to help prevent the swallowing mech.

2.°) contextos de 14 linhas:

```
My strength, and then I would  go out to put some hurt on the pe
a place in orbit where you can  go out for a good time on a Satur
For example, if classified ads  go electronic, make sure they're
with only three more states to  go, Carter let the movement stumb
ibutions, How many politicians  go around voting against churches
on auto loans. And so it would  go, through mayors supporting the
r turn to take a vacation. Now  go. No whimpering. Not another wo
alysis has some merit, I would  go further. I really wonder wheth
s he boarded his helicopter to  go to Camp David shortly afterwar
return for a commitment not to  go the plutonium route has never
omeuppance, and now things can  go back to normal. There could no
prisons. The other group would  go easy on the criminal and try t
ir, and leave him limp while I  go call the police. Would our pre
, but that doesn't mean he can  go around shooting people. The ps
.....
Isaac A. Levi (AP, Mexico City): Elections, Col. Disp., 5 Jul 82
```

Os programas ALPHA e GLOBAL que geraram essas concordâncias (as do inglês) têm a grande vantagem de não precisarem de um computador de grande porte como aquele utilizado em Nancy, França (as concordâncias do francês acima transcritas) para o *Trésor de la langue française*, e rodarem em um microcomputador, o que coloca ao alcance de mais indivíduos e entidades a possibilidade de gerar concordâncias para trabalhos lexicográficos. De fato, é possível executar essas obras de referência básica em lexicografia como *índices verborum* e concordâncias dispondo de equipamento menos sofisticado que aqueles do *Institut de la Langue Française*. Infelizmente os programas criados por Paikeday para o inglês precisariam ser adaptados para a língua portuguesa, a fim de servirem ao tratamento automático do português e isso não é uma empresa fácil.

As concordâncias de texto são um manancial riquíssimo para documentar e ilustrar usos semânticos e gramaticais e atestar o que está ocorrendo de fato na língua, quando se trata de um trabalho sobre Lexicografia contemporânea. Um dos mais importantes serviços que uma concordância pode prestar é relativa à sintagmática. Fornecendo-se em bloco seqüencial as ocorrências de uma palavra em seus contextos,

ela fornece ao lingüista, lexicógrafo, ou gramático, o conjunto das combinações e das colocações que a palavra em epígrafe pode ter. Para serem perfeitas as concordâncias exigiriam uma lematização, trabalho de alta sofisticação para um computador, assunto esse que discutiremos mais abaixo.

4. Uma das empresas mais arrojadas no domínio da automação da lexicografia é o "dicionário de máquina". Um trabalho pioneiro dessa natureza foi o executado por uma equipe de lingüistas e especialistas em computação e matemática do *Istituto di Linguistica Computazionale* de Pisa, equipe essa liderada por G. Ferrari e I. Prodanof. A pesquisa ainda prossegue atualizando-se e aprimorando-se os resultados.

A concepção original do "Dicionário de máquina do Italiano" (DMI) teve como objetivo criar um instrumento básico para lematização automática de textos fornecidos ao computador para análise. Ferrari se baseou num dicionário comum de uso do italiano: N: Zingarelli — *Vocabolario della Lingua Italiana*. A partir desse banco de dados léxicos sobre o italiano contemporâneo, os pesquisadores elaboraram uma lista de formas lematizadas do italiano (forma canônica) de qualquer paradigma ou conjunto das variantes flexionais. Para fazer a máquina operar foi criada uma série de algoritmos, capazes de reconhecer as formas flexionadas e remetê-las à forma lematizada, ou lema.

Para se ter uma idéia da complexidade dessas operações, basta lembrar que o DMI contém 912.618 formas geradas automaticamente (lemas), a partir de 106.152 entradas léxicas. Um dos refinamentos que está sendo elaborado, diz respeito ao difícil problema dos homógrafos no tratamento automático. Os pesquisadores estão tentando reduzir a extensão do dicionário e tornar mais econômico o "look-up", isto é, a busca das formas em todo o dicionário. Para isso tentam unificar certas classes de homógrafos. Convém lembrar que a homografia constitui um dos mais desafiadores problemas para a lingüística computacional. A propósito: até esta data a distinção entre homógrafos ainda está sendo feita manualmente. Os pesquisadores estão tentando refinar o seu analisador sintático-semântico para poder automatizar esta operação ao menos parcialmente.

O DMI tem operado satisfatoriamente no reconhecimento das formas lingüísticas dos textos processados com o seguinte rendimento porcentual: 86% em textos de jornais, chegando a 96% em textos literários. No caso de textos jornalísticos os maiores embaraços são criados por nomes (geográficos e onomásticos) que, naturalmente, constituem pequena parcela do DMI. E em ambos os casos os homógrafos criam os maiores impasses na análise e identificação automáticas. Por enquanto, os pesquisadores analisam os homógrafos manualmente, a despeito da lentidão dessa tarefa.

