LEXICOGRAFIA DE *CORPUS* E A DICIONARÍSTICA CONTEXTUALISTA

Mauro de Salles Villar*

> Penetra surdamente no reino das palavras.
> Lá estão os poemas que esperam ser escritos.
> Estão paralisados, mas não há desespero,
> há calma e frescura na superfície intacta.
> Ei-los sós e mudos, em estado de dicionário.
> — Procura da poesia, *ROSA DO POVO*

Que entender por esse 'estado de dicionário' de que Carlos Drummond de Andrade com fineza se serve?

Nos dicionários semasiológicos tradicionais, as palavras registram-se como entidades congeladas, autonomizadas de contextos e ordenadas alfabeticamente numa macroestrutura. Classificam-se por sua categoria gramatical na estrutura interna do verbete e apresentam convencionalmente acepções fixas, de contornos delimitados, o mais possível variadas.

E como as palavras funcionam na língua? Nas classes gramaticais, a volatilidade de limites no português é grande. As palavras articulam-se dentro de contextos e associam-se, para a expressão de conceitos e idéias, segundo os padrões da sintaxe, tanto livremente como através de estruturas mais ou menos fixas — determinados tipos de seqüências memorizadas que funcionam em padrões combinatórios repetidos, nos quais, por sua enorme freqüência e importância, vale a pena atentar.

Por outro lado, pelo que se tem apreendido do estudo dos *corpora* das línguas naturais, as unidades léxicas sobre que versam os dicionários na verdade apresentam número restrito de acepções autônomas reais. A esse respeito já observara JURI APRESJAN (1974) que "os dicionários exageram muito na minúcia dos sentidos e tendem a estabelecer limites naquilo que um exame mais atento revela não mais que uma área intermediária, vaga, de sentidos que se superpõem." Os registros que os lexicógrafos captam muitas vezes não são

---

* MAURO DE SALLES VILLAR é membro da Academia Brasileira de Filologia, diretor do Instituto Antônio Houaiss de Lexicografia e co-autor do *Dicionário Houaiss da Língua Portuguesa*.

mais que contextualizações desses sentidos básicos. Pelo método tradicionalmente empregado, nos grandes dicionários as entradas acabam com tantas acepções quantas é capaz de distinguir a sensibilidade do dicionarista ou a diversidade do material que tem nas mãos, mas isso não tem fim, uma vez que cada nova tonalidade, cada nova nuança de utilização pode ser registrada como sentido ou emprego diferente.

Veja-se o exemplo do verbete **olho** no *Dicionário Caldas Aulete*, terceira edição brasileira (1974), — mas poderia ter usado qualquer outro dicionário, pois todos incorremos em tal disfunção. Realcei com retícula as acepções que ilustram o que disse acerca de um mesmo sentido (neste caso, o de 'buraco', 'vazio') contextualizado diferentemente e tomado como nova acepção:

> OLHO, *s. m.* (anat.) órgão da visão situado em órbita própria, de forma mais ou menos globular, ordinariamente em número de dois, colocados na parte anterior da cabeça do homem e de quase todos os animais. || Órgão da vista considerado como indício das qualidades ou defeitos do espírito, do caráter, das paixões, dos sentimentos: A bondade brilha nos seus *olhos*. || (Fig.) Olhar, luz, clarão, ilustração: A reflexão é o *olho* da alma. Vendo pelos *olhos* do espírito a desonra e o desprezo, e ouvindo a desesperação gritar. (R. da Silva.) A geografia e a cronologia são os *olhos* da história. || Atenção, esforço da alma aplicado a um objeto: Tem os *olhos* fixos no seu dever, na sua conduta. || Atenção, vigilância, cuidado: O ladrão escapou, aos *olhos* da polícia. Os *olhos* da real benignidade. (Camões.) Traz o *olho* no criado que o rouba. || Ocelo. || Gota de líquido gorduroso que flutua sobre outro liquido mais denso. || **Buraco ou furo em certos objetos por onde se enfiam linhas ou fios.** || **Aro das ferramentas por onde se enfia o cabo: O *olho* da enxada.** || **(Pleb.) O orifício do ânus.** || (Gír.) Tostão. || **Vão nos tímpanos dos arcos da ponte para dar maior vasão à água.** || **Abertura por onde entra a água que faz mover a roda dos moinhos.** || (Tipogr.) A espessura de um caráter de imprimir; **a abertura no *e* que distingue esta letra do *c*.** || **Poro ou buraco que apresentam certas massas e especialmente os queijos.** || **(Arquit.) Abertura circular ou elíptica feita nos tetos ou paredes dos edifícios para lhes dar claridade.** || **(Metalurg.). O buraco da fieira por onde passa o metal que se quer adelgaçar.** || (Alcanena) Porção de qualquer casca, que serviu num tanque de curtimenta. || Batoque ou **orifício na parte superior e anterior do tonéis e que serve para lhes introduzir o líquido e tirá-lo depois de fermentado.** || **O buraco da pedra superior ou girante (falando da mó dos moinhos) por onde cai o trigo e outros cereais para serem reduzidos a fari-**

