Pontifícia Universidade Católica
do Rio de Janeiro

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: Profª Doutora Maria Carmelita Pádua Dias

Rio de Janeiro
Março de 2006

Pontifícia Universidade Católica
do Rio de Janeiro

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

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 córpus 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 córpus 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 córpus que pode ser resumida em três etapas: 1) córpus robusto do PB como base de análise, 2) aplicação de um teste estatístico ao córpus, 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 Córpus; Semântica de Córpus

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

Lista de Tabelas e Quadros

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*)

1
Pulga atrás da orelha

1.1
Caracterização do problema

Na raiz deste projeto está a preocupação com a descrição do uso da língua, principalmente no que diz respeito às Combinações Multivocabulares *(*CMs, *MWE: multi-word expressions)*, um fenômeno lingüístico de grande impacto nas línguas do mundo. Por CMs queremos abranger as combinações de vocábulos recorrentes na língua; embora a expressão *multilexema* seja mais amplamente utilizada nessa área, optamos por *multivocábulo* em função do teor representacional e mental tipicamente associado ao termo *lexema*.

Sob uma ótica descritivista, o não tratamento sistemático das CMs pode vir a acarretar uma avaliação maquiada do uso da língua, comprometendo diretamente domínios da lingüística aplicada, como por exemplo, o ensino de português como segunda língua e de tradução, além de outros campos multidisciplinares, como o que nos serve de cenário neste estudo: a área de Processamento Automático de Linguagem Natural (ou PLN) também conhecida como Lingüística Computacional.

Um dos maiores entraves teóricos e metodológicos no tratamento sistemático de CMs em geral está no fato de elas se caracterizarem por uma alegada difusão de padrões semânticos e estruturais, como, por exemplo, diferentes níveis de opacidade semântica. Há um enorme número de CMs na língua que, além de resistir a uma análise sintática tradicional, também não se alinha aos casos normalmente rotulados na literatura como “semanticamente opacos”. No entanto, salvo algumas visões teóricas não-hegemônicas, decidir se uma combinação sintática qualquer é ou não uma CM depende dos já notórios testes que se baseiam em uma visão absoluta de composicionalidade semântica, a qual questionaremos no capítulo 2. Grande parte de estudos voltados à descrição de CMs vai buscar numa concepção *entitativa* do significado (cf. M. Gross, 1982; G, Gross, 1996; Ranchhod, 2002; Tagnin, 1999; Vale, 2002) a resposta para sua

definição. Sob este olhar, subentende-se que as expressões lingüísticas têm um significado a priori; o foco está, portanto, no significado das expressões lingüísticas (itens lexicais) e suas interrelações. Parte-se, então, de uma perspectiva representacionista do significado, em que se postula uma relação direta entre a expressão lingüística, ou a palavra, e um referente não-lingüístico, uma entidade. Já a forma através da qual esta relação é estabelecida varia entre as diferentes perspectivas representacionistas do significado, como será apresentado também no capítulo 2.

Constitui-se aí o primeiro problema para o pesquisador. Ele se depara com suas intuições, não raras vezes hesitantes, sobre o significado de cada palavra pertencente a uma CM. Tal problema se agrava quando não se está diante de uma composição nominal ou um nome composto, o qual freqüentemente nomeia um objeto até mesmo visível ou palpável (como *pó compacto* ou *alto falante)*, mas sim diante de uma expressão encabeçada por um verbo, muitas vezes descontígua, com todas as suas nuanças aspectuais e abstrações de atos, eventos, ou do que houver de mais intangível (como *dar (muito) mole, fazer (muito) tempo*). De fato, as CMs de base verbal resistem mais bravamente ao rótulo de *expressões fixas* justamente pela presença do aspecto verbal, algo que não aflige os lexicógrafos quando diante de uma CM de base nominal. A grande questão para a lexicografia é que as CMs de base verbal não são raras no Português Brasileiro (PB doravante) a ponto de prescindirem de uma incursão teórica mais profunda. Muito pelo contrário, trata-se de um fenômeno de peso não só no PB, como em outras línguas (cf. Guenthner & Blanco, 2004). E é por esta razão que resolvemos fazer deste estudo um palco para discutir os enfoques teóricos mais comuns em relação ao tratamento semântico de CMs de base verbal e para proposição de uma nova perspectiva que consiga equacionar minimamente problemas que parecem intocados por esses olhares teóricos hegemônicos.

1.2
Desconfianças teóricas e caminhos alternativos

Esta pesquisa foi motivada, essencialmente, por algumas desconfianças relativas a fundamentações teóricas que costumam guiar muitas pesquisas com o

objetivo de descrever CMs verbais nas línguas: i) a nossa primeira desconfiança se deve a uma perspectiva um tanto dogmática que cerca uma certa visão do significado presente em muitas tentativas de distinguir os tipos de co-ocorrências vocabulares nas línguas; ii) a nossa segunda desconfiança está relacionada ao peso que muitas destas perspectivas atribuem a um critério especulativo e intuitivo para interpretação semântica dos enunciados, que minimiza ou até mesmo desconsidera usos reais da língua; iii) a nossa terceira desconfiança recai sobre uma relativa soberania teórica atribuída à visão "sintaticocêntrica" da linguagem, isto é, a uma valorização e a um apego teórico à caracterização gerativista da criatividade do falante, baseada no que chamarei no capítulo 2 de semântica do cálculo; iv) a nossa última desconfiança, intimamente associada a todas as três supracitadas, está na alegada possibilidade de separação clara entre conhecimento enciclopédico e lingüístico, o que, grosso modo, pode ser resumido como aquilo que o falante sabe em decorrência de seu papel social e aquilo que tem internalizado como conhecimento lingüístico, respectivamente.

Em relação à primeira desconfiança, propomos ser teórica e metodologicamente frágil considerar que uma palavra carrega um significado atômico: as tentativas de rotulações semânticas de CMs de base verbal, descritas no capítulo 2, são uma boa medida para esta fragilidade. Sobre a segunda, julgamos ser no mínimo arriscado qualquer tipo de dedução do pesquisador sobre as possibilidades de ocorrência, de usos e de estruturas de CMs sem uma verificação em córpus. Os resultados de abordagens especulativas e intuitivas expostos também no capítulo 2 evidenciam esse risco. Sobre a terceira desconfiança, ressaltamos que uma abordagem empírica é capaz de demonstrar como o uso da língua se baseia em reutilização de sintagmas (algo que será evidenciado pelas medidas estatísticas de CMs relatadas no capítulo 3), e que é remota a possibilidade de esta reutilização depender do modelo do cálculo. Todos esses indícios nos levam a renunciar a possibilidade de separação clara entre conhecimento lingüístico e enciclopédico.

O que nos serviu de base para questionar as perspectivas expostas em i, ii, iii, talvez preponderantes na lingüística contemporânea, foi uma tentativa de descrição de um tipo de CM, a qual rotulamos como expressões cristalizadas do tipo *bater +SN* (relatada no capítulo 2) para a sua inclusão em um dicionário bilíngüe de um tradutor automático em Garrão, 2001. Muito embora o objetivo do

estudo fosse prático, eminentemente dedicado à dicionarização eletrônica, a nossa opção de descrição de CMs com base em um critério dedutivo, baseado na intuição de falante, pôde nos trazer algumas conclusões sintomáticas a respeito do caminho escolhido:

i) O córpus se mostrou rico em contra-exemplos daquilo que os testes geralmente utilizados para detectar uma CM de base verbal se diziam capazes de prever em relação ao fenômeno;

ii) A freqüência de CMs do tipo *bater+SN* não era tão significativa quanto outros tipos de padrões V+SN;

iii) O método utilizado se revelou oneroso e pouco preditivo.

1.3
Objetivos

Nosso objetivo principal é propor uma abordagem alternativa à identificação de CMs de base verbal, abrindo mão do conforto, ou desconforto, de uma visão representacionista do significado. Acreditamos ter razões suficientes para confiar nosso projeto a uma perspectiva teórica alheia à intuição semântica do observador, conforme demonstraremos no capítulo 3.

Esta investigação abarca o fenômeno com nítida ênfase na freqüência das co-ocorrências. Por isso, dependemos inevitavelmente da utilização de córpus para dar conta das falhas de intuições do pesquisador quanto ao que vem a ser uma combinação verbal freqüente na língua. Nossa escolha por uma abordagem a partir de córpus prioriza o não estabelecimento prévio de qualquer tipo de rotulação semântica das CMs do tipo V+SN mais freqüentes no PB. Embora o método com base em córpus não seja incontroverso, defenderemos a sua escolha mais esmiuçadamente no capítulo 3.

A opção pelo padrão de combinação V+SN não é aleatória. Além de haver poucos estudos que se dediquem de forma sistemática às combinações verbais (Tagnin, 1999 e Vale, 2001 são alguns deles), existe um tipo em particular, o padrão V+SN, que se destaca das outras combinações verbais recorrentes no PB tanto pela sua freqüência quanto pelos seus alegados sub-padrões semânticos. Ele abarca, por exemplo, um dos tipos de expressão com verbo leve — como *dar um*

susto (assustar), fazer um discurso (discursar) — um fenômeno reconhecido não só no PB e no Português Europeu (cf. Basílio, Dias & Martins, 1997; Neves, 1999; Salomão, 1990) como na língua inglesa, francesa, espanhola e alemã (cf. Guenthner & Blanco, 2004).

O termo verbo leve, ou verbo-suporte, recebe tais designações na Lingüística por ser considerado um elemento verbal pertencente a uma classe mais ou menos fechada de verbos que se combinam regularmente com nomes, atribuindo à expressão como um todo outros valores aspectuais. Embora o uso do adjetivo *leve* induza a uma visão representacionista do significado, uma vez que implica um esvaziamento de conteúdo semântico intrínseco ao verbo, optamos pelo seu uso para fins meramente metodológicos e descritivos.

Vale (2001) se dedica a diferentes padrões de CMs encabeçadas por verbos e faz um levantamento bibliográfico sobre o estudo de expressões com verbo leve encabeçadas por *ser, estar, ficar, fazer, ter* e *dar*, mas exclui essas expressões da sua análise. De acordo com nossa perspectiva teórica, entretanto, não podemos nos furtar a incluir tais expressões nos nossos dados, uma vez que o nosso objetivo primeiro é listar as CMs do tipo V+SN mais freqüentes no córpus, independentemente de seu perfil estrutural. Além do que, segundo uma das leis de George Zipf (1902-1950) — um teórico dos fenômenos estatísticos relacionados à linguagem e introdutor do Princípio do Menor Esforço (*the Principle of Least Effort)*[1], que segundo ele subjaz a toda a condição humana —, "quanto maior a freqüência de uma palavra ou morfema, maior será o número de combinações possíveis (grosso modo, compostos e formas morfologicamente complexas)" (Zipf, 1949).

Uma outra posição teórica relativamente inovadora desta pesquisa é a perspectiva semântica adotada, compatível com uma leitura não-representacionista para análise de CMs. Muito se especula sobre a importância da aplicação de teorias semânticas já existentes na lingüística para fins de PLN. Contrariamente, ressaltamos a importância de PLN e do córpus para avaliar as CMs encontradas. A explicitação dessa medida semântica será exposta no capítulo

[1] No seu livro *Human Behavior and the principle of least effort* (1949), Zipf defende que as pessoas agem de modo a minimizar seu índice médio de esforço possível. De acordo com sua teoria, o esforço do falante é conservado através da utilização de um vocabulário reduzido de palavras comuns e o esforço do ouvinte é minimizado através da utilização de um vocabulário extenso de palavras mais raras (tornando o discurso menos ambíguo). Para uma leitura mais aprofundada sobre as Leis de Zipf ver Manning & Schütze, 2003: cap. 1.

seguinte.

Em suma, no decorrer deste estudo pretendemos:

i) Questionar as visões mais recorrentes no tratamento do fenômeno multivocabular (capítulo 2);
ii) Demonstrar como é possível lidar com o fenômeno lingüístico, mais especificamente com o fenômeno de recorrência vocabular, sem lançar mão de representações semânticas a priori (capítulo 3);
iii) Utilizar um método estatístico — o logaritmo de verossimilhança (Banerjee & Pedersen, 2003)— ao invés da intuição do pesquisador para a identificação das CMs do tipo V+SN (capítulo 3);
iv) Estabelecer as CMs com padrão V+SN mais freqüentes no PB, através de um recurso estatístico utilizado para detectar este padrão (também descrito no capítulo 3);
v) Propor uma nova medida de composicionalidade semântica destas CMs a partir de técnicas de Recuperação de Informações; isto é, poder avaliar o nível de composicionalidade/opacidade semântica de uma CM com base numa medida de similaridade entre contextos (ou parágrafos) através da implementação do Modelo Espacial Vetorial (Baeza-Yates & Ribeiro-Neto, 1999). Quanto maior a similaridade entre os contextos que apresentam, por exemplo, a CM *fazer amigos* e os contextos que apresentam o nome *amigos* fora da CM, maior será o nível de composicionalidade da CM. Quanto menor for esta medida (como, por exemplo, *tomar partido* e *partido*) mais opaca a CM será e mais polissêmico tenderá a ser o nome (o SN) que compõe a CM. (capítulo 4)

vi) Dentre outros fins, contribuir para a lexicografia, e mais especificamente para uma lexicografia quantitativa, uma vez que oferecemos uma lista das CMs do tipo V+SN mais freqüentes do PB. Auxiliar, conseqüentemente, o ensino de português para estrangeiros, tendo em vista que seria muito mais produtivo para um falante estrangeiro ter domínio das CMs mais freqüentes de uma segunda língua do que aquelas mais esporádicas. Auxiliar domínios de PLN, quais sejam a Tradução Automática, a Recuperação de Informações, por ser

um método relativamente preditivo em relação ao uso das CMs.

Nossa abordagem empírica, portanto, tem por objetivo aprender automaticamente preferências estruturais com base em córpus. Há um claro investimento nas relações entre palavras e o que se pode depreender de grupamentos vocabulares específicos. Segundo Manning & Schütze (2003), modelos estatísticos são robustos, têm bom poder de generalização e se comportam de forma elegante diante de erros ou de novos dados. São métodos de grande valia para resolver também problemas de ambigüidade (cf. Aranha, Freitas, Dias & Passos, 2004).

Num primeiro momento, este caminho não-representacionista pode parecer improdutivo ou conflitante com uma preocupação lexicográfica. Contrariamente, consideramos que através dele nos desviamos de alguns questionamentos sobre o significado que acabam retardando resultados no domínio semântico. Concordamos com Martins (1999, cap.3), quando defende que tal visão pode constituir um ângulo fértil para "um estudo sistemático empírico sobre os usos dos signos nas línguas do mundo". Ao abrir mão de uma posição representacionista, portanto, não estamos assumindo a inexistência dos significados; apenas ratificando a resistência de separação entre significado e uso.

1.4
Organização

O capítulo 2 da tese é destinado às considerações teóricas mais recorrentes sobre o significado de um modo geral e sobre o fenômeno multivocabular de base verbal, em particular. Apresenta, criticamente, alguns olhares distintos sobre o fenômeno multivocabular nitidamente imbuídos da semântica do cálculo; e em seguida volta sua crítica para olhares de inspiração cognitivista. No final desse capítulo, fazemos um balanço dessas duas visões representacionistas em relação a CMs e avaliamos minimamente seu custo teórico.

Apresentamos, no capítulo 3, a pertinência de uma visão de significado que abre mão do tipo de representação presente nas teorias expostas no capítulo 2 e demonstramos como esse ponto de vista é bem-vindo no domínio do fenômeno multivocabular; finalmente, tentamos evidenciar como essa perspectiva se alinha

a uma abordagem com base em córpus.

