



Milena de Uzeda Garrão

**O CÓRPUS NÃO MENTE JAMAIS: SOBRE A IDENTIFICAÇÃO
E USO DE COMBINAÇÕES MULTIVOCABULARES DO
TIPO VERBO MAIS SINTAGMA NOMINAL**

Tese de Doutorado

Tese apresentada ao Departamento de Letras da Pontifícia Universidade Católica do Rio de Janeiro como requisito parcial para obtenção do título de Doutor em Letras (Estudos da Linguagem).

Orientador: Prof^a Doutora Maria Carmelita Pádua Dias

Rio de Janeiro
Março de 2006



Milena de Uzeda Garrão

O CÓRPUS NÃO MENTE JAMAIS: SOBRE A IDENTIFICAÇÃO E USO DE COMBINAÇÕES MULTIVOCABULARES DO TIPO VERBO MAIS SINTAGMA NOMINAL

Tese apresentada ao Departamento de Letras da Pontifícia Universidade Católica do Rio de Janeiro como requisito parcial para obtenção do título de Doutor em Letras (Estudos da Linguagem). Aprovada pela Comissão Examinadora abaixo assinada.

Profª Doutora Maria Carmelita Pádua Dias

Orientadora

Departamento de Letras — PUC-Rio

Profª Helena Franco Martins

Departamento de Letras — PUC-Rio

Profª Violeta de San Tiago Dantas

Barbosa Quental

Departamento de Letras — PUC-Rio

Profª Solange Coelho Vereza

Departamento de Letras Estrangeiras Modernas — UFF

Profª Rove Luiza de Oliveira Chishman

UNISINOS

Prof. Paulo Fernando Carneiro de Andrade

Coordenador Setorial do Centro de Teologia
e Ciências Humanas — PUC-Rio

Rio de Janeiro, 24 de Março de 2006

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização do autor, do orientador e da universidade.

Milena de Uzeda Garrão

Graduou-se em Letras (Tradução-português/inglês) pela PUC-Rio, em 1997. Obteve seu título de Mestre em Estudos da Linguagem pela mesma instituição em 2001, na área de Tradução Automática. Neste mesmo ano atuou como professora de Lingüística no Departamento de Estudos da Linguagem da UERJ. Em 2002, colaborou com a implementação do *CLIC* (Centro de Lingüística Computacional da PUC-Rio). Tem como principais interesses, a Lexicografia, a Tradução Automática e a Lingüística de Córpus.

Ficha Catalográfica

Garrão, Milena de Uzeda

O cópús não mente jamais : sobre a identificação e uso de combinações multivocabulares do tipo verbo mais sintagma nominal / Milena de Uzeda Garrão; orientadora: Maria Carmelita Pádua Dias. – Rio de Janeiro : PUC, Departamento de Letras, 2006.

124 f. ; 30 cm

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

Inclui referências bibliográficas.

1. Letras – Teses. 2. Combinações Multivocabulares. 3. Colocações Verbais. 4. Lexicografia de Córpus. 5. Semântica de Córpus. I. Dias, Maria Carmelita Pádua. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

A meus pais, Nani e Mel.

Agradecimentos

Aos meus pais e ao Ernani, pela ajuda preciosa com a nossa adorada Mel.

À professora Helena Martins, não somente pelos escritos e aulas inspiradores mas, principalmente, por ter me escoltado neste caminho teórico ainda não amplamente explorado dentro da Lingüística.

Aos amigos do **Clic**, sobretudo:

à professora Claudia Oliveira, do Instituto Militar de Engenharia, pela amizade e idealismo, e pela adaptação e realização computacional do Modelo de Espaço Vetorial aos fins propostos.

à professora Maria Claudia Freitas, pelas idéias compartilhadas e pela ajuda com a aplicação do Modelo.

ao Cícero Nogueira, Doutorando do Departamento de Informática da PUC-Rio, pela implementação do extrator V+SN e por ter me socorrido com a realização dos testes estatísticos.