> **nha.** || O botão que vai desenvolvendo-se na planta ou o rebento das árvores: *Olho* da couve. E sejam guarnecidas com *olhos* de alface, ou de chicória. (Domingos Rodrigues, *Arte de Cozinha,* I, c. 1, 4, p. 5, ed. 1693.)

As acepções dos vocábulos têm utilização bastante fluida. Elas se expandem, quer horizontal, quer verticalmente, por meio de analogias, metáforas, metonímias, extensões de sentido e tantas outras formas de contaminação semântica. Diante da flexibilidade da língua, o entendimento entre os falantes só se dá por estes porem em ação a sua habilidade de interpretação. Praticamente, qualquer palavra de uma língua apresenta uma dinâmica semântica aberta, podendo ser sempre empregada e colorida de um novo modo pelos usuários.

PATRICK HANKS (2000) pondera sobre a questão, propondo um modelo de dicionário em que as palavras não tivessem acepções rigidamente marcadas, os tradicionais lexemas separados por números, mas sim que fosse apresentado um grupo de sentidos potenciais de cada unidade léxica, capaz de ser ativado pelos contextos particulares. Na introdução do *New Oxford Dictionary of English* (1998), em que havia posto em funcionamento a sua tese, ele esclarece que as entradas compreendem um ou mais sentidos 'genéricos' e cada um deles faz as vezes de passagem para outras subacepções relacionadas e mais especializadas.

Outra tentativa recente nessa direção foi feita pelo *Macmillan English Dictionary for Advanced Learners* (2002), que aceita o modelo dos dicionários convencionais nas palavras que apresentam acepções claramente distintas, mas em outros casos elabora conjuntos de sentidos mais complexos, em que a base semântica é desenvolvida de maneiras diversas.

Enquanto isso, o que vêm fazendo os dicionários semasiológicos ditos de língua geral? Trabalham a descrição das unidades léxicas como se estas apresentassem sentidos constantes, partilhados por todos os falantes. As relações de sentido, porém, não são nem estáveis nem predizíveis, não havendo, mesmo, maneira de saber como cada um de nós lida com os significados e como os representa internamente. Com freqüência, a proliferação de sentidos dicionarizados não passa de incapacidade de o lexicógrafo atingir o nível de generalização correto, ou então trata-se de incidência da velha indistinção entre sentido diferente e contexto diverso. Os dicionários priorizam também os critérios classificatórios, passando por cima do que é conotativo, pragmático, atitudinal ou reduzem-no à condição de umas poucas notas (SINCLAIR: 2004).

Hoje parece óbvio não poder priorizar-se nos dicionários uma classificação baseada no *sentido* dos vocábulos em detrimento do seu *uso* (SINCLAIR: 2004). Essa foi a revolução introduzida na lexicografia pelos contextualistas britânicos.