O DMI é um arquivo, um tesouro léxico da língua italiana contemporânea, estocado na memória do computador para ser consultado pela máquina quando ela for analisar qualquer texto automaticamente. Por isso a equipe de pesquisadores continua trabalhando na atualização dos dados estocados e no refinamento dos algoritmos que operam a identificação das formas. Esse refinamento diz respeito às tentativas de eliminação de ambigüidades quando o computador propõe mais do que uma lematização. Por isso se trabalha na criação de melhores analisadores sintático-semânticos. O conjunto de operações realizadas pela máquina constitui "uma tentativa de produzir um modelo computacional unívoco do léxico dentro de um modelo de língua natural" no dizer de G. Ferrari. Esse lingüista e sua equipe acreditam que "embora cada realização lingüística se inclua no nível do desempenho, existe um nível de competência para o léxico, que justifica nossa proposta em favor de uma concepção unitária". (2, p. 8)

Por tudo o que foi dito acima o DMI não é uma simples lista de unidades léxicas; e "deve ser considerado como uma máquina composta de um núcleo de dados e de uma

série de procedimentos que operam em vários níveis lingüísticos. Em outras palavras, cada nível da análise exige a criação de um novo modelo lingüístico a ser associado ao dicionário e a ser executado em conexão com ele. Neste sentido, o dicionário está situado entre a lexicologia, a morfologia, a sintaxe e a semântica. Se essa posição em relação às duas primeiras é relativamente clara, entre as últimas a fronteira não é precisa, sobretudo com relação ao léxico". (2, p. 9)

Várias outras tentativas estão sendo feitas contemporaneamente, mais ou menos similares ao DMI, em outros centros de computação lingüística, mas com objetivos diferentes e concepção original bem distinta. Em Grenoble (grupo GETA) do Centro de Tradução Automática, vários transcodificadores estão em operação (e sendo aprimorados) com vistas à tradução automática das várias línguas da comunidade européia. Na equipe do Prof. Vauquois em Grenoble, está trabalhando nosso colega Paltônio Daun Fraga no "autômato de estados finitos não determinados" para a análise automática do português e sua tradução em inglês.

O uso do computador em Lexicografia está revolucionando essa ciência secular. Atualmente grandes editores de dicionários e obras de referência usam o computador como instrumento básico nos trabalhos de coleta, seleção, armazenamento e recuperação de dados. A rigor, hoje não se pode planejar um dicionário-padrão da língua de porte médio, ou seja, 50.000 verbetes, sem o concurso do computador. Essa máquina pode realizar operações fundamentais na confecção de um dicionário com eficiência muito superior à do homem, a saber: 1) seleção das palavras-entrada para a composição do repertório vocabular do dicionário com base em critérios de Estatística Léxica; 2) controle das definições do dicionário, usando modelos uniformes e um vocabulário básico na formulação dos verbetes; 3) controle rigoroso das referências cruzadas; 4) concordâncias de palavras das quais os lexicógrafos extrairão a documentação dos significados e usos registrados no dicionário. Por maioria de razão, as obras de referência de grande porte não podem dispensar a tecnologia e os recursos científicos contemporâneos, pois os grandes dicionários devem partir de corpora de 30 milhões a 100 milhões de ocorrências de palavras, coligidos em uma gama muito variada de textos: literários, jornalísticos, técnicos, científicos etc. Além disso, existem outros recursos técnicos muito úteis em Lexicografia: a) leitura ótica; b) possibilidade de seleção de textos, correção e inserção de informação diretamente através do vídeo; c) classificação e ordenação automática dos contextos que serão utilizados, contextos esses de fácil recuperação; d) enorme potencial de armazenamento de dados com rápido acesso não-seqüencial de um bloco de dados através do uso de discos; e ainda vários meios técnicos de tipo documental que hoje facilitam imensamente a penosa tarefa do lexicógrafo.

Creio que comprovamos suficientemente duas coisas: de um lado, o formidável instrumental que hoje dispomos para a confecção de dicionários e de obras de referência, especialmente o computador; de outro, que a máquina não pode dispensar o concurso do lingüista e do lexicógrafo. Ela chega até um certo ponto e até esse marco trabalha melhor e mais rapidamente que o homem; porém, quando o computador atinge seus limites, o lexicógrafo entra com a sua competência lingüística e com a sua inteligência operadora. Assim, ambos — homem e máquina — podem trabalhar como colaboradores para um produto final de melhor qualidade. Infelizmente, porém, o computador e todos os seus periféricos produzem material numa velocidade muito maior do que uma equipe de lexicógrafos pode digerir, para redigir os verbetes de um dicionário. Provavelmente ainda se passará muito tempo antes que lexicógrafos e analistas de sistemas consigam atingir o estágio ideal em que se criará um analisador automático de uma linguagem natural, capaz de esmiuçar um texto e distinguir como um lingüista compe-

tente as nuanças sintático-semânticas do discurso produzido pelos homens. Talvez esse estágio de perfeito entrosamento entre o homem e a máquina não seja jamais atingido. De qualquer forma, a era dos computadores gerou uma autêntica revolução na Lexicografia. Esperemos que a moderna tecnologia enseje a produção de dicionários cada vez mais perfeitos sem que o lexicógrafo se transforme num obscuro mártir da ciência que pratica.