Já o conteúdo de aplicação prática deste estudo é apresentado na segunda metade do capítulo 3 e no capítulo 4. No capítulo 3 aplicamos o teste estatístico para detecção de CMs do tipo V + SN, que nos foi disponibilizado através do pacote estatístico NSP (Banerjee & Pedersen, 2003). Aliado a ele está um programa feito em linguagem Java™, que recebe como entrada o córpus e fornece como resultado a lista de CMs do tipo V+SN em ordem de ocorrência (Nogueira, 2004). Só a partir de então, é estabelecida a lista dessas CMs que, posteriormente, são ordenadas por freqüência. Um ponto em favor desse teste é o fato de ele ser capaz de detectar um determinante entre o verbo e o nome da construção, um aspecto relevante no estudo das CMs do tipo V+SN (ex: *fazer muito tempo, dar um susto*). Esta primeira etapa irá identificar as 100 CMs mais freqüentes de cada um dos 10 verbos mais recorrentes na estrutura V+SN no córpus; um corte meramente metodológico.

No capítulo 4 nos dedicamos a uma medida de composicionalidade semântica de algumas das CMs identificadas por cada um dos testes aplicados para os 10 verbos. Propomos uma análise do nível de composicionalidade das CMs a partir de uma técnica utilizada no domínio computacional de Recuperação de Informação. A nossa proposta aqui é atribuir ao córpus, mais especificamente a uma Medida de Similaridade (SM, *similarity measure*) entre os contextos (ou parágrafos) em que a CM ocorre, a responsabilidade de avaliar o nível de composicionalidade semântica da expressão (cf. Baeza-Yates & Ribeiro-Neto, 1999, cap. 2).

Além de eliminar o risco da avaliação semântica especulativa do pesquisador, esse recurso permite também a detecção do grau de polissemia dos SNs que aparecem nas CMs. Segundo Aranha, Freitas, Dias e Passos (2004), "palavras com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes". Enfatizaremos aqui um tipo mais restrito de contexto: o microcontexto, ou seja, o parágrafo em que a CM ocorre no córpus. Nossa proposta é a de que o grau de transparência semântica da CM é proporcional ao aumento do grau de similaridade entre os parágrafos contendo a CM e os parágrafos contendo somente o SN presente na CM.Trata-se, portanto, de uma medida de composicionalidade de base empírica.

No capítulo final traçamos algumas conclusões sobre o fenômeno multivocabular para o domínio semântico e fazemos um balanço sobre a relevância do método inaugurado neste estudo tanto para a lexicografia de um modo geral quanto para o domínio de PLN.

2
Combinação Multivocabular: da palavra como representação a seus desdobramentos teóricos

> A semântica constitui para a Lingüística o ponto nodal para todas as suas contradições, porque é nesse ponto, e mais freqüentemente sem reconhecê-lo, que a Lingüística tem a ver com a Filosofia.
>
> Michel Pêcheux

Intuímos que prioritária à defesa de uma perspectiva não-representacionista do significado para lidar com as CMs seja a exposição de algumas caracterizações teóricas acerca do fenômeno multivocabular que, consciente ou mesmo inadvertidamente, optam por um olhar representacionista do significado. O nosso empenho, neste capítulo, é em demonstrar como há um custo teórico nessas perspectivas que, de uma forma bastante generalizada, podem ser divididas em representacionistas e neo-representacionistas (cf. Martins: 1999).

Há três aspectos que dividem águas na teoria semântica e que trazem um impacto direto para o tratamento das CMs: a possibilidade de distinção entre os aspectos literal e figurativo da linguagem; a possibilidade de distinção entre conhecimento lingüístico e conhecimento enciclopédico, e a visão de composicionalidade forte versus composicionalidade fraca do significado. Veremos nesta seção que essas três dicotomias, embora didaticamente bem-vindas, não são tão bem-delineadas e rígidas quanto se pode supor, e que mesmo se focarmos somente um de seus lados seremos capazes de traçar diferentes ângulos sobre a questão do significado.

Talvez fosse prudente reservar um espaço nesta seção para um breve histórico sobre o conceito de palavra antes de nos aventurarmos a traçar qualquer consideração sobre o termo *significado*. Como bem observa Steven Pinker em seu *The Language Instinct, "a word in a word is complicated. But then what in the world is a word?*".[2] Optamos por não enredar por esse caminho, entretanto, pois

[2] Trata-se, na verdade, de uma frase que brinca com os usos de *word* ("palavra") além de fazer um trocadilho com a palavra *world* ("mundo"). Poderíamos arriscar a seguinte tradução: *"Transmitir o significado de 'palavra' em uma palavra é complicado. Em outras palavras, o que é, afinal, uma palavra?"*

imaginamos que não só fugiria às delimitações metodológicas deste estudo, como também correríamos o risco de reproduzir desnecessariamente uma vasta e diversificada bibliografia sobre o assunto, incluindo os seus vários recortes teóricos e comprometimentos metodológicos (cf. Garrão, 2001: cap. 2). Para uma análise mais aprofundada do tema, Cruse, 1986: caps1-3, Jackendoff, 1997: cap 7, Basílio 1999 e Biderman 1999 são alguns exemplos de textos elucidativos. Para uma leitura mais fluente desta tese é essencial ter em mente que a utilização do termo *palavra* refere-se ao ponto de vista gráfico; como vocábulo ou unidade constituída por grafemas (a menor unidade contrastiva num sistema de escrita), delimitada por espaços em branco e/ou sinais de pontuação.

2.1
Sobre o significado e a representação de entidades extra-lingüísticas

> Cada coisa tem por natureza um nome apropriado.
>
> Platão, Diálogos de Crátilo 383-a

Para uma apresentação menos superficial da visão representacionista do significado na Lingüística, não podemos prescindir de expor ao menos uma síntese da sua ascendência filosófica. Resumidamente, o comprometimento da linguagem com a realidade ou com sua função declarativa sobre o mundo guia filósofos de influência incontestável no pensamento ocidental, como Platão, Aristóteles e Locke[3]. A certeza de que a linguagem representa algo que lhe é exterior e de que o significado é uma entidade estável encarregam-se de alijar, por exemplo, as suas manifestações não declarativas e não literais pelos embaraços que trazem para a visão representacionista.

Com efeito, a oposição entre significado literal e figurativo se apresenta em virtude de uma compreensão do significado como entidade. A certeza de que é possível através da linguagem estabelecer relações de verdade sobre as coisas impulsiona esses pensadores a se aterem ao campo da literalidade. Um discurso

[3] Para uma leitura menos abreviada sobre o assunto ver Martins, 1999: cap.1.

figurativo, portanto, seria identificado com falta de clareza e imprecisão; careceria do poder epistemológico presente somente no discurso literal. Podemos dizer com certa segurança que o papel acessório que, tradicionalmente, vem sendo reservado à figuratividade na lingüística tem uma hegemonia histórica no pensamento filosófico. Afinal, é de Aristóteles a frase "*a metáfora consiste em dar à coisa um nome que pertence a outra coisa"*. E uma vez que a ciência e todo o pensamento ocidental são, em grande parte, herdeiros das suas considerações filosóficas, não é de se estranhar que o mesmo seja dito por muitos cientistas contemporâneos da área da linguagem, como veremos na seção abaixo.

2.1.1
Representacionismo na Lingüística

> ...um nome é puro símbolo, parte de um elenco de milhares, rapidamente adquiridos graças à harmonia entre a mente de uma criança, a de um adulto e a textura da realidade.
>
> Steven Pinker, *The Language Instinct*

Inspirada pelo pensamento filosófico de que a linguagem tem eminentemente uma função representativa, grande parte das teorias lingüísticas também exclui do seu campo de estudo qualquer modalidade de linguagem que não representa de fato "a coisa em si". Ou seja, para se construir uma ciência da linguagem há de se excluir todo tipo de expressão lingüística que foge à representação da realidade. Chomsky assim como seus legatários edificaram suas teorias neste **pressuposto universalista.** Toda manifestação lingüística não-literal, ou seja, não-representativa, deixa de ter importância do ponto de vista científico. Bastaria, então, traçar o limite entre significado literal e figurado. E é nesta, aparentemente infindável, tentativa de divisão desses dois campos de significação que a maior parte dos estudos semânticos se situa. De fato, isto tem sido, se não a maior, uma das grandes interrogações da Lingüística.

É bem verdade que mesmo os defensores da chamada **visão literalista** admitem que a questão não é trivial: Ruth Kempson (1995: 73) pondera que, entre o literal e o metafórico, "há um *número grande* de casos duvidosos". J. Sadock

(1993:42), por sua vez, afirma que os princípios subjacentes à metáfora são psicológicos e não estritamente lingüísticos e que, portanto, devem estar fora do escopo da lingüística sincrônica. Sadock, contudo, admite que "em *inúmeros casos* é difícil determinar onde começa o sentido figurado e onde termina o sentido literal" (p.48). No entanto, as incontáveis fronteiras nebulosas não parecem ter dado origem a questionamentos do pressuposto teórico, mas sim, como avalia criticamente Martins (1999:55), apenas consideradas como "percalços naturais do fazer científico".

Uma das grandes conseqüências sintomáticas do conflito literal versus figurado é a inevitável tensão entre homonímia e polissemia. Martins explica que quando há uma resistência para o estabelecimento da fronteira entre literal e figurado é comum entre os literalistas a utilização de duas estratégias: (a) redução do escopo da análise de modo a tentar excluir a variação polissêmica e (b) conversão dos casos de polissemia em casos de homonímia. Tomemos, a título de ilustração dessa tensão, a lista de frases abaixo:

(2.1) Ele *tirou* a camisa.
(2.2) Ele *tirou* a camisa do armário.
(2.3) Ele *tirou* 10 na prova.
(2.4) Ele *tirou* aquela idéia da cabeça.
(2.5) Ele *tirou* o corpo fora.

Dentro de uma concepção literalista do significado, poderíamos supor que somente as frases 2.1), 2.2) e 2.3) seriam consideradas instâncias de fragmentos lingüísticos passíveis de análise lingüística, onde 2.1) e 2.2) apresentariam significados análogos do verbo *tirar* com transitividades distintas e 2.3) talvez fosse equacionada através da estratégia descrita em (b); isto é, seria um verbo distinto de 2.1) e 2.2) apresentando apenas a mesma forma fonológica. Mas o que fazer, então, com 2.4) e 2.5)? A rigor, a estratégia de homonímia também poderia ser aplicada em 2.4). Ou seja, embora 2.4) tenha a mesma estrutura sintática de 2.2), muito provavelmente deixaria de ser considerada analisável dentro da mesma entrada lexical, uma vez que o uso do verbo não é literal. Isto é, por uma perspectiva tradicional, não é possível tirar "literalmente" uma idéia da cabeça. Em 2.5) teríamos ainda o agravante de haver uma aberração estrutural. Por ser uma frase sintaticamente anômala, tal estrutura só poderia ser analisável por uma

perspectiva literalista como uma expressão fixa; com sentido unitário, indivisível. Conclui-se, portanto, que uma frase corriqueira do PB como *pode tirar o seu cavalinho da chuva,* seria interpretada como a retirada de um mamífero quadrúpede da precipitação atmosférica ou como um segmento indivisível em que tem sentido análogo a *Desista*! Esta última interpretação, entretanto, estaria fora do domínio sintático-semântico e seria atribuída a um campo do estudo lingüístico fronteiriço com a sociologia e a psicologia: a pragmática. Seria um exemplo de **conhecimento enciclopédico** do PB.

Intrínseco a esse literalismo presente na visão representacionista do significado está uma **perspectiva imanentista**, ou a idéia de que a entidade significado literal é auto-evidente; nas palavras de Martins (1999:55):

> A opção pela exclusão teórica da metáfora articula-se, então, com o projeto de alcançar aquilo que há de estável e previsível nos significados: a entidade *significado* é, em suma, concebida como equivalente à entidade *significado literal*, presumida como algo, pelo menos até certo ponto, evidente em si mesmo.

A certeza de que o significado literal é auto-evidente viabiliza uma outra característica teórica dessa linha de pensamento, de extrema relevância para a sua concepção de CM, como veremos na seção 2.3: a **visão composicional do significado**. Ao que tudo indica, o termo *composicionalidade* foi cunhado por Katz & Fodor (1963) para dar conta da capacidade semântica do usuário de uma língua. É um modelo de interpretação e produção semântica baseado em cálculo: cada fragmento lingüístico, ou cada palavra, contribui para o significado total da frase. Ou seja, o significado de uma expressão lingüística maior é calculado através do conhecimento do significado das suas partes. Cada vez que uma estrutura ou frase reaparece, é calculada novamente.

Mais tarde, Searle (1978) equacionou a noção de semântica composicional postulando a sua dependência do "sentido literal". Considerou que "o sentido literal de uma frase é inteiramente determinado pelos significados de suas palavras componentes (ou morfemas) e pelas regras sintáticas de acordo com as quais esses elementos se combinam" (p.207). Esse modelo de cálculo fundamenta a análise semântica representacionista. Por essa visão, os exemplos 2.4) e 2.5) não podem ser analisados porque seus itens não são contributos para o significado

total das expressões a que pertencem. Esse tipo de construção passa a ser considerado, portanto, como semanticamente indivisível. A seção 2.3 apresenta mais minuciosamente as implicações teóricas e práticas desse modelo do cálculo.

Em suma, podemos dizer que, por esse olhar:

i) O significado é uma propriedade exclusiva das expressões lingüísticas e se define em termos de referência e verdade assim como independem dos indivíduos que o produzem;
ii) É possível estabelecer o potencial referencial da linguagem; ou seja, como os símbolos se relacionam com a realidade;
iii) Deve-se ater exclusivamente ao domínio literal, não enciclopédico, não pragmático, declarativo e sincrônico;
iv) A linguagem tem prioritariamente a função de expressar (racionalmente) o mundo e, portanto, é possível estabelecer o valor de verdade das sentenças.

Antes de apresentar as conseqüências do estudo de CMs segundo esse olhar teórico, destinamos minhas considerações ao neo-representacionismo e à sua base filosófica. Finalmente, na seção 2.3, partimos para uma apreciação do fenômeno multivocabular de acordo com as duas perspectivas em análise, com ênfase nas implicações teóricas e práticas desse sob um e outro ângulo.

2.2.
Neo-representacionismo e sua ascendência filosófica

> Não possuímos nada mais do que metáforas das coisas, que de nenhum modo correspondem às entidades de origem
>
> Friedrich Nietzsche, *Sobre verdade e mentira no sentido extra-moral*

A nova forma de pensar a representação, ou o neo-representacionismo (cf. Martins, 1999), está parcialmente fundamentada nas considerações filosóficas anti-representacionistas de Friedrich Nietzsche e Ludwig Wittgenstein, dois dos

pensadores mais representativos do século XIX e XX, respectivamente, e críticos afiados à filosofia tradicional.

Enquanto autores como Platão, Locke e Aristóteles apresentam as palavras como representantes de entidades, Nietzsche parece ser o anunciador de que esse pressuposto que mobilizou todo o ocidente não passou de um "placebo" epistemológico. Seu alerta sobre esse pseudoconhecimento indica que a linguagem falsifica a existência de conceitos, que o conhecimento foi inventado, ou melhor, forjado pelo homem e que esse foi o "momento mais soberbo e mentiroso da história universal" (Nietzsche, 1978: 45). A linguagem estaria, então, subjugada a tal invenção. Conseqüentemente, não possuiria nenhum potencial epistemológico:

> O que é uma palavra? A figuração de um estímulo nervoso em sons. Mas concluir do estímulo nervoso uma causa fora de nós já é resultado de uma aplicação falsa e ilegítima do princípio da razão (...). A 'coisa em si' (tal seria justamente a verdade pura sem conseqüências) é, também para o formador da linguagem, inteiramente incaptável e nem sequer algo que vale a pena. (ibd.: 47)

Para Nietzsche, a língua é um grande depósito de metáforas e metonímias e o impulso à verdade se origina da necessidade de evitar "a guerra de todos contra todos": um "acordo de paz" (ibid:46) do qual a humanidade não consegue se desvencilhar. Diferentemente de Nietzsche, entretanto, o introdutor do neo-representacionismo não considera que a língua seja totalmente, mas sim, eminentemente metafórica: "*Although I will argue that a great many concepts like causation and purpose are metaphorical, there is nonetheless an extensive range of non metaphorical concepts. A sentence like 'the balloon went up' is not metaphorical…*"[4] (Lakoff, 1993: 205)

De Wittgenstein, o neo-representacionismo herda, entre outras características, a crítica à idéia aristotélica de que as categorias seriam universais. Em *Women, fire and dangerous things* (1987), Lakoff argumenta que, desde

[4] "Embora eu defenda que diversos conceitos como causa e finalidade sejam metafóricos, em contrapartida, há uma vastidão de conceitos não-metafóricos. Uma frase como 'o balão subiu' não é metafórica."