Desde a década de 1960, um grupo de lingüistas, semanticistas e lexicógrafos britânicos, mais tarde batizados de escola contextualista, percebeu que os computadores podiam ser empregues no armazenamento de textos e nas pesquisas de abonações. A partir dessa prática, a lexicografia e a lingüística puderam analisar em detalhe o funcionamento das palavras da língua, livres da artificialidade das escolhas pessoais de exemplos antes praticada. Com crescentes milhões de exemplos à mão, os *corpora* computadorizados transfiguraram a lexicografia.

Foi assim que se perceberam fenômenos curiosos dentro das línguas. Por exemplo, usamos em nossa comunicação vocabular grande número de combinações fixas ou mais ou menos fixas de palavras, que funcionam como elementos pré-fabricados, prontos a usar. Sua compreensão pelo ouvinte é ótima, pois se trata de recurso repetitivo, e isso poupa energia na expressão. Percebeu-se também, ligado a esse fato, o fenômeno da imantação vocabular, que faz que, entre dezenas de escolhas na língua, as palavras se unam percentualmente em alto grau apenas em determinados sintagmas.

Vamos analisar esses e outros fatos mais à frente. Agora, interrompo o que digo, para transcrever uma pequena série de exemplos desse comportamento vocabular, para sua melhor apreensão.

Quando determinado conceito precisa ser expresso num discurso, o que vem à cabeça de imediato são elementos combinatórios pré-fabricados assim. Repare:

> || efeito perverso, virtudes cardeais, assistência social, casca grossa, mundo civilizado, prova cabal, valores morais, bala perdida, deuses pagãos, honras fúnebres, dinheiros públicos, obra aberta, chamas eternas, cunho religioso, alma penada, vendedor autônomo, gênio incompreendido, junta comercial, sexo seguro, preferência nacional, cálculo estrutural, foro privilegiado, *deficit* habitacional, força policial, pessoa jurídica, criança mimada, vista curta, última vontade, instante supremo, sucesso brilhante, empréstimo compulsório, artilharia pesada, reprodução assexuada, pesca predatória, colorido orquestral, crime passional, última moda, tinta fresca, vida eterna, carros clássicos, meio ambiente, riso amarelo, bilhete azul, língua negra, pensamento positivo, ardor missionário, pecado mortal, tresloucado gesto, mobiliário urbano, estado crítico, pretinho básico, espetáculo circense, cultura popular, música erudita, vontade política, piloto automático, panos quentes, tríduo momesco, carro alegórico, inclusão digital, desenvolvimento sustentável, crescimento sustentado, ponto morto, vaso sanitário, renúncia fiscal, fome zero, lucro cessante, massa falida, imprensa marrom, mudança radical, vontade louca, olho

grande, bandeira amarela, livre concorrência, analfabeto funcional, boi gordo, fogo cerrado, duras penas, notório saber, reserva técnica, fogo cruzado, preços salgados, sigilo bancário, corpo fechado, figurinha carimbada, recurso extraordinário, horário nobre, aula particular, círculo virtuoso, círculo vicioso, arma branca, tiro livre indireto, picanha maturada, bens tombados...

|| piloto de provas, cartão de natal, almas dos justos, profissão de fé, espírito de porco, poder de veto, objeto de desejo, colônia de férias, banho de loja, paciência de Jó, paletó de madeira, febre de feno, quebra de protocolo, papas na língua, ordem de idéias, população de baixa renda, rolamento da dívida, camisa de onze varas, o espetáculo do crescimento, regime dos ventos, duro na queda, poucas e boas, elas por elas, rápido e rasteiro, pau a pau, uma ova, a duras penas, gol de ouro, livro de cheques, voto de confiança, invasão de privacidade, via de regra, lavagem de dinheiro...

