1
Introdução

1.1
Apresentação do tema

Segundo Basílio (1987), as principais funções do léxico são a representação conceitual e o fornecimento de unidades básicas para a construção dos enunciados. Para atender a essas funções, o léxico deve ser um sistema dinâmico, em constante expansão, na medida em que também se expandem continuamente nossas necessidades de novas unidades conceituais e de construção. Assim, pode-se conceber o léxico como o local de interface sociocultural, porque armazena o conhecimento e permite a formação de palavras novas que venham a atender às nossas necessidades de comunicação. Nesse sentido, os processos de formação de palavras são relevantes para as duas funções, uma vez que permitem seu atendimento de modo praticamente automático, a partir de elementos previamente estruturados no léxico. Entre os processos de formação de palavras, nesta pesquisa, faz-se um recorte com prioridade para as palavras candidatas a compostas com estrutura substantivo (ou nome) - "de" - substantivo (ou nome) – doravante N de N, considerando-se duas razões: a) este processo é um desafio para os estudiosos que reclamam por informações sintáticas, morfológicas, semânticas e lexicais para explicar esse fenômeno da língua e b) este processo é muito produtivo no uso da língua, daí a necessidade de critérios formais que possam identificar esse tipo de item lexical, tornando possível a representação dessa categoria de palavra para a construção de um dicionário eletrônico.

Segundo Ranchhod (2001:27-8), as palavras compostas constituem uma porcentagem muito elevada do léxico de qualquer língua. São freqüentes em todos os textos, mas são particularmente abundantes nos de natureza técnica e científica. Em processamento de linguagem natural, torna-se cada vez mais evidente a necessidade de analisar essas unidades lexicais, principalmente porque, muitas vezes, grande parte do sentido de um texto está vinculada à interpretação de nomes compostos.

1.2
Objetivos

Esta tese tem como objetivo geral desenvolver um estudo para a descrição dos tipos de composição nominal do português do Brasil, com a estrutura Nome+de+Nome (NdeN), na perspectiva de se construir um dicionário eletrônico para o processamento automático da linguagem natural. Com esse propósito, para se alcançar essa meta, mais dois objetivos de caráter mais específicos se constroem: o primeiro diz respeito às propriedades estruturais das seqüências de palavras com estrutura N de N. Apresentamos um estudo descritivo dessas estruturas, levando-se em conta as propriedades distribucionais, que estão relacionadas à natureza dos complementos, e as propriedades transformacionais, que dizem respeito às possibilidades de transformações com o apagamento de N2, a substituição de N2, a inserção de um item lexical entre os componentes da seqüência N de N, a ruptura paradigmática e outros procedimentos gramaticais, tais como: o processo de coordenação, a pronominalização e a passiva, etc. O segundo estudo diz respeito ao processamento automático das palavras, ou seja, ao uso desse conhecimento pela máquina. Nesse sentido, faz-se necessário representar as descrições lingüísticas dessas estruturas por meio de códigos que denotem as suas propriedades sintático-semântica e morfológica.

Além disso, pode-se afirmar que a necessidade de concretização desses objetivos se justifica na medida em que boa parte do vocabulário de uma língua consiste de palavras compostas e expressões complexas. Segundo Maurice Gross (1984), o número de palavras compostas constitui a maior parte do léxico de qualquer língua, e por isso a descrição de nomes compostos tem sido a preocupação de muitos lingüistas. Em várias línguas – por exemplo, francês (Gross 1975), inglês (Marchonis 1987), Português europeu (Eleutério et *al*, 1995 e Ranchhod *et al* 1999) –, onde foi feito o recenseamento das palavras simples e das palavras compostas, deu para constatar que o número de palavras compostas é maior do que o número de palavras simples e esses recenceamentos são um suporte estatístico. Essa é uma posição com uma base empírica bastante sólida. De acordo com Pamela Downing (1977), o processo de composição é altamente produtivo, mas apresenta muitas restrições. Há muitos pesquisadores trabalhando

nesse tipo de estrutura para caracterizar essas restrições e propor sugestões de como pode ser incorporado dentro de uma gramática.