Aristóteles até o 1º. Wittgenstein[5], as categorias — substância, quantidade, qualidade, relação, lugar, tempo, situação ou postura, condição, ação, passividade — eram consideradas verdades absolutas, universais e base de qualquer conhecimento. Para Lakoff, não havia um estudo científico sobre as categorizações, já que estas eram uma pressuposição. As coisas pertenciam à mesma categoria se tivessem propriedades em comum:

> *This classical theory was not the result of empirical study* [...] *It was a philosophical position arrived at on the basis of a priori speculation* [...] *Over the centuries it simply became part of the background assumptions taken for granted in most scholarly disciplines*[6]. (Lakoff, 1987: 6)

Já o segundo Wittgenstein, e sua virada lingüística marcada por *Investigações filosóficas* (1979 [1953]), condena a idéia aristotélica de que as categorias são definidas pelas propriedades comuns a todos os seus membros e argumenta que os membros de uma categoria compartilham "semelhança de família", somente (cf. Glock,1996:154-155).

A idéia wittgensteiniana de que a linguagem não é uma representação da realidade ou um instrumento de mediação entre sujeito e realidade, mas sim um fenômeno eminentemente intersubjetivo também é, em alguma medida, advogada pelo neo-representacionismo. De fato, as considerações de Wittgenstein sobre a linguagem são bastante mais profundas e impactantes e vão muito mais além do que as duas características supracitadas. Voltaremos a alguns de seus pressupostos anti-representacionistas na primeira seção do capítulo 3. Esses são apenas argumentos comumente utilizados entre os cognitivistas para refutar a visão universalista, como veremos a seguir.

---

[5] Há dois momentos filosóficos distintos nos textos de Wittgenstein. Nesta tese, volto atenção somente à chamada "virada lingüística", onde está, a meu ver, a sua grande contribuição para o pensamento filosófico.

[6] "Essa teoria clássica não foi resultado de investigação empírica[…]Foi uma posição filosófica estabelecida através de especulação *a priori*[...]Com o passar dos séculos, simplesmente tornou-se parte das suposições tidas como verdadeiras na grande maioria das disciplinas acadêmicas."

2.2.1.
Neo-representacionismo na Lingüística

> *Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.*
>
> Lakoff & Johnson, *Metaphors We Live By*

Após um resumo mais do que abreviado sobre a crítica ao pensamento filosófico tradicional, voltamo-nos agora à sua ressonância na visão lingüística proposta por Lakoff. Segundo a sua proposta, os sistemas conceituais variam, ou seja, não haveria uma visão de mundo unicamente verdadeira; a gramática não seria pura forma; a emoção teria conteúdo conceitual; a mente não seria uma dimensão independente do corpo, e o significado não seria baseado em verdade e referência ou na relação entre os símbolos e as coisas no mundo.

Lakoff introduz o paradigma do **experiencialismo** baseando-se tanto nos aspectos imaginativos quanto nos aspectos corporais da cognição. Argumenta que a visão essencialista da linguagem está equivocada ao afirmar que as categorias são universais e que existe uma linguagem do pensamento capaz de captar essas categorias uniformemente (o "mentalês", cf. Pinker, 1995, cap. 3). Para ele, as categorias lingüísticas são claramente reflexos da experiência, da imaginação e do corpo.

Daí, pode-se também concluir que grande parte do legado da lingüística antropológica norte-americana pode ser encontrada nas idéias neo-representacionistas. Em *Handbook of American Indian Languages* (1911), o antropólogo Franz Boas já trazia uma vasta contribuição de diferenças entre estruturas gramaticais e lexicais entre as línguas[7]. Outro representante da visão antropológica da linguagem, Benjamim Lee Whorf (1998 [1957]), fornece um exemplo ainda mais contundente para defender o relativismo lingüístico. Seu

[7] É de sua autoria o atualmente corriqueiro exemplo dos vários significantes para a palavra *neve* em esquimó, que ilustra uma diferença de estrutura lexical entre as línguas. Já para ilustrar diferenças gramaticais, Boas demonstrou como uma frase simples, como por exemplo, *The man is sick* traduzida para três línguas indígenas americanas (kuaikutl, esquimó e ponca) obrigaria o tradutor a acrescentar informações à frase, como, por exemplo, se o informante pode garantir a informação por ter presenciado o fato, ou se confiou no que ouviu falar, ou até mesmo se o sujeito da frase está deitado ou visível ao falante.

relato sobre a cultura Hopi nos convida (ou nos obriga) a refletir sobre algo que nem concebemos ser visto de outra forma, ou que pressupomos ser universalmente percebido como um fluxo contínuo: a concepção de tempo. Lançando um olhar extremamente relativista sobre a linguagem, Whorf desestabiliza o senso comum ocidental ao relatar a concepção de tempo dessa cultura, onde, dentre outros "devaneios" que desafiam a metafísica ocidental, não há uma fronteira para a distinção temporal passado-presente-futuro.

Seu relato poria em xeque as categorias universais aristotélicas: "*Discrete are number and language;* ***continuous*** *are lines, surface, bodies and also, besides these,* ***time*** *and place"*. (Aristóteles, 2000: 8; grifo nosso). Contrariamente, para o determinismo lingüístico de Whorf, o pensamento passa a ser determinado pelas categorias disponíveis na língua, uma posição teórica radicalmente refutada por estudiosos universalistas (cf. Pinker, 1995: 59-67) e pouco apoiada até mesmo por partidários do relativismo lingüístico. George Lakoff, portanto, acomoda as evidências de Whorf em sua teoria, mas tenta fugir do extremismo presente na idéia de determinismo lingüístico. Na verdade, não traça um limite muito claro entre universal e cultural; parece estar se situando numa interseção entre o determinismo lingüístico e o universalismo do significado. Parece apostar, a nosso ver, não em um "mentalês" de Pinker nem em um "culturalês" de Whorf, mas talvez num ocidentalês: um padrão conceitual herdado por culturas análogas.

Mas talvez a outra base teórica do neo-representacionismo, proposta em co-autoria com Mark Johnson — **a Teoria Cognitiva da Metáfora** —, seja especialmente cara ao presente estudo, uma vez que oferece explicações contundentes sobre a formação de metáforas, o que, de certa forma, responde a algumas questões da Semântica, dentre elas, a formação de CMs.

Segundo a visão representacionista, as metáforas são simples expressões lingüísticas fora do escopo da linguagem corriqueira, assim como as expressões idiomáticas (tradicionalmente consideradas um tipo de CMs) têm sentido arbitrário. Lakoff & Johnson rebatem esse senso comum através da argumentação de que se a metáfora fosse uma mera expressão lingüística não haveria tanta evidência (cf. Lakoff & Johnson, 1980) apontando para um mapeamento entre domínios específicos, o que demonstra uma restrição conceitual para a construção do fenômeno.

Os autores explicam que a construção metafórica seria o mapeamento entre um domínio-fonte e um domínio-alvo, que compartilhariam correspondências ontológicas sistemáticas (A VIDA É UMA JORNADA, DISCUSSÕES SÃO GUERRAS, O CORPO É UM *CONTAINER*, dentre outros)[8]. Esses mapeamentos obedeceriam ao Princípio da Invariância, um princípio que caracteriza a classe de analogias possíveis. Trata-se de um mapeamento convencional, isto é, uma parte fixa do nosso sistema conceitual. Não deve ser, portanto, caracterizado como processos ou algoritmos que transformam automaticamente o input do domínio-fonte em output do domínio-alvo. Tal caracterização também se prestaria a dar conta da polissemia, um outro grande problema dentro de estudos semânticos representacionistas, conforme apresentado na seção 2.1.1.

Para explicar o mapeamento entre domínio-fonte e domínio-alvo é necessário considerar que esses compartilham uma "estrutura de nível geral". E por ser convencional, a metáfora é usada constantemente e automaticamente, sem esforço ou consciência. Lakoff, de fato, responde a uma pergunta constante dentro de estudos sobre o significado. Por que nossas metáforas são como são? Porque, diria ele, seriam baseadas na experiência. E a experiência motiva a metáfora, não a determina. Por exemplo, o mapeamento conceitual MAIS–EM CIMA/ MENOS–EMBAIXO (que viabiliza expressões como *ter alto astral, estar meio cabisbaixo*) prediz que simplesmente será difícil numa língua haver a correspondência "menos-em cima" e que o falante de uma língua que não tenha nenhum desses mapeamentos irá apreender a correspondência MAIS–EM CIMA com muito mais facilidade do que aquela improvável (Lakoff, 1993).

Os autores explicam, por outro lado, que os mapeamentos TEMPO É DINHEIRO, TEMPO É UM RECURSO LIMITADO e TEMPO É UM BEM DE VALOR — formadores de metáforas como y*ou are wasting time/ você está desperdiçando tempo* — seriam válidos somente em culturas que enxergam o tempo da forma como o fazemos. "Isso não é uma necessidade humana de

[8] Em Lakoff & Johnson 2002, *Philosophy in the flesh: the embodied mind and its challenge to Western Thought*, tal abordagem bidominial cede espaço a uma teoria que envolve quatro domínios — a teoria da mesclagem —, segundo a qual nosso cérebro adquire informação do resto do nosso corpo. A relação do corpo com o mundo estrutura os conceitos que usamos para pensar. Não podemos pensar o que queremos, somente aquilo disponibilizado pelo nosso cérebro. Nessa obra, há um incremento dos mapeamentos, mas o conceito cognitivo lançado nos anos 80 parece persistir.

conceituar o tempo, está ligado à nossa cultura. Há culturas em que não há nada dessas coisas" (Lakoff, 1980: 9).

Ao que tudo indica, chegamos a um impasse teórico. O mapeamento metafórico MAIS-EM CIMA seria um princípio candidato a universal e o mapeamento TEMPO É DINHEIRO seria então cultural. Mas como Lakoff traça esse limite? Onde encontramos a diferença formal entre experiência física e experiência cultural? Se as metáforas são baseadas na experiência, por que não pode haver, portanto, metáforas do tipo MENOS –EM CIMA em uma cultura como a Hopi, uma vez que talvez até a noção de tempo (ao menos seguindo o relato de Whorf) seria completamente diferente da nossa?

O fato de a teoria conjugar no modelo domínios diversos como experiência direta, mente, linguagem, história e cultura parece comprometer, de certa forma, a sua formalização. Lakoff (1987), talvez já antecipando essa indagação por parte do leitor, argumenta que "o rigor teórico e a precisão não são deixados de lado. Apenas serão caracterizados de outra forma" (Lakoff, 1987: 9). Martins (1999) credita ao jogo entre esses variados domínios o foco do problema no modelo:

> A experiência direta que alegadamente engendraria a emergência de tais conceitos fundadores recebe uma caracterização um tanto vaga, sobretudo no que diz respeito a seu grau de universalidade. Lakoff & Johnson admitem, por exemplo, que toda experiência *é*, em certo sentido, *cultural*, ressalvando, porém, que "se pode fazer uma distinção importante entre as experiências que são 'mais' físicas (como levantar) e aquelas que são 'mais' culturais (como participar de uma cerimônia de casamento)" (Lakoff & Johnson 1980:57). Não esclarecem, contudo, em que "certo sentido" as experiências "mais físicas" (*diretas*) são também determinadas por fatores culturais, e tampouco nos fornecem critérios bem definidos para decidir o que faz com que uma experiência deixe de ser "mais física" e passe a ser "mais cultural, dando-nos a impressão de que esse esclarecimento talvez não seja importante para o modelo como um todo." (Martins, 1999: 90)

O modelo também não expõe claramente em que medida viabiliza a formação de metáforas novas e metáforas irregulares. Portanto, concomitantemente à clareza intuitiva presente nos mapeamentos conceituais expostos por Lakoff & Johnson, constata-se uma certa omissão de explicações daquilo que foge aos tipos de mapeamentos previstos. Isto é, ao mesmo tempo que há uma vastidão de exemplos que fundamentam a Teoria Cognitiva da Metáfora, há outros tantos que carecem explicação teórica.

Talvez este seja o momento oportuno para retomarmos os exemplos de (2.1) a (2.5) — expostos na seção anterior como ilustração dos recursos comumente utilizados por grande parte de autores da linha representacionista — para fins de comparação com a visão neo-representacionista:

(2.1) Ele *tirou* a camisa.
(2.2) Ele *tirou* a camisa do armário.
(2.3) Ele *tirou* 10 na prova.
(2.4) Ele *tirou* aquela idéia da cabeça.
(2.5) Ele *tirou* o corpo fora.

O mapeamento O CORPO É UM CONTAINER justificaria o uso do verbo *tirar* em (2.1), (2.4) e (2.5). Já a explicação para (2.2) não se pautaria em nenhum mapeamento entre domínio-fonte e domínio-alvo por não se tratar de um exemplo metafórico. Como já mencionado na seção 1.2, Lakoff reconhece que há "uma extensa gama de conceitos não metafóricos" (como "*o balão subiu*"). Na verdade, pode-se dizer que quando se trata somente do domínio-fonte, estaríamos diante de conceitos não-metafóricos.

De fato, há uma elegância explanatória nesses exemplos. Mas o que fazer com (2.3)? Teríamos que conceber *prova* como sendo CONTAINER ? Até onde é possível estender essas analogias? Quando deixam de ser válidas? Ou seja, o que fazer com aquela parte da formação de metáfora que foge ao mapeamento; o que fazer com as irregularidades da formação da metáfora? Essa é uma pergunta que não fica claramente respondida pelo modelo. O autor argumenta que a metáfora nova se constrói dentro de parâmetros impostos pelo sistema conceitual fixo, e "raramente ocorre independente dele" (Lakoff, 1993:228), mas não esclarece os casos em que isso acontece. Este será o nosso foco de crítica ao modelo em relação às CMs, mais adiante.

De uma forma ampla, mas talvez bastante elucidativa, podemos traçar características definidoras dessa linha de pensamento:

i) O significado não é puramente lingüístico; pertence ao escopo das estruturas cognitivas gerais fundadas na nossa experiência concreta (universal/cultural); a linguagem é simplesmente uma de suas manifestações superficiais.

ii) Cabe ao lingüista descobrir e explicar a estrutura conceptual que subjaz à linguagem;
iii) O escopo de análise é, portanto, muito mais amplo do que aquele proposto pelo representacionismo: inclui-se o não literal, o enciclopédico, o pragmático e o diacrônico;
iv) Não há uma ambição verificacionista; não é uma semântica calcada em lógica abstrata;
v) A característica principal da linguagem é servir de sinal concreto para "complexas construções não lingüísticas"
vi) Há ênfase nos processos humanos de categorização, na polissemia, e no poder figurativo da linguagem.

* * *

Note-se que embora as duas grandes perspectivas abordadas até aqui — representacionista e neo-representacionista, respectivamente — reservem para o fenômeno metafórico dois lugares opostos (quais sejam, marginal e central), ambas trafegam por caminhos teóricos que preconcebem um impulso representativo da linguagem, mesmo que sejam bastante distintos.

O fato de Lakoff admitir que há expressões literais na língua, embora seja uma pequena parcela, já nos dá instrumentos suficientes para afirmar que há subliminarmente às considerações neo-representacionistas uma crença de que a função da língua é, de fato, representar algo exterior a ela. Por isso, pode-se dizer que o modelo advoga um composicionalismo fraco.

Argumentaremos mais adiante que a constatação do valor teórico de ambas as perspectivas apresentadas até agora está nos casos contemplados; o seu custo teórico está nos casos deixados de fora. Para tanto, focaremos os dois ângulos através de suas considerações e conceituações do fenômeno multivocabular. Cremos que a partir de então a justificativa da nossa resistência em adotar uma ou outra visão sobre o significado ganhe mais clareza.