|| meter a mão, não é nada não é nada..., uma gota no oceano, o fim da picada, sem sombra de dúvida, na expressão da palavra, marinheiro de primeira viagem, dois dedos de prosa, um belo dia, era uma vez, trocar as bolas, tomar a peito, ter pavio curto, pra ninguém botar defeito, na crista da onda, na corda bamba, ter minhocas na cabeça, ter macacos no sótão, estar fora de si, estar na água, da mão para a boca, meter os pés pelas mãos, abrir todas as portas, falar claro, jogar limpo, descongelar preços, serem favas contadas, estar pela hora da morte, num abrir e fechar de olhos, limpar a barra, pôr em pratos limpos, matar a pau, a vaca foi pro brejo, fazer fita, dar com os burros na água, durma-se com esse barulho, segurar as pontas, o mar não está para peixe, pão-pão queijo-queijo, voltar à vaca-fria, ser pé-quente, tirar o atraso, sem eira nem beira, ter bala na agulha, cair a ficha, passar lotado, soltar os cachorros, partir para o abraço, levar às últimas conseqüências, passar energia positiva, uma química perfeita, na medida do possível etc.

(Muitos desses exemplos, especialmente os últimos, são de linguagem informal, mas tal nível de uso é ocasional. Estes foram apenas os que me ocorreram ao tentar levantar em pouco tempo uma lista desse fenômeno em nossa língua.)

As combinações com que nos deparamos no discurso não são, obviamente, todas desse tipo. Há-as livres, nas quais é regular a soma dos significantes e dos significados do sintagma, e que podem ser substituídas por quaisquer outras combinações suficientemente sinônimas. *Água gelada*, *terra árida*, *chuva*

fria são exemplos de combinações livres. Uma frase como "este dicionário foi feito por um grande grupo" poderia ser dita "este léxico é resultado do trabalho de muitos lexicógrafos e colaboradores", por se tratar de um sintagma de *combinações livres*. Mas estes casos não nos interessam aqui.

Deixemo-los de lado e debrucemo-nos sobre as *co-ocorrências lexicais restritas*, também ditas *combinatórias lexicais não livres* — além de diversas outras denominações. São estas as que demonstram tendência de adotar tão-somente um número limitado de associações com outras palavras, dentre grande quantidade de combinações possíveis. Para fazê-lo, vou utilizar-me da classificação de Igor Mel'čhuk, autor do celebrado *Dictionnaire explicatif et combinatoire du français contemporain*, cuja análise das co-ocorrências é bastante interessante.

As combinatórias lexicais não livres estão genericamente catalogadas por Mel'čhuk em *sintagmas semânticos* e *sintagmas pragmáticos*. Comecemos pelos semânticos. Estes podem ser de três gêneros: *frasemas*, *semifrasemas* e *quase-frasemas*.

Os frasemas completos são a combinação de dois ou mais lexemas A + B, cujo significante é a soma regular dos significantes dos lexemas constituintes / A + B/, mas cujo significado é diferente da soma dos significados constituintes.

Observe as seguintes associações para melhor entender a teoria: *saia justa*, *televisão de cachorro*, *olho grande*, *bafo de boca*, *boca de siri, lua-de-mel*. Qualquer pessoa sabe o que significa o substantivo *saia* e o adjetivo *justo*, mas isso não basta para apreender o significado do sintagma *saia justa*, uma vez que seu sentido é dissemelhante da soma dos significados constituintes: 'situação embaraçosa'. O mesmo ocorre com os outros exemplos. *Televisão de cachorro*, no Brasil, é aquela 'máquina em que ficam girando, nas padarias, os frangos em cozimento'. Esse tipo de associação, cujo sentido vai além da soma dos significados de cada parte constituinte, é, na classificação de Igor Mel'èhuk, o *frasema*, e sua natureza é a das expressões idiomáticas.

Repare, agora, nas co-ocorrências *imprensa marrom, sorriso amarelo, água dura*. São de outro tipo. Nessas combinações de dois lexemas, A + B (que também poderiam ser mais de dois), o significante é a soma regular dos significantes dos lexemas constituintes /A + B/, mas apenas o sentido do adjetivo é diferente de sua acepção original, o que faz que a soma dos significados constituintes resulte em outra coisa. *Marrom* aqui não é 'cuja cor é a da castanha', mas 'sensacionalista, caluniadora'. *Amarelo* não é 'da cor da gema do ovo', mas 'contrafeito'. Duro não é 'não é flexível ou macio', mas 'que contém sais de cálcio, magnésio e ferro em quantidades tais que dificilmente produz espuma com sabão'. Esse tipo de associação é, na classificação de Mel'čhuk, o *semifrasema*, equivalente à *collocation* dos lingüistas anglófonos, e tem uma

