Aucione Das Dores Smarsaro

Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico

Tese de Doutorado

DEPARTAMENTO DE LETRAS
Programa de Pós-Graduação em
Estudos da Linguagem

Rio de Janeiro
Março de 2004

Aucione Das Dores Smarsaro

Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Letras do Departamento de Letras da PUC-Rio como parte dos requisitos parciais para obtenção do título de Doutor em Letras.

Orientador: Maria Carmelita Padua Dias (PUC-Rio)
Co-orientador: Eric Laporte (Univ. Marne-La-Vallé-Paris)

Rio de Janeiro
Março de 2004

Aucione Das Dores Smarsaro

Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo programa de Pós-graduação em Letras do Departamento de Letras do Centro de Teologia e Ciências Humanas da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Dra. Maria Carmelita Padua Dias (PUC-Rio)
Orientadiora
Departamento de Letras - PUC-Rio

Prof. Dr. Eric Laporte
Co-orientador
Univ. Marne-La-Vallé

Profa. Dra. Violeta de San Tiago Dantas Barbosa Quental
Departamento de Letras – PUC-Rio

Profa. Dra. Margarida Maria de Paula Basílio
Departamento de Letras – PUC-Rio

Prof. Dr. Carlos Alexandre Victorio Gonçalves
UFRJ

Prof. Dr. Oto Araújo Vale
UFRJ

Prof. Dr. Paulo Fernando Carneiro de Andrade
Coordenador Setorial do Centro de
Teologia e Ciências Humanas

Rio de Janeiro, de de 2004

Ficha catalográfica

Smarsaro, Aucione das Dores

Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico / Aucione Das Dores Smarsaro ; orientador: Maria Carmelita Padua Dias ; co-orientador: Eric Laporte. – Rio de Janeiro : PUC-Rio, Departamento de Letras, 2004.

154 f. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras.

Inclui referências bibliográficas

1. Letras – Teses. 2. Composição. 3. Linguistica computacional. 4. Dicionário eletrônico. 5. Léxico-gramática. I. Dias, Maria Carmelita Padua. II. Laporte, Eric. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. IV. Título.

CDD: 400

Agradecimentos

Inicialmente, agradeço à *Profa. Dra. Maria Carmelita Paduas Dias*, que tem me orientado desde o Mestrado, com observações críticas e sugestões sempre de forma muito respeitosa, sendo estímulo constante para a conclusão desta pesquisa.

Quero agradecer, também, de forma muito especial, a orientação do *Prof. Dr. Eric Laporte*, que, sem dúvida, é responsável pelos novos conhecimentos que adquiri ao longo desta pesquisa. As suas informações, críticas e sugestões sempre abriram novos horizontes.

Manifesto meus agradecimentos à *Profa. Dra. Margarida Basílio*, que por sua competência me motivou para os estudos do léxico.

Agradeço também, de forma muito expressiva, às *Profas. Dras. Violeta Quental* e *Helena F. Martins*, pelas sugestões na apresentação do exame de qualificação.

Não poderia deixar de agradecer, também, aos amigos e colegas de trabalho, *Prof. Santinho Ferreira de Souza* e *Lúcia Helena P. da Rocha*, que sempre disponibilizaram tempo e atenção para as leituras e trocas de idéias. Sem o apoio deles nem tudo teria o mesmo sentido.

À *Zuleme Maria da Cruz*, pelo apoio emocional. Amiga do coração.

A *Alcides Paredes*, pelo apoio incondicional em todos os sentidos.

Ao *Departamento de Línguas e Letras da Universidade Federal do Espírito Santo*, que me concedeu afastamento para realização do Curso de Doutorado.

À *CAPES*, pelo apoio financeiro.

Resumo

Smarsaro, Aucione Das Dores; Dias, Maria Carmelita Padua (Orientador). **Descrição de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico**. Rio de Janeiro, 2004. 154p. Tese (Doutorado) - Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ.