2.3.
As CMs sob os dois ângulos de representação

Qualquer tratamento teórico sobre as CMs esbarra inevitavelmente no tema da figuratividade. Antes, porém, de dar início a uma avaliação crítica sobre o que costuma ser dito em relação aos multivocábulos, não podemos deixar de abordar um tópico que precede qualquer discussão sobre o tema: a escolha terminológica. Existe uma fartura de rótulos — tradicionalmente apresentados por autores representacionistas — que, de certa forma, destinam-se ao mesmo "conceito": Expressão Multivocabular ou MWE (*Multi-word expression*, Blanco & Guenthner, 2000), Expressão Cristalizada (Neves, 1999; Vale, 2002), Expressão Fixa (Ranchhod, 2003), Colocação (Manning & Schütze, 2001; Tagnin, 1999), para citar alguns.

Alguns representantes da linha de pensamento literalista que se dedicam a esse tema, como Gaston Gross (1996) e Elizabeth Ranchhod (2003), concordam com a idéia de que o conceito de *expressão fixa* está longe de ser consensual, tanto do ponto de vista analítico e conceitual quanto terminológico. Neves (1999), Vale (2002*)* e Garrão (2001), dentro do mesmo pressuposto teórico, optam pelo termo *expressão cristalizada* e o diferenciam do termo *expressões com verbo-suporte*. Outros olhares, reconhecidamente influentes do ponto de vista teórico, como o do semanticista lexical David Cruse (1986) e do gerativista Ray Jackendoff (1997), utilizam termos que rotulam de forma diferenciada *expressões idiomáticas*, *metáforas cristalizadas* e *colocações*.

Talvez haja uma lista infindável de escolha terminológica e conceitual nesse domínio. A propósito deste problema, Gaatone (1990: 296) avalia que a noção de *expressão fixa* (*expression figée*), como prefere se referir ao conceito, encontra obstáculos pelo simples fato de já existirem na literatura vários termos, geralmente pré-teóricos, para as designar. A meu ver, contudo, o problema da explosão terminológica no domínio multivocabular é sintoma da **impossibilidade** de se chegar a um modelo com potencial teórico que dê conta desse conceito (ou não-conceito). Portanto, é por desconfiar que a conceituação clara do termo dependa de exemplos mais emblemáticos do que outros que hesitamos em apostar na sua tentativa de teorização.

2.3.1.
Multivocábulos e o representacionismo: a profusão de rótulos da semântica da inocência

Como já mencionado, o compromisso com a literalidade, com o imanentismo e com a composicionalidade do significado pauta a forma com que os multivocábulos são identificados e explicados por autores como Jackendoff (1997), Cruse (1986) e Gross (1982), por exemplo. A palavra seria o ponto de partida para a análise composicional do significado justamente porque ela é internamente composta de propriedades do significado necessárias e suficientes. O significado da palavra está dentro dela. E, se assim é visto, é possível, através de testes de composicionalidade, identificar muitos multivocábulos, uma vez que, de uma forma geral, eles não obedecem ao padrão de organização de significação linear na língua; ou seja, exemplificam níveis distintos de figuratividade ou de "opacidade semântica".

O grau máximo desta alegada opacidade seria exemplificado pelo conceito "expressão idiomática": Cruse (1986) afirma que os itens que a compõem não contribuem para o significado total da expressão. A checagem desse nível máximo de opacidade é possível através de testes com base na substituição, posposição, inserção de constituinte para se comprovar a impossibilidade de se compreender a expressão sem que todos os seus itens estejam presentes na ordem original (é o caso do exemplo clássico *bater as botas*).

O que argumentamos aqui é que mesmo esses casos considerados mais nitidamente impermeáveis do ponto de vista semântico não são tão indivisíveis assim se recorrermos a evidências de córpus. Ou seja, os testes de opacidade semântica, a nosso ver, não são teoricamente conclusivos porque: 1) os avaliadores são lingüistas e não falantes desavisados; 2) a noção de *opacidade* versus *transparência semântica* é escorregadia e também carece de uma delimitação teórica precisa e incontroversa.

Em uma busca ao sistema de Recuperação de Informação *Google*™, talvez o maior córpus digitalizado existente, pudemos constatar um fato curioso. A expressão *bater a caçuleta* também é utilizada no PB (principalmente no nordeste do país) com o mesmo sentido de *bater as botas*, o que nos faz supor que o SN da expressão não seria tão fixo quanto se imagina: "E o Doutor Morte finalmente

bateu a caçuleta", (*Casseta & Planeta online).* O mesmo detectamos sobre a expressão *quebrar galho.* Diferentemente de *bater as botas,* alega-se que essa expressão tem um grau de flexibilidade maior, por aceitar inserção/modificação de determinante (*quebra um galho, quebrar esse galho, quebrar mais esse galho*). Isso indicaria, segundo a concepção representacionista, um grau maior de transparência semântica, mas preservando algum nível de opacidade, uma vez que seria inviável, por exemplo, a substituição de *quebrar* por *partir* ou de *galho* por *vareta.* Seria, portanto, uma expressão semi-fixa.

No *Google*™, entretanto, pudemos encontrar frases do tipo "Quebra essa pra mim", "me quebra essa urgente", num teor bastante coloquial, é verdade, mas que parece revelar que o falante enxerga o "quebrar" como "fazer um favor". O vocábulo *galho* também é utilizado individualmente como algo paliativo, um "jeitinho". Isso seria resolvido por uma visão representacionista como evidência de um aumento do grau de transparência de *quebrar galho,* que ao longo dos anos, passou a ocupar a posição de uma **metáfora cristalizada.**

Mas, afinal o que é metáfora cristalizada? Searle (1979,p. 115) define o termo: "o sentido original da expressão metafórica é ignorado e a expressão adquire um novo sentido literal [...] Há um desvio de elocução metafórica para a elocução literal". Para Cruse (1986: 43), é possível diferenciar expressão idiomática de metáfora cristalizada através da aplicação do teste de substituição, que é mais aceitável no segundo caso. *Ele perdeu a razão/cabeça/juízo. Ele deu/disse adeus/tchau à reeleição* Em alguns casos, o autor explica, a substituição volta a dar vida a metáfora. *Ele quebrou a cabeça para resolver o problema /Ele rachou o cérebro para resolver o problema.* Já a expressão idiomática resiste a qualquer tipo de modificação de seus itens. *Ana bateu perna, Ana ? perna, Ana bateu?; Ana deu o braço a torcer, Ana deu o braço a ?, Ana ? o braço a torcer.* Cruse defende também que a tradução literal de uma metáfora cristalizada pode dar certo: *to break the ice; quebrar o gelo,* por exemplo. A utilização da mesma estratégia para expressões idiomáticas seria desastrosa: *bater perna, to beat leg*.

Esses exemplos parecem dar conta da diferença. Mas se *bater perna* e *dar o braço a torcer* são, de fato, expressões indevassáveis, a inserção de qualquer constituinte tornaria as expressões literais, mas não é o que os exemplos abaixo parecem demonstrar:

> *Resumindo, quem quiser economizar, ou fica em casa, ou vai ter que bater **muita** perna para achar onde comer e onde ficar.*
> (http://www.bemtevibrasil.com.br/diarioviagem18.htm)

> *Para não dizerem que sou um fanático apenas pela aviação militar, dei **meu** braço a torcer e consegui alguns interessantíssimos anúncios de companhias americanas.* (http://www.jetsite.com.br/aviacao_mkt.asp)

Ambas as expressões acima teriam comportamento idiomático nos testes, mas sua utilização pelo usuário da língua parece bem mais flexível. No ponto de vista de Cruse (1986), a expressão idiomática é uma unidade lexical elementar: "embora consista em mais de uma palavra, apresenta uma coesão interna de palavras simples" (Cruse, 1986:38). Embora o autor considere Expressões Idiomáticas, Metáforas Cristalizadas e Colocações como tipos de expressões cristalizadas distintas, reconhece que há casos limítrofes. Mas como, então, teorizar sobre um fenômeno que é escorregadio?

Com isso, pode-se dizer também que, quando Neves avalia que a expressão "tomar partido", em "Valéria *tomou partido* da tia" (Neves, 1999: 99), seria uma expressão cristalizada, está, na verdade, desconsiderando o fato de a construção admitir intercalação de advérbio dependendo do teor aspectual da frase em que se insere. Segundo a autora, a expressão não admite inserção de nenhum tipo de constituinte. O mesmo ela diz para a expressão "ter cabeça" em "O capitão Aparício *tem cabeça* para tudo". O Google, contudo, oferece contra-exemplos para o que a autora propõe:

> *A Quarta é um meio termo, uma sinfonia que parece não **tomar muito partido** desta relação, pois está montada sobre um afresco extremamente original de **...*** (www.mnemocine.com.br/filipe/ensaios.htm)

> *Tem que ser um exame de nível nacional para entrar quem tem **mais cabeça**, **...***
> (www.museudapessoa.net/MuseuVirtual/hmdepoente/
> depoimentoDepoente.do?action=ver&idDepoente=63&key)

Numa perspectiva alinhada à de Neves (1999), procuramos em Garrão & Dias (2001) encontrar em versões *on-line* dos periódicos *Veja*, *JB* e *O Globo* e posteriormente em Basílio, Oliveira & Garrão (2003), no córpus NILC[9], dados

[9] Núcleo Interinstitucional de Lingüística Computacional, contendo cerca de 37 milhões de palavras. É considerado um dos córpus mais abrangentes do PB, por incluir uma diversidade de gêneros discursivos.

qualitativos que espelhassem a noção de Expressão Cristalizada proposta por Neves. A idéia inicial era detectar as estruturas semanticamente opacas do tipo *bater+SN* para que estas pudessem figurar em um dicionário eletrônico (como *bater as botas, bater perna, bater pino, bater boca, bater os olhos, bater o martelo, bater ponto, bater papo,* entre outras).

Notamos que muitas das construções seriam consideradas pelos testes de composicionalidade propostos pela autora — substituição, coordenação, posposição, elipse e inserção de constituinte — como uma unidade indevassável. De fato, elas resistem à substituição das partes (*chocar boca), posposição (*boca bater), coordenação (*bater boca e as botas), elipse (*ele bateu boca e ela as botas) e intercalação (*bater minha boca). O que nos surpreendeu foi o fato de esse conjunto de testes ser insuficiente para dar conta de uma boa parte dessas construções, já que algumas delas parecem admitir intercalação de intensificador ou marcador de freqüência, conforme detectamos nos córpus. Abaixo seguem exemplos do córpus NILC e Portugal Natura Publico, respectivamente:

> *Quem pretende ter peixe à mesa durante a Semana Santa precisa* ***bater muita perna.***
> *Para os comunistas, o grave é que não estão em condições de* ***bater demasiado o pé.***

Os dados demonstram que há expressões supostamente indevassáveis que admitem inserção de advérbio, mais especificamente, de um marcador de freqüência ou um intensificador, o que, ao menos, nos faria ampliar os testes geralmente feitos para detectar o teor de fixidez dessas construções. É importante ressaltar, ainda, que a sua alegada opacidade semântica parece não ser definidora do nível de indivisibilidade da expressão, visto que há expressões que admitem marcador de freqüência e cujos constituintes não parecem ter o que se chama de papel composicional (como *bater perna, bater boca, dar trela, fazer questão).* Pudemos constatar também que o aspecto verbal da expressão como um todo parece ser muito mais preditivo em relação à sua fixidez do que a sua suposta opacidade, uma vez que tais construções verbais com aspecto pontual tenderiam a um grau de fixidez elevado (*bater o martelo/?bater muito martelo*) e aquelas com aspecto durativo seriam menos rígidas (*bater boca/ bater muita boca*). Mas

observamos também que isso parece ser um padrão de comportamento, uma tendência, não uma regra.

Não raras vezes, também, autores que optam pelo mesmo rótulo oferecem explicações díspares quanto ao seu conceito. Manning & Schütze (2003) e Tagnin (1999) ilustram de maneira clara esta constatação. Os primeiros, com foco em PLN, conceituam *colocação* como sendo uma combinação freqüente na língua com sentido específico entre, no mínimo, duas palavras — um bigrama — abrangendo expressões totalmente opacas, cujos itens não contribuem para o significado total da expressão (expressões idiomáticas), ou cujos itens contribuem parcialmente para o significado total da expressão.

Já para Tagnin (1999), numa abordagem voltada para a lexicografia, uma colocação não é necessariamente uma combinação de alta freqüência na língua. Considera que no fenômeno de colocação verbal, por exemplo, o verbo utilizado é consagrado, na medida em que sua ocorrência é preferencial, quando outros verbos semanticamente relacionados também poderiam ocorrer mas não ocorrem. Seria o caso, segundo ela, de *quebrar uma regra* ao invés de **romper uma regr*a. Esta última construção, no entanto, é facilmente encontrada no sistema de busca *Google*:

> *Por fim, do **Unplugged** saiu "The Man Who Sold the World", de David Bowie, que r**ompeu uma regra** dos Acústicos da MTV: Kurt eletrificou seu violão nessa música.*
> (http://www.cornflakepromises.hpg.ig.com.br/nirvana.htm)

Trata-se, portanto, de simples intuição da lingüista; mas não uma regra. Além disso, a autora defende, mais adiante no seu artigo, uma separação conceitual entre colocação e expressão idiomática, um argumento de difícil constatação pelo próprio teor informalizável dos dois fenômenos.

Numa passagem de seu texto, Tagnin ilustra o conceito de colocação através de uma estrutura que ocorre tanto na língua inglesa quanto no português: “Em português dizemos *levantar acampamento*, enquanto em inglês a colocação é “*break camp”*. (p.15). Mais adiante, ela segue explicando a sua metodologia:

> As ocorrências eram anotadas em fichas e posteriormente discutidas em grupo até se chegar à **noção clara** do que seriam as colocações verbais, pois muitos (dos seus alunos pesquisadores) as **confundiam** com expressões idiomáticas. Entrava aqui a **competência**. (Ibid:17; grifo nosso).

Ao ratificar que as ocorrências eram discutidas para não haver *confusão* entre o conceito de expressão idiomática e colocação, a autora se furta a descrever a tão valiosa "noção clara" dos conceitos envolvendo os dois fenômenos. Se considerarmos a definição atribuída por ela ao conceito de colocação (convencionalidade com possibilidade de substituição), poderíamos questionar a adequação do exemplo escolhido para representá-lo *(?*subir acampamento, ?*levantar tenda*). Em outras palavras, *levantar acampamento* seria um dos casos em que não se é possível estabelecer o que ela chama de "noção clara".

O grande incômodo que muitas co-ocorrências convencionais causam aos semanticistas representacionistas é, sem dúvida, a gradação de solidariedade entre os componentes da expressão. Há casos, minoritários, em que se pode detectar um grau máximo de coesão entre os itens e que, por isso, são mais facilmente rotulados, como, por exemplo: *assoar o nariz, arregalar os olhos, água potável* (dependência unildateral, Borba, apud Vale, 2001). A grande maioria dessas combinações, entretanto, não demonstra esse mesmo padrão e só pode ser detectada como uma seqüência convencional se comparada a combinações livres; o que, de certa forma, gera um novo problema, uma vez que a noção de combinação livre também pode ser considerada teoricamente frágil, como veremos na seção seguinte.

Vale (2001: 16), numa proposta de tipologia de *expressões cristalizadas* para o PB, também expõe sintomaticamente a arbitrariedade da intuição do pesquisador em relação aos testes de composicionalidade. A sua argumentação deixa clara a falta de força teórica distintiva entre opacidade e transparência semântica. Ao explicar a aplicação dos testes, recorre ao uso de "asterisco para inaceitabilidade; ponto de interrogação para aceitabilidade duvidosa; dois pontos de interrogação para aceitabilidade ainda mais duvidosa do que a precedente; três pontos de interrogação para aceitabilidade no limite da inaceitabilidade". Sua tentativa parece ser um sintoma de que não há como teorizar sobre as noções de opacidade/transparência semântica tendo como base a intuição do falante.