característica especial: as palavras usadas nos sentidos "afastados dos originais" só se empregam com tais acepções nas associações aqui registradas. Não se pode usá-las com a mesma acepção em outras ocorrências. Por exemplo, não é possível dizer "Fulano sentiu-se amarelo" por "Fulano sentiu-se contrafeito", nem citar uma "carta marrom", querendo significar uma "carta caluniadora".

Outra curiosidade nos semifrasemas é que, mesmo que diferentes adjetivos signifiquem a mesma coisa, eles não são permutáveis nas co-ocorrências em que são usados. Por exemplo, em *atividade febril*, *luta encarniçada*, *ódio mortal*, *vontade louca*, todos os adjetivos foram usados no sentido de 'acentuado'. Tente, porém, trocá-los nas citadas colocações e verá que o uso não confirma tal possibilidade: Atividade mortal? Luta louca? Vontade encarniçada... Já não significam a mesma coisa. O fenômeno da imantação é, portanto, "pessoal" e (praticamente) "intransferível".

O levantamento dos frasemas e semifrasemas é fundamental na língua, especialmente nos dicionários bilíngües e plurilíngües, uma vez que quem aprende um idioma ou quem tem de verter ou traduzir textos carece de que tais combinatórias lexicais sejam esclarecidas e seus equivalentes ou descrições parafrásticas sejam informadas.

Vimos acima o caso dos sintagmas cujo significado é diferente da soma dos significados constituintes, quer porque um dos elementos 'funciona diferentemente', quer porque todo o conjunto o faz. Vejamos agora a terceira modalidade de sintagmas semânticos de associação restrita, o *quase-frasema*. Um bom exemplo deste é a locução *centro comercial*. Repare que, em ambos os componentes, é regular a soma dos significantes e dos significados, uma vez que se trata de um *centro* (porque para ali convergem lojas) que é *comercial* (porque nele se fazem negócios). Mas a Saara, no Rio de Janeiro, e a 25 de Março, em São Paulo, seriam igualmente centros comerciais, uma vez que ali existe convergência de variadas lojas e naquele local se mercancia — mas ninguém lhes atribuiria esse epíteto. Por quê? O motivo é que, embutida na locução, existe a presunção elíptica de que um centro comercial seja composto de lojas em andares superpostos, com garagens, que exista o oferecimento de serviços (bares, restaurantes, cabeleireiros, supermercados, praças de alimentação etc.). Essa composição entre sentidos expressos e não expressos é o que caracteriza os quase-frasemas da classificação de Mel'čuk.

Faltou, então, falar dos sintagmas pragmáticos, os *pragmatemas*. Eles são os conjuntos empregados na língua de modo fixo e que se repetem para fins práticos, como as seguintes fórmulas: agite antes de usar; este lado para cima; proibido fumar; graças a Deus; o Ministério da Saúde adverte: fumar causa...; ver validade no fundo da garrafa, consumir de preferência antes de...; se persistirem os sintomas, o médico deve ser consultado; saída de emergência etc. É

também considerada pragmatema a linguagem fática — aquela usada não para comunicação de informações, mas apenas para assinalar que o canal de comunicação está aberto. Por exemplo, quando você encontra alguém e pergunta "como vai", não quer sabê-lo de fato. Se o outro responde que vai bem, não lhe está afirmando isso, mas apenas utilizando uma fórmula padronizada de comportamento socialmente aceitável. Exemplo de uma conversa com esse tipo de linguagem (em versão informal):

— Como é? Tudo nos conformes?

— Tudo em cima.

— Beleza! É isso aí.

— Então tá.

Nada foi efetivamente perguntado e nada foi verdadeiramente respondido — nem se espera que o indagado vá contar-nos a sua vida naquele instante. Trata-se de simples pragmatemas.