Nesta pesquisa são estudadas as estruturas de nomes candidatos a compostos que podem ser descritos como compostos formados por justaposição, particularmente aqueles com a estrutura do tipo (N+de+N), em que N1 representa o primeiro nome e N2 o segundo, como, por exemplo, *boca de urna, bodas de ouro, prata da casa, dor de cabeça, dor de cotovelo, rato de porão, cartão de crédito, pai de família, lua de mel, pulo do gato, água de cheiro, água de coco, dona de casa, etc,* que, embora sejam seqüências constituídas por mais de uma palavra, funcionam lexicalmente como uma unidade. Essa característica é uma das questões mais relevantes que movimentam este estudo para a representação, pois daí decorre a necessidade de se identificar e se distinguir um grupo nominal livre de um grupo nominal composto, com a mesma categoria interna de palavras, ou seja, com a estrutura NdeN, visto que as estruturas compostas devem ser interpretadas como um bloco, um único item lexical. Essa propriedade faz uma grande diferença para o processamento da linguagem natural, pois o processamento adequado das palavras compostas pode evitar, por exemplo, a geração ou tradução de textos de forma imprecisa e incoerente nas relações de sentido, pois, numa análise automática de texto, uma seqüência de palavras, quando forma uma palavra composta, tem de ser analisada como um bloco, para que se possam construir representações adequadas de estrutura sintática e semântica das frases em que se encontra. Pretendemos, portanto, observar e descrever quais são as características dessas seqüências de palavras; quais são os fatores que interferem e determinam o uso como uma composição, para se definirem critérios que possam identificá-las num ambiente de geração automática de textos, isto é, para o processamento automático da língua portuguesa. Embora a composição se configure como um processo muito produtivo, há poucos estudos que possam explicar as relações estruturais e pragmáticas desse fenômeno, especialmente para processamento automático. Para o português do Brasil, ainda há muito por fazer. Esperamos encontrar explicações para esse fenômeno e dar, com este estudo, uma contribuição para a formalização desse tipo de unidade lexical.

Nesse sentido, trabalhamos com a hipótese de que há possibilidades de se definirem critérios formais para identificação de estruturas do tipo (N de N) compostas, o que pode ser de grande relevância para elaboração de um dicionário eletrônico de nomes compostos, utilizado em todo tipo de programa que lida com o processamento automático da linguagem natural. Segundo Dias (1994:2), o Processamento da Linguagem Natural é um campo de estudos multidisciplinar, pois engloba conhecimentos da Lingüística e da Informática, bem como de outras áreas. A tarefa da Lingüística dentro do Processamento de Linguagem Natural é possibilitar, com os recursos de que dispõe, utilizar ao máximo o que é previsível e determinado dentro da língua e explorar o que ela oferece em termos de interpretação e expressividade, nem sempre previsíveis, pois dependem também de fatores extralingüísticos, como a situação e o conhecimento compartilhado. Assim, levando-se em conta esses aspectos, esta tese propõe, como já dissemos inicialmente, um estudo descritivo das estruturas (N de N) candidatas a compostos, do ponto de vista de uma descrição lingüística e, a partir dessa descrição, apresenta uma codificação das propriedades estruturais, representando esse conhecimento de forma que possa ser utilizado numa base de dados para processamento automático.

Para descrever e codificar as descrições das propriedades de estruturas candidatas a compostos com vistas à elaboração de um dicionário eletrônico, alguns passos deverão ser seguidos:

a) selecionar as seqüências de estruturas candidatas a compostos;
b) descrever critérios lingüísticos que permitam identificar de modo operativo e reprodutível as estruturas compostas;
c) descrever as relações morfossintáticas;
d) descrever estruturas com ambigüidades lexicais provocadas pela homografia;
e) codificar o conhecimento a respeito das estruturas compostas, de modo que possam ser utilizadas por sistemas de processamento automático.

Em síntese, com essa descrição, pretendemos chegar a um conjunto de propriedades definidoras e diferenciadoras das unidades compostas, levando em conta as restrições distribucionais, a não-composicionalidade semântica, a convencionalização e a utilização terminológica, entre outros aspectos. A partir dessas propriedades, pretende-se formalizar esse conhecimento, criando um conjunto de regras e critérios de delimitação de unidades lexicais, bem como a incorporação de novos itens ao léxico. O resultado prático desse estudo é a elaboração de um dicionário eletrônico de palavras compostas do português do Brasil.

1.3
Desenvolvimento da pesquisa

A tese será desenvolvida sob dois pontos de vista: o das descrições lingüísticas das estruturas N de N candidatas a compostos e o da representação formal dessas descrições para processamento automático. Como resultado, será proposta uma codificação das propriedades dos nomes considerados compostos

O capítulo 2º expõe uma reflexão sobre a relevância do léxico para o processamento automático da linguagem natural e sobre a interação entre os campos da Lingüística e da Informática. O capítulo 3º apresenta os pressupostos teóricos que dão sustentação à descrição e formalização dos nomes compostos. O capítulo 4º apresenta o recurso eletrônico utilizado para selecionar as estruturas que compõem o *corpus*. O capítulo 5º faz uma revisão de literatura, sobre o tratamento dado às palavras compostas. O capítulo 6º descreve os critérios de identificação dos nomes compostos. O capítulo 7º discute os conceitos de composicionalidade e não-composicionalidade. O capítulo 8º descreve as propriedades dos nomes compostos por meio de uma codificação das propriedades. O capítulo 9º apresenta as conclusões do trabalho e as possíveis realizações de novas pesquisas, a partir de descrições de outras estruturas também para fins computacionais.