4
Metodologia

O método de descrição do léxico das palavras compostas se apóia em textos do português do Brasil processados em suporte eletrônico, anotações de jornais e revistas. Esses textos são representativos da língua falada e escrita, de conteúdo irrestrito. Será utilizado também o dicionário eletrônico português de Portugal (Ranchhod *et al*., 1999), elaborado no LABEL (Universidade de Lisboa), a ferramenta INTEX (Silberztein, 1994), elaborada no LADL (Universidade de Paris 7) e a ferramenta UNITEX elaborada por Sebastien Paumier (2002), no Instituto Gaspard-Monge (Universidade de Marne-la-Valée) para consulta a dicionários eletrônicos.

As formas extraídas dos textos constituem listas e resultados de concordâncias. O estudo lingüístico dessas formas se baseia nos critérios de aceitabilidade dos enunciados do locutor nativo e na consulta a dicionários. As formas efetivamente compostas são selecionadas nessas listas. Durante o processo, novos compostos que não figuravam nas listas extraídas dos textos foram acrescentados, e as ambigüidades lexicais relativas às propriedades gramaticais e morfológicas são detectadas. As entradas obtidas por esse processo serão classificadas e codificadas de modo a especificar as propriedades gramaticais e morfológicas essenciais: gênero e número, variações eventuais e formas flexionadas. Tal codificação permite a geração automática de todas as formas flexionadas das novas entradas, bem como a possibilidade de serem reconhecidas automaticamente por sistemas de consulta a dicionários eletrônicos (Gross, 1986).

Nesta pesquisa as seqüências de palavras serão analisadas sob os procedimentos metodológicos da teoria do léxico-gramática apresentada por Gross (1975). A metodologia do léxico-gramática foi elaborada numa perspectiva de tratamento automatizado da língua e se propõe estabelecer um inventário de informações lingüísticas: explícitas, precisas e exaustivas.

4.1
O Corpus

Para realizarmos esta pesquisa, foi selecionado, inicialmente, um *corpus* de 30.000 ocorrências de palavras, com estrutura NdeN, colhidas de textos do "Jornal do Brasil", "O Estadão", "A Folha de São Paulo" e a "A Gazeta" por meio de programa computacional e 10.000 ocorrências de uso do português europeu.

A escolha pela literatura jornalística se justifica por sua importância na medida em que é aí que há não só variedade de autores, mas, principalmente, grande variedade de assuntos e enfoques.

Segundo Borba (1999), um dicionário de língua deve apresentar, topicamente, a estrutura e o funcionamento da língua, se possível num sistema bem nítido de notação. Um dicionário nunca deverá ser tomado como apenas um simples repertório ou acervo de palavras; ao contrário, deve ser um guia de uso e, como tal, torna-se um instrumento pedagógico.

Com esse propósito e com propósitos descritivos que precedem a montagem de um dicionário eletrônico, utilizou-se uma metodologia que privilegia a função de interação social da linguagem, procurando observar como as combinações de palavras circulam na língua escrita e falada no Brasil.

4.2
O Software Unitex

Unitex é um conjunto de programas que possibilitam o tratamento de textos em língua natural utilizando recursos lingüísticos. Esses recursos encontram-se sob a forma de dicionários eletrônicos, de gramáticas e tábuas de léxico-gramática e têm origem nos trabalhos desenvolvidos por Maurice Gross no Laboratoire d'Automatique Documentaire et Linguistique (LADL). Esses trabalhos têm sido desenvolvidos também em outras línguas pela rede de laboratórios RELEX. Os resultados dessas pesquisas podem ser validados com diversas ferramentas de Processamento da Linguagem Natural. Dentre essas ferramentas destaca-se o Unitex, elaborado por Sebastien Paumier (2002).

Espera-se que os resultados desta pesquisa possam ser validados pelo

Laboratoire d'Automatique Documentaire et Linguistique (LADL), constituindo informações do dicionário eletrônico de palavras compostas do português do Brasil inseridos no Unitex.