BIDERMAN, M.T.C. — The science of Lexicography. **Alfa**, São Paulo, **28**(supl.):1-26, 1984.

ABSTRACT: Three aspects are discussed: 1) a short history of Lexicography in Spanish, French and Portuguese. The most important dictionaries of these languages are evaluated, from the XVIIth to the XXth century. 2) Typology of lexicographical works. The main types of existing dictionaries in Latin languages and English are commented on. 3) Use of computers in modern Lexicography. The computer caused a revolution in Lexicography owing to its capacity to perform basic and tedious tasks such as: compiling, classifying and ordering lexical and contextual data for the organization of dictionaries. In addition, they provide rapid retrieval facilities.

KEY-WORDS: Lexicography; thesaurus; dictionary; historical dictionary; etymological dictionary; conceptual dictionary; technical dictionary; scientific dictionary; frequency dictionary; lexical data bank; corpus; indices verborum; concordance; machine dictionary.

REFERÊNCIAS BIBLIOGRÁFICAS

1. AITKEN, A.J. — Historical dictionairies, word frequency, distributions and the computer. *Cahiers de Lexicologie, 32*(1): 28-47, 1978.
2. FERRARI, G. & PRODANOF, I. — Machine dictionary and lexicon. *IN*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Otawa, 1976 (Comunicação)
3. KELLER, H.H. — The American heritage word frequency book (review). *Language Learning, 25*: 173-178, 1975.
4. MARCOS-MARÍN, F. — *Curso de gramática española*. Madrid, Cincel-Kapelsz, 1980.
5. MATORÉ, G. — *Histoire des dictionnaires français*. Paris, Larousse, 1968.
6. MELO, G.C. de — *Dicionários portugueses*. Rio de Janeiro, S.O.M.E.S., 1947.
7. MORAIS SILVA, A. de — *Diccionário da língua portuguesa*. — Fac-simile da segunda edição, 1813, photographada pela Revista de Lingua Portuguesa. Rio de Janeiro, Oficinas de S.A. Littro Typographia Fluminense, 1922. 2v.

BIBLIOGRAFIA CONSULTADA

BALDINGER, K. — Semasiologia e onomasiologia. *Alfa, 9*: 7-36, 1966.

BARNHART, C.L. — Plan for a Central Archive for Lexicography in English. *Annals of the New York Academy of Sciences, 211*: 302-319, 1973.

BIDERMAN, M.T.C. — *Teoria lingüística*. Rio de Janeiro, Livros Técnicos e Científicos, 1978.

BLUTEAU, R., Pe. — *Vocabulário português e latin*. Lisboa, Colégio das Artes da Cia. de Jesus, 1712-1728. 8v., 2 supl.

CAPPELLI, A *et alii* — *Parsing an Italian text with an ATN Parser*. Pisa, Instituto di Lingüistica Computazionale, 1978.

CASARES, J. — *Introdución a la lexicografía moderna*. Madrid, C.S.I.C., 1950.

COSERIU, E. — Vers une typologie des champs lexicaux. *Cahiers de Lexicologie, 27*(2): 30-51, 1975.

DELATTE, L. *et alii* — Le traitement automatique de la langue française au laboratoire d'analyse statistique des langues anciennes. *Revue des Études Anciennes, 4*:1-55, 1977.

DICCIONARIO DE LA LENGUA ESPAÑOLA /por/ Real Academia Española. 18.ed. Madrid, Espasa-Calpe, 1956.

DRETTAS, G. — Les théoriciens allemands du champ. *La Linguistique, 17*(2): 3-22, 1981.

DUBOIS, J. — *Introduction à la lexicographie: le dictionnaire*. Paris, Larousse, 1971.

DUNCAN, J. — *Frequency dictionary of Portuguese words*. Stanford, Stanford University, 1972. (Ph. D. Dissertation)

FERRARI, G. — Dictionnaire automatique et dictionnaire-machine: une hypothèse. *In*: COMPUTATIONAL AND MATHEMATICAL LINGUISTICS. PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Pisa, 1973. Firenze, Leo S. Olschki Ed., 1977. p. 257-262.