Ao corpo docente da pós-graduação do Departamento de Letras da PUC-Rio, principalmente às professoras Margarida Basílio e Violeta Quental, pela clareza nos ensinamentos.

À Chiquinha, da secretaria de pós-graduação do Departamento de Letras da PUC-Rio, pela ajuda sempre pontual durante os quatro anos de curso.

À Capes, pela bolsa de estudos que me foi concedida.

À Banca Examinadora, pela sua excelência e pelos comentários e sugestões preciosos.

E especialmente, à professora Maria Carmelita Pádua Dias, pelo incentivo, pela amizade, mas sobretudo, por ter aberto mão de algumas de suas convicções teóricas para, mais uma vez, presentear-me com sua orientação sempre precisa e terna.

Resumo

Garrão, Milena de Uzeda; Dias, Maria Carmelita Pádua. **O corpus não mente jamais: sobre a identificação e uso de combinações multivocabulares do tipo *verbo mais sintagma nominal***. Rio de Janeiro, 2006. 124 p. Tese de Doutorado - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

Muitos estudos recentes sobre a identificação e uso de combinações multivocabulares (CMs) adotam uma perspectiva representacionista do significado da palavra. Este estudo propõe que é muito mais interessante identificar as CMs por um olhar não-representacionista. A metodologia proposta foi testada em CMs do tipo V+SN, um padrão bastante freqüente no português do Brasil (PB). Trata-se de uma análise estatística com base em corpus que pode ser resumida em três etapas: 1) corpus robusto do PB como base de análise, 2) aplicação de um teste estatístico ao corpus, a saber, teste de Logaritmo de Verossimilhança (Banerjee & Pedersen, 2003), para detecção das CMs mais freqüentes com padrão V+SN (como *tomar café*) e exclusão de co-ocorrências sintáticas aleatórias dos mesmos itens lexicais, 3) aplicação de Medidas de Similaridade (Baeza-Yates & Ribeiro-Neto, 1999) entre todos os parágrafos contendo uma certa CM (por exemplo, *fazer campanha*) e todos os parágrafos contendo o substantivo fora da CM (*campanha*). Esta última etapa foi utilizada para avaliar o grau de composicionalidade da CM. Pôde-se concluir que quanto maior a similaridade entre os parágrafos contendo a CM e os parágrafos contendo o substantivo fora da expressão, maior será o grau de composicionalidade da CM. Por essa razão, este estudo tem um impacto tanto teórico quanto prático para a semântica.

Palavras-chave

Combinações Multivocabulares; Colocações Verbais; Lexicografia de Corpus; Semântica de Corpus

Abstract

Garrão, Milena de Uzeda; Dias, Maria Carmelita Pádua (Advisor). **The corpus never lies: on the identification and use of multiword expressions of the pattern *verb plus noun phrase***. Rio de Janeiro, 2006.124 p. PhD Thesis - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

A considerable amount of recent researches on defining multi-word expressions' (MWE) phenomenon has an underlying representational framework of word meaning. In this study we claim that it is much more interesting to view MWE from a non-representational perspective. By choosing this path, we avoid the time-consuming and controversial human intuitions to MWE identification and definition. Our methodology was tested on Brazilian Portuguese verbal phrases of V+NP pattern. It is a statistically-based corpus analysis which could be summed up as the following three sequent steps: 1) robust linguistic corpora as output, 2) application of a probabilistic test to the corpora, namely Log Likelihood test (Banerjee & Pedersen, 2003), in order to spot the Portuguese MWEs of V+NP pattern (such as *tomar café*) and disregard casual syntactic and not otherwise motivated co-occurrences of the same lexical items, 3) application of Similarity Measures (Baeza-Yates & Ribeiro-Neto, 1999) between all the paragraphs containing a certain MWE and all the paragraphs containing its separate noun. This latter step is crucial to assess the MWE compositionality level. We conclude that the higher are the similarity measures between the MWE (such as *fazer campanha*) and its separate noun (*campanha*), the more compositional will be the MWE. Therefore, we believe that this work has both a practical and a theoretical impact to semantics.