Esse tipo de verificação nos impulsiona a concluir que caracterizar deterministicamente uma expressão como opaca ou semi-opaca é elevar um olhar dedutivo, baseado em uma intuição interessada do pesquisador, a uma supremacia que talvez não mereça. Enquanto se priorizar uma abordagem dedutiva, que se pretende capaz de caracterizar a expressão tendo por base a própria intuição de quem a descreve, estaremos ignorando o fato de que é o discurso do falante desavisado, sem pretensões nem comprometimentos teóricos, a fonte mais segura para tanto.

2.3.2.
Multivocábulos e o neo-representacionismo: sinais de difusão teórica

Talvez Fillmore (1979) seja o texto mais apropriado para iniciar a crítica cognitivista à visão composicional do significado ou ao modelo do cálculo, de certa forma, preponderante dentro da Semântica. Argumenta Fillmore (1979: 63) que essa se baseia naquilo que ele conceitua como "uma segunda idealização" dentro da Lingüística. Assim como a sintaxe se vale do falante ideal proposto por Chomsky, Fillmore propõe que o falante idealizado pela Semântica representacionista seja melhor caracterizado por *falante inocente.*

Segundo essa idealização, o falante sofre sérias limitações, como o total desconhecimento de recursos idiomáticos lexicais (como, por exemplo, a diferença conceitual entre *caixa e caixão)* além de recursos idiomáticos sintagmáticos ou de qualquer princípio de construção de linguagem metafórica. Esse falante não teria informações prévias sobre ditados, fórmulas situacionais e comunicação indireta. Como ouvinte, o usuário inocente equaciona o significado de cada frase através do conhecimento das partes da frase e de sua organização. O autor questiona a confiabilidade dessa noção de composicionalidade e constata: "uma vez detectados os significados atômicos (*core meanings*) de tudo, não há parâmetro para se identificar quais combinações de palavras têm quais significados" (ibid:71).

O modelo lingüístico proposto por Ronald Langacker (1991, cap. 1) também é crítico à perspectiva composicional do significado. O autor é radicalmente

cético à idéia de que existe uma diferença demarcada entre combinações sintáticas e o que chamamos aqui de CMs. Segundo ele, o sistema lingüístico não é autônomo nem pode ser descrito sem referência ao sistema cognitivo. Propõe que as estruturas gramaticais não constituem um sistema formal ou um nível de representação autônomo. "O léxico, a morfologia e a sintaxe formam um ***continuum*** de unidades simbólicas, divididos **arbitrariamente** em componentes distintos".

Embora nenhum desses dois autores esteja advogando em favor do abandono de uma visão representacionista do significado, mas apenas de uma revisão desse tipo de representação, consideramos que suas reflexões abriguem de certa forma a nossa escolha teórica. Cremos que seja de extrema importância esclarecer que, embora resistamos em utilizar o arcabouço cognitivista, temos simpatia por muitas das suas idéias, que põem em xeque a visão semântica tradicional.

A partir de agora, entretanto, passamos a questionar as explicações oferecidas por autores neo-representacionistas sobre a formação de CMs, algo que seria motivado e, portanto, passível de explicação.

Enquanto as CMs, em geral, são tratadas pela tradição como um fenômeno irregular, por uma concepção neo-representacionista há uma explicação cognitiva para essas construções, um embasamento teórico que se pretende capaz de dar conta do impulso metafórico cognitivo gerador dessas estruturas multivocabulares. Portanto, uma das críticas que se faz em relação à primeira abordagem é o fato de somente ser possível a verificação do estatuto da expressão, mas não de previsão de estruturas multivocabulares.

De acordo com Lakoff (1991: 211), "as expressões idiomáticas não são automaticamente geradas por regras de produção lingüística, mas obedecem a padrões conceptuais". Como vimos, o autor prevê um mapeamento bidominal na construção de metáforas. Em Gibbs (1995) encontram-se inúmeras instâncias desses tipos de mapeamentos metafóricos na formação de expressões idiomáticas.

Ele inicia o seu artigo avaliando criticamente a contribuição teórica da Semântica tradicional e de seus testes sintáticos para caracterizar as expressões idiomáticas. "Artifícios sintáticos como esses citados acima (transformações em passiva, nominalização e movimento) são utilizados por uma teoria formal da

gramática como uma das deficiências transformacionais de expressões idiomáticas. Mas a visão tradicional de idiomaticidade não fornece explicações de como o falante adquire as regras de transformação para cada tipo de expressão idiomática. O falante não aprende formalmente quais são as expressões idiomáticas sintaticamente produtivas e improdutivas."(Gibbs, 1995: 272)

Ressalta também que as expressões idiomáticas não são completamente opacas como se costuma postular. Ele argumenta que algumas delas não são tão indecomponíveis quanto aparentam:

> Pesquisas indicam que os falantes americanos geralmente consideram algumas expressões idiomáticas como *miss the boat* e *button your lip* como altamente analisáveis, ou decomponíveis, e julgam outras como *kick the bucket* e *shoot the breeze*, como semanticamente indecomponíveis. (Gibbs, 1995: 279)

Embora o autor advogue em favor da transparência de algumas expressões idiomáticas, com a intenção de criticar a visão tradicional, alinha-se também à explicação de Cruse (1986) sobre o fato de haver algumas expressões mais transparentes do que outras. A única diferença está na nomenclatura: Cruse conceitua as expressões mais transparentes como *metáforas cristalizadas* enquanto Gibbs utiliza o termo *expressão idiomática* para todas essas CMs. Este, portanto, mesmo que inadvertidamente, está se baseando nos mesmos pressupostos teóricos para explicação do fenômeno: literalidade (uma vez que considera algumas expressões idiomáticas mais literais do que outras) e composicionalidade (já que algumas expressões são menos opacas e mais facilmente analisáveis do que outras).

Salomão (1990: 286), num estudo cognitivista sobre CMs encabeçadas pelo verbo *dar*, reconhece, assim como Gibbs, que há níveis de opacidade semântica, embora acredite que a situação lingüística ideal não esteja nos extremos desse *continuum*: "opacidade absoluta e total transparência são os dois extremos de um *continuum*, mas nenhum deles representa a situação lingüística ideal (*the linguistic situation at its best)*". Mais adiante a autora conclui:

> A gramática é conceptualmente motivada, mas não necessariamente conceptualmente transparente. Entre a forma e o significado, o processo de convencionalização intervém e afeta a expressão lingüística de tal forma que os usuários da língua são obrigados a buscar o frescor da representação de seus

> pensamentos. É nesse ponto que a atividade de conceptualização recomeça, e a riqueza da experiência humana de alguma forma emerge na forma lingüística. (ibid, p. 286)

Essa busca pelo frescor da representação do pensamento, contudo, é um mecanismo um tanto enigmático no modelo cognitivista. Qual é o ponto de partida para o recomeço da conceptualização? Até onde pode ir? Como novos conceitos e novos mapeamentos são formados?

O incontestável mérito do modelo cognitivista em relação às CMs seria o esclarecimento sistemático da motivação de construções possíveis, tendo em mente o mapeamento entre dois domínios, em que um deles, em geral mais abstrato, é entendido em termos de outro, em geral mais concreto. Scherer (2002) exemplifica o mapeamento em nossa cultura entre o domínio mais abstrato IDÉIAS e o mais concreto COMIDA através das seguintes CMs: *digerir uma idéia*, *devorar um livro*, *engolir uma história*. Por essa perspectiva, no entanto, o que impediria CMs como *comer uma opinião* e *devorar um raciocínio* de serem formadas?

Conclui-se, portanto, que o modelo é capaz de caracterizar uma forte tendência da capacidade de conceptualização das línguas. Contudo, seu poder preditivo é fraco, uma vez que mesmo entre dois domínios há uma explosão de possibilidades de combinações. E talvez seja um pouco leviano dizer que isso seja uma falha do modelo, já que seu o objetivo é caracterizar a motivação. No entanto, exatamente por essa razão, optamos por renunciar a uma abordagem neo-representacionista para o tratamento de CMs.

2.4
Discussão preliminar

Sob uma perspectiva teórica enxergamos imediatamente um problema em relação aos dois modelos expostos até aqui; respectivamente:

i) uma visão de composicionalidade de significado que está sintomaticamente invalidada pelos próprios exemplos escolhidos pelos autores nos córpus e pelo tipo de explicação calcada em uma gradação de nível de opacidade semântica

que não é incontroverso nem intuitivamente compartilhado pelos falantes (relatados na seção 2.3.1);

ii) uma visão de significado muito inclusiva que, embora seja elucidativa em muitos aspectos semânticos, acaba desqualificando seu poder explicativo por não determinar formalmente os limites de cada faceta da construção do significado. Em outras palavras, não caracteriza explicitamente em que medida o significado é cultural, mental, histórico ou calcado em experiência direta.

Já sob uma perspectiva lexicográfica ou descritiva também vemos implicações:

i) As CMs seriam descritas com base em testes de composicionalidade, um método que não só é trabalhoso e demorado como também pouco confiável, uma vez que não é raro o córpus contradizer o que os testes predizem (como *tomar **muito** partido, em 2.3.1*);

ii) As CMs seriam descritas com base em um modelo muito inclusivo e multifacetado. Um modelo que daria conta de expressões como *devorar um livro* mas que também poderia gerar outras como *comer um raciocínio.*

Por essa razão, escolhemos um caminho minimamente comprometido com representação de significado. Tal escolha leva a uma perspectiva lingüística amplamente amparada pela faceta estatística do significado. Trata-se de uma forma de enxergar as recorrências lingüísticas abrindo mão de rotulações semânticas pré-concebidas ou de considerações experiencialistas.

Esperamos poder demonstrar que esse caminho é bastante interessante para a lexicografia. Diana Santos (1990) argumenta que no domínio de PLN não se deve traçar a distinção entre as CMs: “As fronteiras entre restrições colocacionais, expressões idiomáticas e leituras metafóricas são difusas e talvez impertinentes para o tratamento automático da língua” (p.3). Não só subscrevemos a avaliação da lingüista portuguesa como desconfiamos que a não-pertinência dessa divisão

conceitual em PLN talvez seja um sintoma de que ela também deva ser revista no domínio lingüístico.

No capítulo seguinte propomos uma avaliação para as CMs que contorna minimamente esses recorrentes desafios apresentados por abordagens com viés representacionista. Trata-se de um caminho que não se pretende incontroverso mas é capaz de gerar resultados rápidos, com um grau elevado de precisão.

3
Por um caminho não-representacionista para a detecção dos multivocábulos

We begin to feel, or ought to, terrified that maybe language (and understanding and knowledge) rests upon very shaky foundation — a thin net over an abyss

S. Cavell

Neste terceiro capítulo da tese nos aventuramos por uma vereda relativamente pouco trilhada nos estudos lingüísticos, mas que já havia sido sinalizada pelas considerações sofistas da Grécia antiga. Trata-se de um olhar sobre o fenômeno da linguagem que privilegia o seu estatuto convencional. Dispensa, em contrapartida, qualquer tipo de teor simbólico ou representativo da linguagem em relação à realidade ou à mente.

Seguiremos aqui a mesma organização do capítulo anterior. Reservamos a primeira seção para breves considerações filosóficas sobre essa linha de pensamento; apresentamos na segunda seção o seu encaminhamento no âmbito lingüístico e no âmbito de PLN; já na terceira seção, lançamos esse olhar sobre o fenômeno da CM, apresentando a vantagem ou, talvez seja melhor dizer, a conveniência desse ponto de vista se comparado às perspectivas representacionistas discutidas no capítulo 2. Propomos, finalmente, em alinhamento com essa visão, uma avaliação de base estatística para detecção de CMs do tipo V+SN.

3.1
A herança filosófica

O discurso é um grande soberano, que com o mais diminuto e inaparente corpo as mais divinas obras executa.

Górgias*, Elogio a Helena*

Não é tarefa simples caracterizar as considerações não-representacionistas sobre a linguagem. Na história do pensamento filosófico, a visão sofista sobre o

discurso é talvez a primeira manifestação documentada dessa concepção de linguagem. Muito do que sabemos sobre o pensamento sofista em relação à linguagem deve-se aos escritos platônicos.[10] Platão apresenta as considerações sofistas — em que a linguagem é vista à luz da célebre máxima de Protágoras, "o homem é a medida de todas as coisas" — como contraponto ao pensamento socrático. Este formato dialógico apresenta, de certa forma, um teor didático para reafirmação de sua crença representacionista. Em outras palavras, o fato de Platão sustentar, em oposição violenta ao ideário sofista, a existência de verdades únicas e fixas, em alguma medida pode ter minimizado ou desqualificado a importância e teor das considerações sofistas ao longo da história do pensamento filosófico (cf. Marcondes,1997 e Martins, 2005). Logo, esse investimento de Platão contra os sofistas deve ter contribuído para minimizar e desqualificar também o ângulo anti-representacionista sobre a linguagem.

Dentre as características privilegiadas por essa visão de linguagem está a mutabilidade do significado de uma expressão em virtude de esta depender das práticas humanas, como também a incapacidade de a expressão representar algo exterior à linguagem.

Esse viés pragmático sobre a linguagem é retomado já na história mais recente da filosofia[11], notadamente, por Ludwig Wittgenstein[12]. De forma ostensivamente resumida, pode-se dizer que o filósofo austríaco se recusa a enxergar o estudo da linguagem dentro dos mesmos moldes metafísicos propostos para as ciências, principalmente porque o "cientista", neste caso, é uma das peças envolvidas no que chama de "jogos de linguagem" (Wittgenstein, 1979; Glock, 1996). Ele desenvolve este conceito através de uma analogia com a idéia de "jogo", e aponta certas características e semelhanças com a linguagem: jogos possuem regras, são práticas compartilhadas por uma comunidade, possuem peças, são autônomos, não requerem justificativas. Sua proposta é a de que os jogos de linguagem "são a totalidade da linguagem e das atividades com as quais está interligada" (Wittgenstein, 1979: §§ 7 e 23). Os jogos de linguagem seriam atividades autônomas que prescindem de

[10] O *Sofista* e *Crátilo* são exemplos de diálogos em que Platão tematiza a linguagem.
[11] Para uma boa apreciação sobre o assunto ver Martins, 1999.
[12] Refiro-me ao segundo Wittgenstein e sua visão pragmática expressa em *Investigações Filosóficas,* São Paulo Cultural, Coleção **Os Pensadores**, 1979.

explicação; “são parte de nossa história natural, assim como andar, comer, jogar, etc”. (ibid.: §25).

E o fato de o próprio homem estar encarcerado nos rituais lingüísticos que regem toda e qualquer manifestação lingüística compromete o seu julgamento teórico ou sua tentativa de explicação acerca do assunto. O estudo sobre a linguagem, portanto, seria ele mesmo mais um desses jogos. E como o significado de uma palavra encontra-se na execução da linguagem, não há nada a ser provado ou justificado. Não se pode explicar o que está explícito. A significação é habilidade em lidar com as palavras na linguagem. A explicação é resultado da nossa “ânsia de generalidade”. Ela pode até ser utilizada, mas não deveria ser entendida como uma meta-regra; ou uma regra fora do jogo. Toda explicação é interessada. (ver Scherer, 2002 sobre o conceito de *explicação*).

Em *Da certeza* (§559 in Sumares, 1994) — obra em que o autor tematiza a questão das regras que se cristalizam e passam a funcionar como pressupostos para a construção do conhecimento e como base para nossas ações — ele retoma o conceito de jogo: “O jogo de linguagem é, por assim dizer, imprevisível. Quero dizer: não está fundamentado. Não é racional (ou irracional). Está aí - como a nossa vida”.

Sua visão nega, deste modo, a vocação representacionista da linguagem tão defendida por filósofos como Platão, Aristóteles e Locke. Não há, para ele, uma essência do significado que preceda o uso das palavras. É somente na prática do uso que o significado se dá. E mesmo assim, ele não poderia ser “coisificado”. Por essa razão, é instável, contingente e fragmentado. Falar sobre a linguagem, ou fazer um relato sobre o relato, será um relato do jogo, e também faz parte dele. Portanto, nunca será definitivo.