FERREIRA, A.B. de H. — *Novo dicionário da língua portuguesa*. Rio de Janeiro, Nova Fronteira, 1975.

LARA, L. F. — Méthode en lexicographie: valeur et modalité du dictionnaire de machine. *Cahiers de Lexicologie, 29*(2): 103-28, 1976.

LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH. London, Longman Group, 1978.

MC NAUGHT, J. — Terminological data banks: a model for a British linguistics data bank (LDB). *ASLIB Proceedings, 33*(7/8): 297-308, 1981.

MACHADO, J.P. — *Dicionário etimológico da língua portuguesa*. Lisboa, Ed. Confluência, 1956. 2v.

MOLINER, M. — *Diccionario de uso del español*. Madrid, Gregos, 1975.

MORAIS SILVA, A. de — *Grande dicionário da língua portuguesa*. 10. ed. rev. por J.P. Machado. Lisboa, Ed. Confluência, 1949-1957. 12v.

MURAKAWA, C. de A.A. — *O primeiro dicionário de língua portuguesa de Antonio de Morais Silva. Estudo crítico da edição de 1813*. Araraquara, Instituto de Letras, Ciências Sociais e Educação, Unesp, 1984. (Dissertação de Mestrado)

NASCENTE, A. — *Dicionário etimológico da língua portuguesa*. 2a. tiragem da 1.ª ed. Rio de Janeiro, Francisco Alves, s.d.

PAIKEDAY, T.M. — Language analysis and lexicography by microcomputer. (Comunicação feita no encontro ADS-MLA, 1981).

PAIKEDAY, T.M. — *The New York Times everyday dictionary*. New York, Times Book, 1982.

PORTUGUÊS FUNDAMENTAL. Lisboa, Instituto Nacional de Investigação Cientifica, Centro de Lingüística da Universidade de Lisboa, 1984. v.1, Vocabulário e Gramática; T. 1, Vocabulário.

PRODANOF, I. — A la recherche d'un modèle de derivation en italien. *In*: COMPUTATIONAL AND MATHEMATICAL LINGUISTIC. PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Pisa, 1973. Firenze, Leo S. Olscki Ed., 1977. p. 297-301.

RAVEN, I. *et alii* — Lexicography in English. *Annals of the New York Academy of Sciences, 211*, 1973.

REY-DEBOVE, J. — Le domaine du dictionnaire. *Langages*, (19):3-34, spt., 1970.

REY-DEBOVE, J. — Lexique et dictionnaire. *In*: LE LANGAGE: Les dictionnaires du savoir moderne. Paris, Centre d'Étude et de Promotion de la Lecture, 1973. p. 82-108.

RICHMAN, B. *et alii* — *The American heritage word frequency book*. New York, Boston American Heritage Publ., Houghton Mifflin, 1971.

ROBERT, O. — *Dictionnaire alphabétique et analogique de la langue française*. Paris, SNL, Dictionnaire Le Robert, 1972.

SHERMAN, D. — Retrieving lexicography citations from a Computer Archive of Language Materials. *Annals of the New York Academy of Sciences, 211*:137-142, 1973.

SHERMAN, D. — Special purpose dictionnaires. *Cahiers de lexicologie, 32*(1): 82-102, 1978.

TERMINO GRAMME: Bulletin de la Direction de la terminologie. Québec, Office de la Langue Française, 1979.

THE OXFORD ENGLISH DICTIONARY. Oxford, Claredon Press, 1933.

VENEZKY, R.L. — Computer applications in lexicography. *Annals of the New York Academy of Sciences. 211*:287-291, 1973.

VENEZKY, R.L. — Storage, retrieval and editing of information for a dictionary. *American Documentation, 19*:71-79, 1968.

VIEIRA, D. Frei — *Grande diccionario portuguez ou thesouro da lingua portuguesa*. Porto, Ed. Ernesto Chardon e Bartholomeu H. de Moraes, 1871.

VITERBO, J. de S.R. de, Frei — *Elucidario de palavras e frases que em Portugal antigamente se usarão* (sic) *e que hoje regularmente se ignorarão*. Lisboa, 1978-1979.

WEINREICH, U. — Lexicographic definition in descriptive semantics. *In*: HOUSEHOLDER, F.W. & SAPORTA, S., eds. — *Problems in lexicography*. Bloomington, The Hague, Mouton, 1967. p. 25-44.

WEINREICH, U. — Webster's third: a critique of its semantics (review). *International Journal of American Linguistics, 30*:405-409, 1964.

WICLOW, C.K. — Advanced English vocabulary. *Language Learning, 24*(1): 167-170.

ZGUSTA, L. — *Manual of lexicography*. The Hague, Mouton, 1971.

ZINGARELLI, N. — *Vocabolario della lingua italiana*. 10. ed. Bologna, Zanichelli, 1971.