Keywords

Multiword Expressions; Verbal Collocations; Corpus Lexicography, Corpus Semantics.

Sumário

1. Pulga atrás da orelha	12
1.1. Caracterização do problema	12
1.2. Desconfianças teóricas e caminhos alternativos	13
1.3. Objetivos	15
1.4. Organização	18
2. Combinação Multivocabular: da palavra como representação a seus desdobramentos teóricos	21
2.1. Sobre o significado e a representação de entidades extra-lingüísticas	22
2.1.1. Representacionismo na Lingüística	23
2.2. Neo-representacionismo e sua ascendência filosófica	26
2.2.1. Neo-representacionismo na Lingüística	29
2.3. As CMs sob os dois ângulos de representação	35
2.3.1. Multivocábulos e o representacionismo: a profusão de rótulos da semântica da inocência	36
2.3.2. Multivocábulos e o neo-representacionismo: sinais de difusão teórica	42
2.4. Discussão preliminar	45
3. Por um caminho não-representacionista para a detecção dos multivocábulos	48
3.1 A herança filosófica	48
3.2. Ecos do não-representacionismo na Lingüística e em PLN	51
3.3. A inevitabilidade do paradoxo do cópup	53
3.3.1. O cópup utilizado: CETENFolha	55
3.4. O teor estatístico do fenômeno lingüístico	56
3.4.1. Mãos à obra	57
3.5. A identificação das CMs	64

3.5.1. Testagem de hipóteses	65
3.5.1.1. O teste e a avaliação dos resultados	67
4. Composicionalidade com base em corpus	100
4.1. Passo-a-passo do método	102
4.2. Aferição do grau de composicionalidade das CMs	104
4.3. Avaliação dos resultados	110
5. Discussão, aplicação e trabalhos futuros	116
6. Referências Bibliográficas	121

Lista de Tabelas e Quadros

Tabela 1- freqüência absoluta dos 30 verbos mais recorrentes no corpus	60
Tabela 2 - freqüência de verbos mais recorrentes do corpus CETENFolha seguidos facultativamente de determinante e obrigatoriamente de um nome	62
Tabela 3 - freqüência de verbos seguidos facultativamente de determinante e obrigatoriamente de um nome posposto por marcas de pontuação ou advérbio	63
Tabela 4 - resultados com verbo <i>fazer</i>	107
Tabela 5 - resultados com o verbo <i>ter</i>	107
Tabela 6 - resultados com o verbo <i>dar</i>	108
Tabela 7 - resultados com o verbo <i>perder</i>	108
Tabela 8 - resultados com o verbo <i>usar</i>	109
Tabela 9 - resultados com o verbo <i>receber</i>	109
Tabela 10 - resultados com o verbo <i>deixar</i>	109
Tabela 11 - resultados com o verbo <i>tomar</i>	110
Tabela 12 - resultados com o verbo <i>ganhar</i>	110
Tabela 13 - resultados com o verbo <i>criar</i>	110
Tabela 14 - resumo qualitativo dos resultados	111
Quadro 1 - alguns microcontextos de <i>partido / tomar partido</i>	112
Quadro 2 - alguns microcontextos de <i>bandeira / dar bandeira</i>	112
Quadro 3 - alguns microcontextos de <i>banho/ tomar banho</i>	113
Quadro 4 - microcontextos de <i>camisinha / usar camisinha</i>	113
Quadro 5 - microcontextos de <i>café/ tomar café</i>	114

A coisa parece fácil:
o fora em torno do dentro,
o alto em cima do baixo.

Mas essa ordem serena
é coisa dura e avessa,
uma máquina perversa.

Para instaurar esse mundo
precisa a vontade mais crassa,
a desfaçatez de quem sempre
procura aquilo que acha.

Precisa de olhos sem trégua
e mãos cegas, abissais,
com dedos destros,
capazes de gestos antinaturais.

Paulo Henriques Britto
(*Trovar Claro*)