A lingüística de corpus e o dicionário de língua

Francisco da Silva Borba – UNESP/Araraquara

A denominação lingüística de corpus (LC) é recente, mas ela é um desenvolvimento da noção estruturalista de corpus, pois o define como uma coletânea grande e criteriosa de textos de linguagem natural. A LC volta-se para o uso na tentativa de descoberta de padrões de associação, pois se apóia no pressuposto de que o ser humano não é dotado da capacidade de perceber o que é típico, ao contrário, é equipado para notar aquilo que se destaca, isto é, o atípico. A abordagem baseada em corpus permite buscar respostas à questão da tipicidade porque faz uso do computador, o qual é naturalmente programado para detectar ocorrências e co-ocorrências.

Ocupando-se do produto lingüístico a LC só pode chegar ao sistema por meio de generalizações indutivas.

Qualquer dicionário orienta-se por princípios estabelecidos a partir de seus objetivos. Se o dicionário de língua (DL) resulta da análise de um corpus previamente estabelecido, então o lexicógrafo precisa tomar algumas posições para levar a cabo, de modo coerente, sua tarefa. Entre elas contam-se as relações entre dicionário e gramática, entre dicionário e texto e entre dicionário e ideologia.

Como instrumentos pedagógicos, dicionário é gramática têm pontos em comum, mas não se superpõem: o dicionário é o lugar do particular, do tópico, e a gramática é o lugar do genérico, das regras; o dicionário enumera palavras, a gramática enumera regras; o dicionário é um acervo de formas livres, a gramática contém um conjunto de regras que, aplicadas, mostram como a língua funciona. A gramática apresenta, de forma sistemática, um conjunto de regras de combinatória e de interpretação dos constituintes da língua, em seus diversos níveis.

A contextualização de acepções em dicionários não faz parte de nossa tradição lexicográfica, que segue a tradição românica de listagens e acervos de palavras e acepções. O fato de o DL resultar da análise real dos textos apresenta vários níveis de dificuldades, entre as quais, o registro dos intertextos. Não existe texto neutro quanto à ideologia, se se entende esta como um conjunto de idéias, opiniões, valores, crenças etc., que expressam, explicam ou justificam a ordem social, as condições de vida do homem em suas relações com os outros homens. Quem fala ou escreve pretende sempre colocar [sugerir, propor, impor, inculcar], mesmo que implicitamente, seu modo de ver e sentir o universo, seus pontos de vista e suas convicções, seu sistema de crenças etc. Quem recebe o texto, pode aceitar ou discutir o que recebe como também pode captar totalmente, parcialmente ou mesmo nulamente o que está implícito.

A adoção de um corpus facilita a organização objetiva da macroestrutura do DL. Assim, por ex, o DLP baseia-se na ocorrência real nos textos [lg escrita no Brasil, a partir de 1950] associada à freqüência de certos tipos. Por esse critério a nomenclatura fica composta de três camadas: (i) itens que compõem a base do léxico da língua e que ocorrem em qualquer contexto e em qualquer registro. Consta de cerca de 14000 tipos com freqüência mínima de 6. Reconhece-se esse núcleo pela presença maior de vocábulos polissêmicos. (ii) itens que compõem a base ampliada, que alcança vários setores da vida social e vários aspectos culturais. Têm freqüência mínima 4 e conta com ±14000 tipos. Aí, ao mesmo tempo em que aumenta a presença de vocábulos monossêmicos, diminuem os polissêmicos e (iii) itens que circulam na língua escrita como um todo, com freqüência

diferente de zero, e abrangendo todos os setores da vida social. Abrange cerca de 30000 itens e se caracteriza pela predominância de itens monossêmicos.

O primeiro passo para a organização de um DL é descrever o uso, quer dizer, mostrar como funcionam efetivamente os diversos setores da língua num determinado estágio de seu desenvolvimento. Satisfeita esta etapa, passa-se à seleção e arranjo dos elementos descritivos que deverão figurar nos verbetes. Tal seleção está relacionada com o objetivo específico do dicionário. Se o objetivo é registrar todo o uso, então caberiam informações relacionadas com todos os níveis da estrutura lingüística. De fato, é por aí que se pode guiar, dar condições de controle e agilizar o uso. A seleção desses elementos vai dar num modelo descritivo de cada verbete. Ex DUP DLP