Outras categorias de palavras são consideradas pragmatemas. Por exemplo, os verbos operadores de ações, também ditos verbos-suportes, que constituem com o substantivo (que na gramática tradicional faz de seu objeto direto) um todo semântico, tendo o seu sentido original esvaziado. Quando você diz: *dei um prêmio à Joana*, o verbo *dar* está utilizado em sentido pleno. Mas em *dar um pulo* (= pular), *dar um sorriso* (= sorrir), *soltar uma gargalhada* (= gargalhar), *fazer questão*, *passar um descompostura* etc. os verbos fazem apenas de operadores.

Há autores que incluem também os epítetos, as antonomásias e os provérbios na categoria de pragmatemas.

É bom lembrar, porém, que os exemplos que aqui dei de frasemas, pragmatemas etc. são obviamente exemplos-tipo. Escolhi-os por serem muito característicos dessas classificações, mas não é tão simples qualificar as coocorrências nos *corpora*, uma vez que esses fenômenos não são discretos, mas sim contínuos e graduais no tecido da língua, sendo difícil a sua exata categorização para os lingüistas e lexicógrafos.

Mas por que estou falando de todas essas coisas aqui? Tudo isso, e mais fenômenos em que não toquei, como as descobertas da recente *prosódia semântica*, que estuda o modo como uma classe semântica inteira pode ter forte tendência a associar-se com determinadas palavras, mostra que a língua carece de ser detalhadamente entendida para ser melhor descrita nos dicionários, e foram as grandes bases de dados de abonações e a pesquisa das estruturas fraseológicas que permitiram esses desenvolvimentos e percepções. Não teria sido possível observá-los, analisando as palavras individualmente.

E com que bancos de palavras conta a lexicografia? Na França, o Inventaire General de la Langue Française colecionou, entre 1936 e 1968, cerca de 6 milhões de abonações de palavras empregadas em textos literários e técnicos. Esse material e muitas outras fontes serviram, mais tarde, para elaborar os alentados 16 volumes do *Trésor de la langue française*. Na Inglaterra, o Brown Corpus, na década de 1960, recolheu 1 milhão de abonações, mas logo ficou claro que isso era absolutamente insuficiente. Vieram então, na década de 1990, o Bank of English e depois o British National Corpus, que reuniu 100 milhões de palavras e a seguir, 250 milhões, num inventário aberto ao público que registra também vocábulos e fraseologia do inglês norte-americano. Há três anos ele já estava em 400 milhões de ocorrências inventariadas.

Nos *corpora* acima de 100 milhões de itens com simples concordâncias à direita e à esquerda da palavra focada, surgem os padrões de emprego da quase totalidade de vocábulos da língua, com exceção dos mais raros, observa PATRICK HANKS (2002). Outros padrões emergem através de elaboradas análises computacionais. Os bancos ingleses de palavras que citei são públicos, mas há também os particulares, de grandes companhias jornalísticas, de universidades, e ainda todo o oceano vocabular da internet, que pode ser usado. Em 2002, Gregory Grefenstette, cientista pesquisador da Clairvoance Corp. (Penn.) dava conta de haver na rede 76 bilhões de ocorrências de inglês (e já 1.333.664 palavras do português). E estamos há três anos desse cômputo.

E o que se está fazendo na língua portuguesa? Pouco. O fenômeno das combinatórias lexicais está mal desenvolvido. No Brasil, as tentativas bem-sucedidas citáveis de trabalhos feitos com o auxílio de computadores são, por exemplo, o projeto NURC, de 1996, ligado ao Proyecto de Estudio Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de la Península Ibérica. Também o banco de palavras que Francisco S. Borba vem desenvolvendo em Araraquara e que já gerou um *Dicionário de usos do português do Brasil* (2002) e um recente *Dicionário Unesp do português contemporâneo* (2004), elaborado sobre uma base que também levou em conta o *Corpus* de Referência do Português Contemporâneo, do Centro de Lingüística da Universidade de Lisboa, segundo informa o seu texto de introdução. Na USP, professores como Tony Berber Sardinha, Heloísa Collins, Maria Adélia Ferreira Mauro, Zilda Maria Zapparoli, João Martins Ferreira e outros mergulharam na lingüística computatorial, enquanto Ronaldo Martins procura desenvolver para o português o aventuroso projeto da Universal Networking Language (UNL) de ligar as línguas do mundo por um sistema de tradução automática — uma patente da ONU.