Neste trabalho estudam-se os nomes com a estrutura NdeN que podem ser descritos como nomes compostos por justaposição. São observadas 1.500 seqüências de palavras, com o objetivo de contribuir na descrição formal do léxico do português do Brasil e de definir os critérios de identificação de um nome composto com essa estrutura. O critério geral está baseado no conceito da não-composicionalidade semântica. Os testes são feitos a partir das propriedades sintáticas e semânticas que há na relação entre os elementos que constituem o grupo nominal, mostrando as distinções entre um grupo nominal livre e um grupo nominal composto. Entre as propriedades, podem ser destacadas: o bloqueio distribucional, a inseparabilidade, a inserção lexical, o apagamento de N2, a substituição de N2 e as variações em gênero e número. Essa descrição mostra-se útil na medida em que um conjunto de regras e critérios de delimitação de unidades lexicais foi definido, constituindo uma base para a incorporação de novos itens ao léxico. Por fim, as propriedades das entradas incorporadas receberam uma representação formal, resultando na criação de um dicionário eletrônico utilizável em processos eletrônicos.

Palavras-chave

Composição; léxico-gramática; lingüística computacional; dicionário eletrônico.

Abstract

Smarsaro, Aucione Das Dores; Dias, Maria Carmelita Padua (Advisor). **Description and formalization of compound word in Brazilian Portuguese for an electronic dictionary**. Rio de Janeiro, 2004. 154p. Tese (Doutorado) - Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ.

This paper is a study of the NofN structure nouns, which may be described as compound nouns by juxtaposition. 1500 word sequences are observed, aiming at contributing to the formal description of the Brazilian Portuguese lexicon, and defining the identification criteria of a compound noun with such a structure. The general criterion is based on the concept of semantic non-compositionality. The tests are made from the syntactic and semantic properties existing in the relationship between the elements that constitute the nominal group, showing the differences between a free nominal group and a compound nominal group. Among such properties, the following can be pointed out: distributional blockage, inseparability, lexical insertion, N2 erasing, N2 substitution, and gender and number variations. Such description proves to be useful in the sense that a set of lexical units delimitation rules and criteria has been defined, constituting a basis for the incorporation of new items to the lexicon. Finally, the incorporated entries' properties received a formal representation, which resulted in the creation of an electronic dictionary that can be used in electronic processes.

Keywords

Compounds; lexicon-grammar; computational linguistics; electronic dictionary.

Resumée

Smarsaro, Aucione Das Dores. Dias, Maria Carmelita Pádua (Directeur de thèse). **Description et formalisation de mots composés du brésilien en vue de l´élaboration d´un dictionnaire électronique**. Rio de Janeiro, 2004. 154p. Tese de Doutorado - Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ.

On étudie dans ce travail les noms formés par la structure NdeN qui caractérise un nom composé par juxtaposition. On a observé 1.500 séquences de mots, pour définir les critères d'identification d'un nom composé par cette structure. Le critère general est fondé sur le concept de la non compositionalité sémantique. Les tests ont été faits à partir des propriétés syntaxiques et sémantiques qu'il y a dans la relation existante entre les éléments qui constituent le groupe nominal, en démontrant les distinctions entre un groupe nominal libre et un groupe nominal composé. Parmi ces propriétés, on peut remarquer le bloquement distributionnel, l'inséparabilité, l'insertion lexical, l'effacement de N2, le remplacement de N2 et les variations en genre et en nombre. En examinant ces propriétés, on a pu observer qu'il y a des irrégularités dans la formation de ce procès. Cette description devient utile dans la mesure que cette reconnaissance peut être formulée et qu'un ensemble de règles et de critères de délimitation d'unités lexicales peut être defini, et que, par là, on peut aboutir à l'incorporation de nouveaux items lexicaux. Au bout, en tenant compte de la possibilité de formalisation, les mots composés peuvent être processés automatiquement dans les diccionnaires électroniques.

Mots-cles

Composition; lexicon-grammar; linguistique informatique; dictionnaire électronique.

Sumário