Se num primeiro momento sua visão parece amputar a Lingüística, e mais especificamente, eliminar a Semântica, uma vez que sugere uma total impossibilidade de tratamento do fenômeno da significação, concordamos com Martins (1999:147) quando pondera que adotar uma visão wittgensteiniana na Lingüística “não corresponde à negação da possibilidade de qualquer estudo sistemático e empírico das línguas do mundo”. A autora sugere, portanto, que uma Lingüística sob a perspectiva wittgensteiniana “é viável e teria como propósito geral descrever as regularidades — parciais e contingentes —

observáveis nos jogos de linguagem que constituem as línguas do mundo"(idem).

A descrição de uma língua, segundo o filósofo, estará sempre condicionada à parcialidade imposta pelas infindáveis possibilidades de lances dos jogos de linguagem. Sobre a completude da linguagem, Wittgenstein confronta o leitor com o seguinte questionamento no §18 de *Investigações Filosóficas*:

>][...] (e com quantas casas ou ruas, uma cidade começa a ser cidade?) Nossa linguagem pode ser considerada como uma velha cidade: uma rede de ruelas e praças, casas novas e velhas, e casas construídas em diferentes épocas; e isto tudo cercado por uma quantidade de novos subúrbios com ruas retas e regulares e com casas uniformes.

Esse parágrafo de *Investigações Filosóficas* é particularmente relevante para o presente estudo porque se alinha à nossa escolha teórica e também metodológica: uma abordagem com base em córpus. Um exemplário da língua sempre pode ser maior, é verdade; mas nunca deixará de ser um extrato válido do seu uso. Portanto, a idéia de abarcar a língua de uma forma totalizante é uma utopia. Conclui-se, daí, que a parcialidade é inevitável.

Ao longo do tempo, algumas expressões são construídas, outras demolidas. Isto é parte do jogo. Todo córpus de uma língua seria, portanto, uma caracterização válida da mesma[13]. Voltaremos a esta questão mais adiante na seção 3.3.

3.2
Ecos do não-representacionismo na Lingüística e em PLN

> Na falta de uma explicação satisfatória para a noção do significado, os lingüistas que atuam na área da Semântica encontram-se na situação de não saber do que é que estão falando
>
> W. Quine

A posição do lingüista Roy Harris (1996), em relação ao funcionamento da língua alinha-se às idéias não-representacionistas solidificadas por

[13] para uma melhor apreciação das idéias wittgensteinianas ver Glock, 1996; Martins, 1999.

Wittgenstein. Em Scherer (2002) é possível encontrar uma farta exposição da simpatia de Harris às idéias do filósofo austríaco. Um dos pontos em comum entre as duas posições é a impossiblidade de distinção entre semântica e pragmática, ou entre conhecimento lingüístico e enciclopédico.

Note-se que a visão pragmática oferecida aqui se distancia de muitas abordagens pragmáticas da Lingüística. Estas conduzem suas explicações sobre o "uso da língua" com base na distinção *Semântica-Pragmática*, que com freqüência se apóia na distinção *significado literal-significado de uso*; algo indissociável por uma visão não-representacionista. Scherer (2002:26) avalia, portanto, que Wittgenstein e Harris compartilham uma visão pragmática radical, em que o uso lingüístico não é um dos componentes da linguagem, mas a única forma produtiva de se pensar os fenômenos lingüísticos.

Adam Kilgarriff, estudioso do léxico do ponto de vista computacional, também advoga este pragmatismo radical. Em seu contundente artigo *"I don´t believe in word senses"* (2000), ele atribui ao córpus o poder de desambiguação de significados. Demonstra que as palavras, em suas diversas acepções, estão desatreladas do seu alegado "sentido atômico" (*core meaning*). Propõe, como alternativa, uma ontologia, ou um conjunto de palavras semanticamente relacionadas, com base em uma convergência (*cluster*) estatística dos seus vários usos em córpus, em detrimento do seu sentido atômico. Os extratos do córpus são agrupados em sentidos específicos de acordo com os objetivos de tarefas específicas. Acredita, portanto, que os significados só existem dependentes de propostas ou tarefas *(tasks*) Ele conclui:"i*n the absence of such purposes, word senses do not exist*".

A visão de CMs proposta aqui insere-se neste pragmatismo radical. Dessa forma, o intuito é caracterizar as CMs com base no seu uso. É bem verdade que ao optarmos por considerar somente as CMs do tipo V+SN, estaríamos inevitavelmente reificando ou coisificando esse padrão de fragmento lingüístico. Mas é essa a nossa "tarefa", pelo apelo prático desse padrão estrutural tão recorrente no PB.

A abordagem que seguimos, portanto, é, em parte, filiada àquela proposta por Manning & Schütze (2003), pois, assim como esses estudiosos de PLN, estamos priorizando a freqüência das CMs e iremos utilizar o mesmo método

estatístico proposto por eles para a sua detecção. Os autores também creditam sua proposta a uma inspiração wittgensteiniana (p.17):

> *Philosophically, this brings us close to the position adopted in Wittgenstein (that is, Wittgenstein, 1968), where the meaning of a word is defined by the circumstances of its use (a use theory of meaning)[...]Under this conception, much of NLP research directly tackles questions of meaning.*

Diferentemente de Manning & Schütze, no entanto, não faremos nenhum tipo de caracterização apriorística do grau de opacidade semântica das CMs, que os autores rotulam por *colocações*: "uma combinação freqüente na língua com sentido específico entre, no mínimo, duas palavras — um bigrama — abrangendo expressões *totalmente opacas*, cujos itens não contribuem para o significado total da expressão (expressões idiomáticas), ou cujos itens *contribuem parcialmente* para o significado total da expressão" (p. 141, grifo nosso). Acreditamos que essa definição implica uma abordagem representacionista da linguagem. Propomos, em contrapartida, um novo tipo de aferição do grau de composicionalidade, guiada por medidas de similaridade entre adjacências lingüísticas, explicitada no capítulo 4.

3.3
A inevitabilidade do paradoxo do córpus

Em um estudo lexicográfico para o francês, Verlinde & Selva (2001) compararam a abordagem tradicional da construção de um dicionário de aprendizes de segunda língua, baseada na intuição do lexicógrafo, a uma abordagem baseada em córpus. De um modo geral, verificaram que é atribuído ao número de entradas (macroestrutura) de um dicionário uma importância maior do que ao conteúdo de cada entrada (microestrutura). Portanto, a ênfase recai nas palavras simples em detrimento das combinações de palavras, o que vem a ser um contra-senso, já que, para fins comunicativos, os aprendizes necessitam muito mais de informações sobre combinações do que sobre palavras isoladas.

Os pesquisadores puderam detectar, por exemplo, que, mesmo com a difundida repressão do governo francês aos anglicismos, o córpus apresentava

uma freqüência bastante representativa desses estrangeirismos. Os autores, no entanto, utilizaram como fontes periódicos jornalísticos, o que, segundo John Sinclair, pode ser questionável: “um córpus é uma coletânea de um material amplamente homogêneo, mas retirado de fontes diversas de forma que a individualidade de uma fonte se perde, a não ser que o pesquisador queira isolar um texto em particular. A diversidade de fontes é uma garantia de segurança dos dados” (Sinclair, 1991: 17-18).

Dentro dessa mesma perspectiva teórica, Verlinde & Selva (2001) destacam que a lexicografia de córpus é uma evidência empírica necessária à intuição do pesquisador, que serviria para preencher as lacunas do córpus que chamam de “não equilibrado”. Portanto, eles ressaltam a urgência da construção de um córpus equilibrado para o francês e outras línguas.

Podemos notar em grande parte dos lexicógrafos que se dedicam ao estudo de córpus o sentimento de que a compilação de fontes diversas evita o paradoxo do córpus. Contudo, essa caracterização do córpus ideal ignora a imagem da “cidade” apresentada por Wittgenstein e revela uma ilusão corrente na Lingüística: a de que é possível fugir do paradoxo do córpus. A questão intocada é a de que a língua em si não é completa; sempre é possível acrescentar mais uma casa ou mesmo uma rua, o que torna o córpus um fragmento de algo já fragmentado. Portanto, não há como fugir do paradoxo do córpus uma vez que a completude da língua também é algo inatingível.

Tal constatação, entretanto, não desabona um estudo descritivo da língua; o fato de não ser possível exaurir todas as possibilidades de jogos da linguagem não impede um lingüista ou um lexicógrafo de descrever o que geralmente é constatado nos jogos (cf. Martins, 1999: 147). Por outro lado, é desejável que o pesquisador se cerque de alguns critérios para aferir a adequação de um córpus em função dos objetivos pragmáticos que se deseja alcançar.

Por essa razão, dentre um escopo restrito de opções de córpus significativos no PB, recorremos a um córpus jornalístico. Concluímos que seria um extrato de língua atraente para a nossa tarefa.

3.3.1
O córpus utilizado: CETENFolha

O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo) é um córpus jornalístico de cerca de 24 milhões de palavras em PB, parte integrante do córpus NILC[14], fornecido pelo projeto Linguateca (www.linguateca.pt). A *Folha de S. Paulo* é um jornal diário brasileiro de grande circulação. Além das habituais seções diárias, o jornal contém igualmente diversos cadernos não-diários, a maior parte dos quais foram incluídos no córpus. Existe também desde 1996 uma edição electrônica (http://www.folha.uol.com.br). Dentre as seções incluídas no córpus CETENFolha estão, por ordem alfabética, Agrofolha, Brasil, Caderno Especial, Cotidiano, Dinheiro, Empregos, Esporte, Folha Ciência, Folhateen, Folhinha, Fovest, Ilustrada, Imóveis, Informática, Mais!, Mundo, Opinião, Revista Folha, TV Folha, Tudo, Turismo, Veículos.

Não foram incluídos no córpus:

- os artigos de primeira página que apenas chamam o artigo principal nas páginas interiores: *NORDESTE Chuvas voltam ao sertão da Paraíba e trazem esperança aos agricultores PÁG. 3;*
- os artigos com menos de 250 caracteres;
- algumas chamadas para outras páginas dentro de um artigo: *(leia mais na página 3);*
- anotações internas que se considerou não fazer parte de notícias ou outro texto do jornal, e que não tinham relevância para o corpus: (*Edição: São Paulo*);
- artigos duplicados (i.e., quando existiam duas cópias do mesmo artigo, só uma foi incluída).

A nosso ver, portanto, a predominância do discurso semi-formal, a transcrição de discursos diretos em entrevistas e a variedade de assuntos disponibilizada pelas diversas seções do jornal tornam esse gênero uma fonte significativa de jogos de linguagem no PB.

[14] O corpus NILC -USP, contém textos brasileiros do registro jornalístico, didático, epistolar e redações de alunos. Trata-se de um extrato com 37 milhões de palavras.

De fato, a opção por um córpus de teor jornalístico tem suas implicações: a língua fica prioritariamente associada àquilo que é considerado notícia em detrimento, por exemplo, de um bate-papo despretensioso entre adolescentes. Entretanto, a escolha por esse tipo de extrato da língua também está associada à falta de um córpus mais robusto do PB. Uma outra razão da opção pelo córpus CETENFolha está no fato de ser o único significativo no PB disponível para *download*; e, portanto, o único passível de aplicação dos testes probabilísticos que virão mais adiante.

3.4
O teor estatístico do fenômeno lingüístico

Embora proposta antes do advento de córpus computadorizados, a Teoria Contextual do Significado proposta pelo lingüista britânico J. R. Firth (1957), subjaz a uma abordagem a partir de córpus. Ele propõe que o estudo do significado e do contexto devem ser centrais para a lingüística. Recusa-se a admitir qualquer tipo de distinção entre *langue* e *parole* (estabelecida antes dele por Saussure) ou entre *competência e desempenho* (estabelecida mais tarde por Chomsky), já que não enxerga a língua como uma entidade autônoma. Os *eventos lingüísticos*, nos seus termos, são recorrentes e repetidamente observáveis. Dessa forma, sua visão se alinha à perspectiva pragmática adotada neste estudo — que, por sua vez, leva a uma perspectiva computacional, já que só podemos constatar mais precisamente certos eventos na língua com auxílio de um aparato que possibilite o armazenamento desses dados e sua recuperação.

No mesmo ano em que Firth profetizou "*You shall know a word by the company it keeps",* Chomsky tornou pública, em *Syntactic Structures,* a sua crença de que a língua deveria ser analisada em um nível muito mais profundo do que sua estatística superficial. Isto é, numa crítica à visão empirista, argumentou que uma abordagem com base em dados da língua consideraria como identicamente remotas as sentenças "*Colorless green ideas sleep furiously" e "Furiously sleep ideas green colorless"*. Ou seja, em qualquer modelo estatístico, as duas sentenças seriam igualmente excluídas, embora, como falantes, nós saibamos que a primeira obedeceria a algumas regras

gramaticais. O que Chomsky defendia é que a primeira não poderia ser totalmente desconsiderada e equiparada à segunda simplesmente pelo fato de sua probabilidade de ocorrência ser remota[15]. Essa crítica, conhecida como *problema dos dados esparsos,* foi altamente influente para uma mudança de perspectiva teórica nos anos 60.

Mas a evolução de técnicas estatísticas aliada à possiblidade de construção de córpus mais robustos de alguma forma equacionou aquilo que Chomsky utilizava como emblema de sua crítica. Além disso, a forma de Firth enxergar a língua não se presta à comparação com o modelo chomskyano. Para ele a língua é um *evento*, uma forma de ação (*a way of "doing things"*); e é por isso que seria um ato legítimo por parte do lingüista se ater aos eventos discursivos propriamente ditos. Firth acreditava que o evento era um fim em si mesmo e não uma forma de acessar o "verdadeiro" objeto de estudo: o sistema subjacente a ele, como Chomsky advogava. E o nosso estudo alinha-se a esse perspectiva eventiva da língua.

3.4.1
Mãos à obra

A seguir está a primeira etapa prática desta pesquisa. Nesta seção apresentamos uma listagem dos verbos mais recorrentes no córpus CETENFolha em três situações distintas: sua freqüência absoluta (Tabela 1); sua freqüência seguida facultativamente de deteminante e obrigatoriamente de um nome, formando a estrutura V+(det)+N (Tabela 2); sua freqüência seguida facultativamente de deteminante e obrigatoriamente de um nome seguido obrigatoriamente de marcas de pontuação ou advérbio ou conectivo, formando a estrutura V+(det)+N intransitiva (Tabela 3). Já em 3.5.1 apresentamos um teste estatístico (Banerjee & Pedersen, 2003), para detectar as CMs, encabeçadas por cada um dos 10 verbos mais freqüentes com esse padrão, em

[15] Ironicamente, a probabilidade de ambas as frases ocorrerem é grande em virtude de sua relevância ilustrativa na Lingüística.

detrimento de combinações sintáticas aleatórias. Após essa detecção, o método lista as CMs por ordem de freqüência.

O critério estatístico para detecção de CMs na língua tem uma dupla função. Primeiramente, pela própria natureza freqüencial do fenômeno, consideramos ser um recurso primordial para trazer à tona os padrões de co-ocorrências mais utilizados; em segundo lugar, ao lançar mão de recursos estatísticos para detecção das CMs mais freqüentes, não apenas poupamos tempo e trabalho de que necessitaríamos para verificar se uma seqüência é ou não freqüente na língua, como também contamos com um valioso aliado para corroborar a listagem: dados reais da língua.