A maioria dos dicionários no português é construída sobre cópia e descaracterização de obras anteriores, em vez de se basear em aprofundamento de

estudos, o que é péssimo, pois além de perpetuar erros, acaba por introduzir impropriedades em definições que estavam boas. "Os dicionários portugueses geralmente adotados no uso e no ensino são maquinalmente copiados uns dos outros", observava já em 1881 F. J. Caldas Aulete, no texto do plano do *Dicionário contemporâneo da língua portuguesa*, em sua primeira edição. Em grande parte, continua-se a exercer a lexicografia como uma 'arte do plágio'. Para nos livrarmos disso, é fundamental que um grande banco público de ocorrência de palavras seja encetado em bases científicas, a fim de que a língua seja analisada e eficientemente descrita.

Um banco capaz de atender à demanda da língua terá de voltar-se para o português do presente, mas também do passado, mergulhar na leitura e registro de ocorrências em nossa literatura e no que vive nos jornais, revistas, manuais técnicos, pesquisar a área da terminologia etc., e registrar a pragmática, o nível de uso das palavras da língua, as suas combinatórias lexicais. Isto se faz não em poucos, mas em muitos anos de trabalho perseverante, e exige uma equipe de dedicados informatas, lingüistas, semanticistas e lexicógrafos em trabalho conjunto. O inglês vem desenvolvendo há décadas ferramentas de pesquisa automática e de armazenamento cada vez mais eficientes: analisadores gramaticais, *crawlers*, analisadores morfológicos, identificadores de linguagem, *taggers* para segmentos de discurso, classificadores de domínio e gênero, etc. É preciso, em grande parte, adaptá-las ou criar ferramentas afins, para dar conta de nossas especificidades.

O português, repito, tem de ser estudado em suas estruturas e nas suas relações semânticas mais finas, para ser entendido e explicado apropriadamente. Sem isso, os nossos dicionários continuarão a passar na tangente da exatidão desejável — tanto os monolíngües como especialmente os multilíngües. Pelo mundo, o levantamento de frasemas, quase-frasemas e semifrasemas das línguas segue adiantado. Estamos atrasados nisso.

Quem deveria investir nesse trabalho fundamental, já que a língua é um dado fundamental da cultura e da coesão de qualquer povo? No Brasil, as universidades estão carentes de fundos e não se dispõem a fazê-lo. A Academia Brasileira de Filologia seria perfeita para isso, mas mal tem dinheiro para publicar a sua própria revista e livros. Estou certo de que esse esforço terá de vir de alguma parceria entre uma instituição como a Academia Brasileira de Letras ou a Fundação Casa de Rui Barbosa e dinheiros públicos para a pesquisa e a cultura, e é isso que se espera que ocorra o mais proximamente possível, pois seria irresponsabilidade alienante preterir a política da língua. Com tal aperfeiçoamento, a nossa lexicografia irá atingir novos patamares de excelência e instalar-se-á confortavelmente no convívio das mais desenvolvidas do mundo.

Conclusão: Os grandes bancos de ocorrências de palavras têm revelado que as unidades léxicas dos dicionários contam, na verdade, com menos lexemas puros do que se costuma registrar; que as classes gramaticais são bastante fluidas e que os vocábulos demonstram forte tendência para se associarem em sintagmas mais ou menos restritos que se repetem. O uso das palavras parece mais importante que os sentidos fechados que lhes dão os dicionários, em vista da dinâmica de fluxos semânticos e de níveis em que os vocábulos são empregados, o que sugere outro projeto para os léxicos mono e plurilíngües. É preciso, por isso, rapidamente desenvolver no português extensos bancos de palavras num esforço conjunto público-privado, a fim de descongelar o 'estado de dicionário' convencional e torná-lo numa ferramenta dinâmica. Assim conseguiremos fazer que o 'estado' dos nossos dicionários ganhe, afinal, a exatidão de que na verdade carece.

Referências bibliográficas

APRESJEAN, JURI D. (1974). 'Regular polysemy'. *Linguistics*, 142, 9. Mouton, Haia.

HANKS, PATRICK (2000). 'Do word meaning exist?' *Computers and the Humanities* 34: 205-215, cit. por MICHAEL RUNDELL in *Lexicography and Natural Language Processing - A Festschrift in Honour of B.T.S. Atkins* (2002). EURALEX, p. 148-8.

HANKS, PATRICK (2002). *Lexicography and Natural Language Processing - A Festschrift in Honour of B.T.S. Atkins*. EURALEX, p. 157.

MEL'ČUK, Igor, A. •OLKOVSKIJ (1970). Towards a Functioning Meaning-Text Model of Language. In: *Linguistics,* 57 pp 10-47.

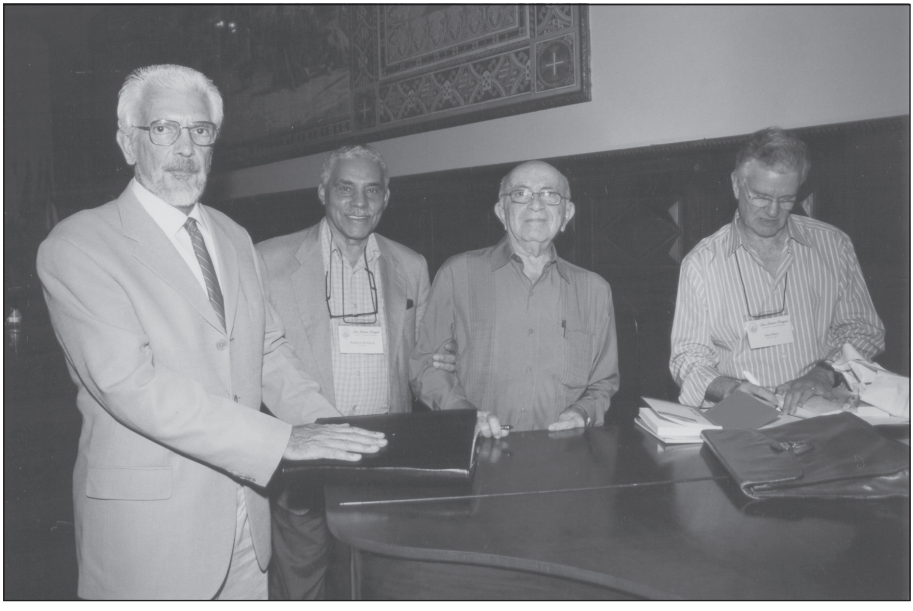
MEL'ČUK, Igor et *al*. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, vol. I, II, III, IV, Montréal, Les Presses de l'Université de Montréal.

MEL'ČUK, I; A. CLAS, A. POLGUÈRE (1995). *Introduction à la lexicologie explicative et combinatoire*. Coll. Champs linguistiques/Université francophones, Louvain-la-Neuve/Paris: Éditions Duculot/AUPELF-UREF.

MEL'ČUK, Igor (1995). "Phrasemes in Language and Phraseology in Linguistics", *Idioms*: *Structural and Psychological Perspectives*, M. Everaert, E.J. van der Linden, A. Schenk et R. Schrender (éds), Hillsdale/ Hove, Lawrence Erlbaum Associates, pp. 167-232.

NEW OXFORD DICTIONARY OF ENGLISH (1998). Ed. Judy Pearsall. Oxford: Clarendon Press.

SINCLAIR, JOHN (2004). 'To complement the dictionary'. The Tuscan Word Centre.



Da esquerda para a direita
Mauro Villar, Rosalvo do Valle, Adriano Kury, Dino Preti.