Numa avaliação do córpus CETENFolha, listamos os 30 verbos mais recorrentes. Isto é, sua freqüência absoluta:

CETENFolha (24 milhões de palavras)			
ranqueamento	***ocorrências***	***lema***	***percentual no córpus***
1	109282	ser	0.46%
2	39422	ter	0.16%
3	36743	estar	0.153%
4	20668	poder	0.086%
5	20440	ir	0.085%
6	14660	fazer	0.061%
7	12880	haver	0.054%
8	11928	dever	0.049%
9	8941	querer	0.037%
10	7037	dar	0.029%
11	6852	ficar	0.028%
12	6822	vir	0.028%
13	5576	dizer	0.023%
14	5049	chegar	0.021%

15	4836	passar	0.020%
16	4369	saber	0.018%
17	3952	começar	0.016%
18	3818	acontecer	0.016%
19	3523	conseguir	0.015%
20	3483	receber	0.014%
21	3369	ver	0.014%
22	3279	levar	0.013%
23	3041	deixar	0.012%
24	2999	existir	0.012%
25	2901	precisar	0.012%
26	2744	ocorrer	0.011%
27	2734	trabalhar	0.011%
28	2710	sair	0.011%
29	2663	pretender	0.011%
30	2662	ganhar	0.011%

Tabela 1: freqüência absoluta dos 30 verbos mais recorrentes no córpus

De acordo com Zipf (1949) "quanto maior a freqüência de uma palavra ou morfema, maior será o número de combinações possíveis (grosso modo, compostos e formas morfologicamente complexas)." Seguindo essa hipótese, portanto, as CMs verbais mais freqüentes na língua seriam encabeçadas pelos verbos apresentados na Tabela 1, que são os mais freqüentes no córpus.

Vejamos agora a freqüência desses mesmos verbos seguidos facultativamente de determinante e obrigatoriamente de nome, formando a estrutura V+(det)+N[16]. Em negrito estão os verbos que não figuram na primeira tabela.

[16] Tomando como exemplo o verbo *fazer*, o formalismo para tal detecção em *linguateca.com.pt* seria **([lema="fazer" & pos="V"] [pos="DET.*"]? [pos="N"] [classe= "JOCF"])**. A fórmula JOCF se refere ao córpus CETENFolha, em detrimento de outros também disponíveis no Linguateca.

CETENFolha V+(det)+N		
ranqueamento	***ocorrências***	***lema***
1	29522	ser
2	24570	ter
3	20101	fazer
4	8545	haver
5	7037	dar
6	**4241**	**usar**
7	3483	receber
8	3369	ver
9	**3320**	**criar**
10	**3237**	**tomar**
11	3035	deixar
12	**2888**	**perder**
13	**2841**	**pedir**
14	2824	levar
15	2682	pagar
16	2625	ganhar
17	**2268**	**comprar**
18	2047	passar
19	**1850**	**vender**
20	1528	conseguir
21	**1471**	**tirar**
22	**1323**	**causar**
23	1286	existir
24	1209	dizer
25	**1121**	**jogar**
26	**1064**	**matar**
27	**1019**	**virar**
28	914	estar
29	866	querer
30	818	começar

Tabela 2: freqüência de verbos mais recorrentes do córpus CETENFolha seguidos facultativamente de deteminante e obrigatoriamente de um nome.

Note-se que exceto por “ser” e “estar”, a posição dos verbos é bastante diferente da sua freqüência absoluta (ver Tabela 1). Além disso, a Tabela 2 apresenta verbos que nem mesmo figuram entre os trinta mais freqüentes da Tabela 1 (expostos em negrito).

A Tabela 2, entretanto, não representa a freqüência ideal de V+(det)+N com teor intransitivo no córpus. Como essa lista se refere a qualquer segmento V+(det)+N, sem qualquer outra restrição de recorrência vocabular, pode incluir também CMs de padrão V+SN+SPrep (***dar o braço*** *a torcer;* ***dar o ar*** *da graça;* ***tirar o cavalinho*** *da chuva*) assim como V+det+N+pron (***tem gente*** *que adora*...).

A saída menos comprometedora para restringir o escopo da busca por uma estrutura V+(det) +N com teor intransitivo foi a inclusão de uma restrição com base em marcas de pontuação, conectivo e advérbio (para que fosse possível, por exemplo, a detecção da expressão *Ele* ***teve alta***; mas também *ele* ***teve alta*** *ontem*; assim como *ele* ***teve alta*** *e já está em casa*)[17]. Abaixo segue a freqüência dos verbos mais recorrentes com esse padrão no córpus. Em negrito estão os 10 verbos selecionados para implementação dos testes estatísticos com padrão V+(det)+N. Em cinza estão os verbos que foram descartados da implementação estatística, como justificado mais adiante.

CETENFolha

V+(det)+N com padrão intransitivo

ranqueamento	*ocorrências*	*lema*
1	**7511**	**ser**
2	**5132**	**fazer**
3	**4958**	**ter**
4	**1993**	**dar**
5	**1749**	**haver**
6	**1075**	**perder**
7	**1038**	**usar**
8	**981**	**receber**
9	**940**	**deixar**

[17] Tomando como exemplo, o verbo *ter*, o formalismo de busca seria: **([lema="ter" & pos="V"] [pos="DET.*"]? [pos="N"] [word="\.|,|;|:|\?|!|"]** | **"pos=KC.*"** | **"pos=ADV.*" [classe= "JOCF"])**

10	**901**	**tomar**
11	**861**	**ganhar**
12	**817**	**ver**
13	**770**	**criar**
14	732	pagar
15	728	comprar
16	564	pedir
17	541	levar
18	525	vender
19	480	dizer
20	438	matar
21	423	passar
22	402	causar
23	400	conseguir
24	387	virar
25	375	jogar
26	371	estar
27	342	tirar
28	262	querer
29	249	começar
30	237	comentar

Tabela 3: freqüência de verbos seguidos facultativamente de deteminante e obrigatoriamente de um nome posposto por marcas de pontuação, conectivo ou advérbio.

Os 30 verbos mais freqüentes expostos na Tabela 3 são essencialmente os mesmos da Tabela 2, muito embora o ranqueamento da Tabela 3 não espelhe as ocorrências dos verbos da Tabela 2, exceto pelo verbo “ser”. O único verbo que não figura nessa terceira tabela é “existir”, cujas ocorrências (234) foram excedidas pelo verbo “comentar” (237).

Segundo Vale (2001), a expressão cristalizada mais recorrente em PB é a V+(det)+N de teor intransitivo — que dentro de outra proposta teórica, rotula como V_0N_0C. Ele afirma também que os verbos mais recorrentes em CMs em geral, que ele rotula como *Expressões Cristalizadas*, são os verbos-suportes ou verbos leves como “dar”, “ter”, “fazer”, “ficar”, “levar”, “tomar”, “tirar”, “pôr”.

De fato, “dar”, “fazer”, “ter” são verbos que figuram no topo da lista da

Tabela 3. Por outro lado, os nossos resultados desmontram que há verbos que não se caracterizariam propriamente como leve, mas que são altamente freqüentes com a estrututa V+(det)+N. Por exemplo, “perder” (6°) “usar” (7°) “deixar” (9°), “ganhar” (11°), “criar” (13°), dentre outros, superam a freqüencia de verbos tradicionalmente rotulados por leves ou suporte, como “levar”,” tirar”, por exemplo, que estão em 17° e 27°, respectivamente.

Além desses, figuram o verbo de cópula “ser” (1°) e o verbo impessoal “haver” (5°). Como o verbo “ser” além de ser sempre o mais freqüente dos verbos em qualquer córpus de prosa, também apresenta uma farta variedade de combinações estruturais (funcionando quase como uma palavra funcional) optamos por excluí-lo da nossa pesquisa, já que seu ecletismo sintático ou sua explosão combinatória acabaria falseando o teste de verificação de CM. Já o verbo “haver” foi excluído da análise pelo seu teor impessoal. Ao contrário dos outros verbos presentes na Tabela 3, ele prescinde de sujeito (*Houve tumulto; há jurisprudência)*, o que o distancia de estruturas de teor agentivo (como, por exemplo N *fazer N,* N *criar* N) ou passivo (como, por exemplo N *receber* N*,* N *ganhar* N).

Uma outra exclusão também foi feita em relação às CMs encabeçadas pelo verbo “ver” em função de seu teor metalingüístico no córpus, como em *ver tabela, ver resultados, ver fotos, ver textos, etc.* Muito embora a descrição do córpus afirmasse que esse tipo de informação tivesse sido excluída — “algumas chamadas para outras páginas dentro de um artigo” —, isso parece não ter sido feito de forma exaustiva.

Portanto, as 10 estruturas V+(det)+N que serão listadas mais adiante dizem respeito aos verbos que figuram em negrito na tabela 3. São eles: “fazer”, “ter”, “dar”, “perder”, “usar”, “receber”, “deixar”, “tomar”, “ganhar”, “criar”. Note-se que este é um recorte meramente metodológico e que sua arbitrariedade, necessária para os limites desta pesquisa, deixará de avaliar outras estruturas com o mesmo padrão que também são bastante freqüentes com outros verbos.

3.5
A identificação das CMs

Existem alguns métodos estatísticos disponíveis para a detecção de CMs. Dentre eles, o mais simples e previsível é a seleção baseada na freqüência. Consiste na computação das freqüências de pares de palavras (bigramas). Os bigramas mais freqüentes seriam candidatos naturais a CMs. O grande problema desse método é que as palavras mais freqüentes da língua tendem a se combinar mais do que aquelas não tão freqüentes (é o caso do verbo "ser"*)*.

No caso das combinações V+det+N, por exemplo, o padrão contém uma palavra funcional: os determinantes. Trata-se de uma classe de palavras com freqüência tão elevada que sua combinação com nomes freqüentes poderia ser sempre analisada como uma CM. Para que isso não ocorra, aplica-se um filtro de classes de palavras, onde somente se consideram padrões candidatos potenciais a sintagmas. Tal método apresenta algum grau de eficiência; porém, dada a sua simplicidade, os resultados não são considerados suficientemente precisos e destinados somente a combinações fixas (ex: nomes compostos).

No caso de ocorrências descontíguas, ou seja, de combinações mais flexíveis, como é o caso de CMs de base verbal (*fazer muito tempo*; *tirar todas as dúvidas)*, são necessários recursos mais sofisticados, dentre os quais estão métodos como Média & Variância e Testagem de Hipóteses. Este último método é subdividido em Teste t, Testagem de hipóteses de diferenças, Teste χ^2 de Pearson e Logaritmo de Verossimilhança (Manning & Schütze, 1999: cap. 6)

O método de Média & Variância, contudo, não é destinado para detecção de uma estrutura sintática específica. É mais apropriado para detecção de vários padrões sintáticos entre duas palavras relacionadas, como por exemplo "bater", "porta". As palavras que aparecem entre os dois termos variam e a distância entre eles também não é constante. Mas a regularidade nos tipos de ocorrências permite determinar que "bater" é o verbo utilizado neste tipo de situação, e não "golpear" ou "topar".

O método funciona com base na contagem de palavras vizinhas àquelas que estão sendo testadas. Calcula-se a média e a variância do deslocamento

entre as duas palavras em um córpus. Tomemos o seguinte córpus de 4 sentenças como exemplo:

(3.1.) Ela bateu na minha porta.
(3.2.) Eles bateram na sua porta.
(3.3) Três mulheres bateram de novo naquela outra porta.
(3.4) Um homem bateu quinhentas vezes na sua porta.

Contam-se quantas palavras a partir de "bater" existem antes de "porta" em cada uma das quatro frases e calcula-se, a partir de então, a média de deslocamento[18] entre as quatro frases. A média é simplesmente o deslocamento médio.

Já o método de Testagem de Hipóteses se revelou mais adequado para os nossos fins. Tanto pela sua natureza amigável, clara e explicativa quanto pela robustez dos resultados, demonstrou ser bastante pertinente para a detecção de CMs verbais, como demonstraremos a seguir.

3.5.1 Testagem de hipóteses

Um problema típico de avaliações estatísticas é determinar se algum evento se deve ao acaso ou é motivado. Para tanto, aplica-se a metodologia de Testagem de Hipóteses. É preciso formular uma hipótese nula que apresenta a relação entre duas palavras que não formam uma CM. Para tanto, assumimos que as duas palavras são completamente independentes. A probabilidade de co-ocorrerem é dada por:

$$H_0: P(w_1 w_2) = P(w_1)(w_2)$$

Onde H = hipótese, P = probabilidade, w = palavra

[18] A média de deslocamento entre *bater* e *porta* é calculada assim: $1\ \frac{(3+3+5+5)}{4} = 4.0$

Dentre os métodos de Testagem de Hipótese, o *Logaritmo de Verossimilhança* tem sido um dos mais utilizados. Ele tem por objetivo detectar se um bigrama é uma CM ou uma co-ocorrência casual na língua. Esse tipo de testagem requer a formulação de dois tipos de hipóteses:

Hipótese 1 (H1): a probabilidade de a ocorrência da primeira palavra de uma seqüência depender da ocorrência da segunda palavra é a mesma do que sua ocorrência independente da segunda palavra;

Hipótese 2, (H2): a probabilidade de a ocorrência da primeira palavra de uma seqüência depender da ocorrência da segunda palavra não é a mesma do que sua ocorrência independente da segunda palavra.

Essas duas hipóteses estão formalizadas abaixo:

H1: $P(w_1 | w_2) = P(w_1 | \neg w_2)$

H2: $P(w_1 | w_2) \neq P(w_1 | \neg w_2)$

Onde H = hipótese, P = probabilidade, w = palavra

Por exemplo, assumindo que a expressão *fazer sucesso* seja uma CM,

H2*:P(fazer | sucesso) ≠ P(fazer |¬ sucesso)*

espera-se que a hipótese de independência

H_1*:P(fazer | sucesso) = P(fazer |¬ sucesso)*

seja falsa. Portanto, o método avalia a probabilidade de H2 ocorrer em detrimento de H1.

A aplicação do Logaritmo de Verossimilhança, disponibilizado através do pacote estatístico NSP (Banerjee & Pedersen, 2003), foi viabilizado, nesse projeto, através de um programa, feito em linguagem Java™, que recebe como entrada o córpus e fornece como resultado a lista de todas as co-ocorrências do tipo V+(det)+ N (Nogueira, 2004). Só a partir de então, aplica-se o teste estatístico e é estabelecida a lista das candidatas a CMs que, posteriormente, são ordenadas por freqüência. Esse teste será aplicado aos dez verbos com o padrão procurado (em negrito na tabela 3) .

3.5.1.1
O teste e a avaliação dos resultados

De forma clara, o teste é capaz de trazer à tona combinações do tipo V+(det)+N consideradas pela literatura representacionista como distintivas se comparadas a uma combinação sintática tradicional. Embora estejamos evitando este tipo de consideração, o resultado a que chegamos já se mostra produtivo por demonstrar que um método semanticamente cego é capaz não só de extrair as CMs do córpus, como também de identificar aquelas mais freqüentes. Uma outra virtude do método é a sua eficácia na detecção de determinante, um aspecto relevante no estudo das CMs do tipo V+SN (cf. Garrão, 2001, Vale, 2002). Além disso, o córpus também se revelou bastante profícuo para a tarefa proposta. A variedade das seções, colunas e matérias do jornal garantiram um resultado abrangente.

Sob uma perspectiva quantitativa, portanto, o método se revelou satisfatório. Em outras palavras, dentre as 1000 candidatas a CMs apontadas pelo método, apenas 128 foram consideradas ruído. Um acerto de 87, 2%[19].

As listas das 1000 CMs detectadas, divididas em 100 ocorrências por verbo, seguem no final desta seção. As "pseudo-CMs" extraídas pelo método, ou seja, os "deslizes" por ele cometido (12,8%), estão indicados na lista da seguinte forma:

i) Erro de avaliação estrutural. Este tipo de erro pode ter sido cometido pelo método por duas razões principais:

- (ETQ)

Em função da etiquetagem equivocada no córpus: por exemplo, algumas combinações nominais compostas, como *reforma agrária*, são consideradas como tais. Isso auxilia a eficácia do método, que detecta satisfatoriamente a CM *fazer a reforma agrária*. Em contrapartida, há palavras que estão erradamente etiquetadas, como por exemplo *livre (em Romário* ***recebeu livre*** *dentro da área),* que está anotada no córpus como nome (SN), talvez em função de sua

[19] Não incluímos nessa avaliação quantitativa o teste 10 (VER). O verbo CRIAR, por ser o 11° mais freqüente para o padrão de CM procurado, foi alçado à 10° colocação.

posição nominal, quando seria adjetivo (Adj). A identificação de CMs como g*anhar líquido* e *ganhar real* também é conseqüência de erros de etiquetagem ("ganho" foi identificado como primeira pessoa do singular do verbo, quando, na verdade, trata-se de um nome).

- (JAN)

Em função de o método ter considerado uma janela sintática menor do que a expressão representa: *ter um papel*, por exemplo, foi detectado pelo método como uma CM do padrão procurado quando, na verdade, sua estrutura vai além de V+(det)+N. Seria V+(det)+N+(Adv)+Adj: *Ter um papel muito importante.* O mesmo acontece com *fazer um trabalho*, *(Paulo Roberto, ainda sem forma física,* ***faz um trabalho*** *apenas burocrático),* e *ganhar linhas (A carroceria* ***ganhou linhas*** *mais atuais).* Esse deslize é conseqüência do formalismo utilizado para restringir o escopo da busca por uma estrutura V+(det) +N que antecedesse um advérbio. Uma outra instância de ruído foi gerada em função de estruturas sintáticas invertidas, deslocadas ou descontínuas: um exemplo é ***ganhar este ano*** *(a campanha* ***ganhou este ano*** *novo fôlego diante da participação dos jornais*) que foi detectada ao invés de ***ganhar novo fôlego***. Um outro exemplo é ***ter acesso*** ("*O mundo dos fanzines tem algo de sociedade secreta, na qual apenas alguns iniciados* ***têm acesso."***) Nesse último caso, o ruído foi gerado em virtude da restrição de pontuação, que, por outro lado, foi altamente econômica e compatível com o que queria ser alcançado, como demonstramos anteriormente nesta seção (p.47).

ii) Outros ruídos foram atribuídos exclusivamente ao córpus:

- (DAT)

CMs claramente datadas: São elas: *criar a URV, usar a URV, tomar AZT.*

- (MET)

Metatexto do córpus: encontrado exclusivamente no teste 10 (*ver resultados, ver tabela*, *ver fotos*, etc), que foi excluído pelo teor metalingüístico de muitas de suas CMs.

Há outros dois tipos de interferência na detecção de CMs que não foram considerados propriamente ruídos. São eles:

- (COE)

Recursos coesivos, como a utilização de anáfora: alguns exemplos são *fazer a denúncia, dar a notícia, ter a doença*. Na verdade, não consideramos esses casos como pseudo-CMs. Ressaltamos apenas que a presença do determinante (na maioria dos casos, um artigo definido) não foi detectada em todos os casos em que a CM ocorre. O mesmo não acontece com CMs como *fazer o contrário, dar as caras, perder a virgindade, usar o bom senso,* dentre outras, cujas estruturas só foram detectadas dessa forma.

- (MAN)

Omissões de artigo (tanto definido quanto indefinido), características de manchetes de jornal, como *Presidente da Shell* ***deixa cargo*** *amanhã*. Embora a descrição do córpus argumente ter eliminado todas as manchetes, isso parece não proceder em todos os casos. Esse tipo de recurso também não foi avaliado como pseudo-CM.

Finalmente, é importante atentar para o fato de que algumas CMs têm mais de uma estrutura transitiva. É o caso, dentre outras CMs, de *fazer campanha*: V+N (*Até agora, apenas Munhoz tem* ***feito campanha*.)** e V+N+Prep+(det)+N, (que foi detectada em sua estrutura deslocada em *O secretário Antonio Felix Domingues* ***faria campanha,*** *em Goiás e Tocantins, para o ex-governador Orestes Quércia nas prévias do PMDB***)**. Como há a possibilidade de a expressão desempenhar um papel intransitivo, não foi identificada na listagem como uma pseudo-CM.

Em suma, a grande vantagem deste método está no seu teor preditivo. Através dele, podemos constatar preferências de usos das expressões presentes no córpus. A seguir, apontamos exemplos que ilustram o grau de eficiência do método. Incluímos também alguns exemplos de ruído, ressaltando a necessidade de uma revisão humana:

No teste 1 (“FAZER”):

A CM *fazer parte* foi apresentada pelo córpus também na forma intransitiva, embora, intuitivamente, talvez pudéssemos afirmar que a expressão exige um complemento transitivo indireto lingüisticamente explícito.

- *Ciúme dosado* ***faz parte.***
- *O medo* ***faz parte.***
- *Perder gol* ***faz parte.***
- *Lógico que tem um lado burocrático, mas* ***faz parte;***

Por essa razão, foi considerada uma CM.

Já *fazer comentários*, embora também tenha sido avaliada como uma CM (*Iasser Arafat não foi encontrado para* ***fazer comentários****)*, parece estar atrelada a uma janela estrutural maior encabeçada pelo ADV *não* ou alguma consideração de teor negativo como a preposição *sem,* o verbo *evitar* ou a expressão verbal *não querer* :

- *Iberê Camargo as leu, mas não* ***fez comentários.***
- *Roseana também não* ***faz comentários.***
- *Itamar e June trocaram presentes, mas não* ***fizeram comentários.***
- *Ao sair, Cerqueira preferiu não* ***fazer comentários.***
- *Tim ignorou a regra de receber o troféu e deixar o palco sem* ***fazer comentários.***
- *Luxemburgo evitou* ***fazer comentários.***

No teste 2 (“TER”):

A CM *ter uma relação* também parece ter um uso intransitivo, além do seu teor transitivo indireto tradicional.

- *Quando eu estava sozinha, sentia falta de* ***ter uma relação,*** *não de trepar por trepar.*

É possível detectar, também, um padrão de ocorrência característico de CMs encabeçadas pelo verbo *ter*. V+N (*ter caráter*, *ter valor*) e V+(det)+N+(adv)+Adj (*ter um caráter Adj*) (*ter um valor Adj*):

- *Chicago é uma cidade que tem alma,* ***tem caráter.***

- *Foram tenistas que não precisavam de «personalidade», porque* ***tinham caráter.***
- *Esta punição ao tenente-coronel* ***tem um caráter exemplar*** *?*
- *Portanto, considero que isso* ***tem um caráter institucional****.*
- *Ser negro é ser a paz,* ***ter valor.***
- *Os camaroneses* ***tiveram valor,*** *os brasileiros, classe», afirma o jornal .*
- *Um aluno desse* ***tem um valor incrível.***
- *No final, quando acabo, sinto que eles podem* ***ter um valor alegórico,*** *simbólico.*

O mesmo ocorre com *ter validade*. Já *ter defeito* não segue o mesmo padrão. Aparece também intercalado do artigo indefinido com uma estrutura intransitiva:

- *Os figos só* ***têm um defeito****: não são recomendados para quem está de dieta .*
- *"Quando um filme* ***tem um defeito****, é condenado como obra mal realizada" .*

No Teste 3 ("DAR"):

As expressões *dar banho*, *dar medo*, *dar voto*, por exemplo, foram consideras CMs pela sua intransitividade em algumas ocorrências.

- *Troco fraldas,* ***dou banho,*** *e depois me tranco lá dentro por pelo menos três horas por dia .*
- *Você vai ter que* ***dar banho,*** *talvez trocar fraldas .*
- *Mulher independente* ***dá medo.***
- *O Brasil* ***dá medo.***
- *É, ferrovia nunca* ***deu voto.***
- *Quer aquilo que* ***dá voto.***

Embora inusitada, a expressão *dar outras providências* é freqüente no córpus em virtude de seu teor jurídico:

- *Lei 6.766, de 19.12.79 - Dispõe sobre o parcelamento do solo urbano e* ***dá outras providências*** *(CCLCV 352, Lex 1979/1.008, Bo)l .*

• *215: 2 a. LC 73, de 10.2.93 - Institui a Lei Orgânica da Advocacia-Geral da União e* ***dá outras providências*** *(Lex 1993/82, RF 321 / 420) :*

• *Lei 8.021, de 12.4.90 - Dispõe sobre a identificação dos contribuintes para fins fiscais, e* ***dá outras providências*** *(Lex 1990/518, RDA 179 / 348):*

No Teste 4 ("PERDER"):

De uma forma ampla, o teste não gerou ruídos em relação a CMs encabeçadas pelo verbo "perder". Isso demonstra não só que o formalismo utilizado para detectar as CMs tem um bom poder de abrangência mas também que o verbo "perder" é vocacionado a encabeçar CMs do padrão procurado.

Pode-se questionar a identificação de "perder móveis" como uma CM freqüente; o que pode ser explicado pelo teor jornalístico do córpus.

• *Vizinho de Deosdeti, o bombeiro* ***perdeu móveis,*** *mas retornou ontem à rua onde mora para se assegurar de que não roubaram o que lhe restou.*

Notamos um padrão enumerativo com esta expressão:

• *Na favela da Santa Cruz 2, 18 famílias* ***perderam móveis,*** *roupas e mantimentos .*

• *Pelo menos cem famílias de classe média na Vila Ema foram atingidas e* ***perderam móveis,*** *eletrodomésticos e mantimentos .*

• *Os moradores* ***perderam móveis,*** *objetos e até carros .*

No teste 5 ("USAR"):

Aqui houve um ruído específico gerado não pelo método estatístico implementado, mas pela época em que o córpus foi compilado. As CMs *usar a URV* e *usar URV:*

• *A Receita Federal não quer* ***usar a URV.***

• *A outra é sobre como será o relacionamento com os fornecedores, já que uns* ***usarão URV*** *e outros não .*

• *Com isso as empresas podem* ***usar a URV****, mas não são obrigadas a adotar esse procedimento», explicou Roberto Padovani, da Secretaria de Política Econômica do Ministério da Fazenda .*

No Teste 6 (“RECEBER”):

A expressão “receber tratamento”, por exemplo, ocorre freqüentemente com padrão intransitivo, com um uso bastante definido:

• *Para o ministério, estudantes com visão inferior à normal devem* ***receber tratamento*** *.*

• *A foca foi levada para o zôo, em São Cristóvão, onde* ***receberá tratamento*** *.*

• *Até as 6h de sábado, ele permanecia em uma maca* ***recebendo tratamento*** *.*

Já quando o padrão é “receber +um+ tratamento”, a janela sintática parece ser “receber +um+ tratamento+ (Adv)+Adj”, o que aumenta a possibilidade de aplicações para seus usos:

• *E, a de uma clínica de emagrecimento, a Proforma, a cantora de 19 anos* ***recebeu um tratamento*** *gratuito .*

• *Sendas já adiantou que os fornecedores alinhados com a promoção* ***receberão um tratamento*** *vip .*

• *Os candidatos à Presidência da República e a vice* ***recebem um tratamento*** *ainda mais vip.*

• *E vai* ***receber um tratamento*** *diferenciado .*

• *Molina afirmou que o leite in natura distribuído para a população não* ***recebe um tratamento*** *adequado .*

• *A grande inovação é o tecido de fibras sintéticas que* ***recebe um tratamento*** *antichamas .*

De uma forma geral, o verbo “receber” gerou poucos ruídos (3 pseudo-CMs: *receber um tratamento, receber livre, receber este ano*).

Teste 7 (“DEIXAR”):

O verbo “deixar” também demonstra ser vocacionado a encabeçar CMs do padrão procurado. Houve apenas 5 ruídos. Dentre eles, um merece destaque: “deixar patente”. A ambigüidade do termo e a posição que exercia na estrutura causaram uma etiquetagem equivocada pelo córpus. “Patente”, nesse caso, é adjetivo e não um substantivo:

- *O que há de alegórico não propriamente no filme, mas na sua difusão televisual, é que ele* ***deixa patente*** *(pela inserção dos peixinhos coloridos na cena monocromática) e ao mesmo tempo tematiza no enredo a lacuna que é o preto-e-branco televisual* .

Por outro lado, não houve nenhuma ocorrência no córpus da possível CM “deixar (det) patente” (“deixar o cargo militar”).

“Deixar pistas” e “deixar rastros”, por sua vez, foram consideradas pseudo-CMs (JAN) por só terem sido detectadas precedidas pelo advérbio *não* ou pela preposição *sem* :

- *«A explicação é que o garoto resolve assaltar o cliente e, para não* ***deixar pistas****, acaba matando», diz Mott* .
- *O jogo consiste em matar pessoas violentamente e não* ***deixar pistas*** .
- *Há cinco anos, um Cessna 720, com capacidade para quatro passageiros, foi levado do aeroporto sem* ***deixar pistas*** .
- *À moda de um bom criminoso hitchcockiano: sem* ***deixar rastros*** .
- *que a ex-senhora Mick deixou algumas jóias no cofre do hotel do amigo, o Morgans Hotel, em NY, e elas desapareceram sem* ***deixar rastros*** .
- *Tinha, pois, charme suficiente para cumprir breve passagem no Santos, aquele Santos de Pelé e cia. Tão breve que não* ***deixou rastros*** .

No Teste 8 (“TOMAR”):

A expressão “tomar conta”, por exemplo não foi considerada uma pseudo-CM, uma vez que aparece em estruturas intransitivas com um teor específico:

- *As pessoas pensavam que, sem as crenças religiosas tradicionais, o cinismo ia* ***tomar conta.***

- *Durante os primeiros meses, pratica Tai Chi com entusiasmo mas aos poucos o cansaço vai* ***tomando conta,*** *e resolve desistir .*
- *Havíamos atracado num pequeno desembarcadouro, sede de um projeto governamental abandonado, onde a margem arenosa se estendia por mais de 30 metros antes da floresta* ***tomar conta.***
- *Se o ator não souber parar, a vida profissional vai sempre* ***tomar conta.***
- *Aí vem a mente, bocejando, e* ***toma conta.***

Por outro lado, "tomar o AZT", "tomar AZT" foram consideradas ruídos pela sua freqüência ter sido gerada em função da época em que o córpus foi coletado:

- *Um remédio usado contra herpes, o acyclovir, aumentou em 44 % a sobrevivência de pacientes de Aids que também* ***tomavam o AZT,*** *segundo estudo publicado ontem na revista «Annals of Internal Medicine» .*
- *Dois meses atrás, a Newsweek publicou uma notícia afirmando que ele estava* ***tomando AZT*** *e que tinha visitado uma clínica no Quênia para tratamento .*

No Teste 9 ("GANHAR"):

Tanto a expressão "ganhar vida" quanto "ganhar a vida" foram consideradas CMs. A primeira, em razão de sempre desempenhar uma função intransitiva no córpus:

- *Ela também divide o palco com objetos que* ***ganham vida.***
- *Espero que a experiência tome curso diferente e* ***ganhe vida.***

Já a segunda, por freqüentemente desempenhar uma função intransitiva:

- *Fazer filmes é uma profissão, um jeito de* ***ganhar a vida.***
- *"Eu estou* ***ganhando a vida!*** *", me diz o cu-a-jato- com um jeito meio sindicalista .*
- *"Temos que* ***ganhar a vida.*** *";* além de seu papel transitivo indireto tradicional:

• *Quatro vezes campeão mundial de F-1, Alain Prost* ***ganha a vida,*** *hoje em dia, como comentarista da TF1, canal de TV da França .*

Um ruído provocado pelo verbo "ganhar", foi gerado pela ambigüidade do termo "ganho" (substantivo), detectado como forma flexionada do verbo. Houve, conseqüentemente, a identificação de pseudo-CMs como "ganhar real", "ganhar líquido" e "ganhar financeiro".

• ***Ganho líquido:*** *4,94 % .*

• *Os analistas recomendam a retirada dos recursos somente dez dias depois do depósito, no mínimo, para garantir* ***ganho real.***

• *Com a entrada da URV, os supermercados deixaram de ter* ***ganho financeiro,*** *diz Diniz .*

No Teste 10 ("VER"):

Conforme já argumentamos, as CMs encabeçadas pelo verbo "ver" foram excluídas em virtude de sua freqüente função metatextual. "Ver quadro", por exemplo, ocorre 150 vezes no córpus. "Ver texto", por sua vez, ocorre 65 vezes. Claramente, a sua posição entre os 10 verbos mais recorrentes com o padrão procurado está comprometida pelo córpus.

No Teste 11 ("CRIAR"):

A expressão *criar chances* foi considerada uma CM pela sua função freqüentemente intransitiva em parágrafos referentes às seções de Esportes.

• *Na prorrogação, os dois times* ***criaram chances****, mas ninguém marcou .*

• *O Corinthians teve maior posse de bola, atacou,* ***criou chances*** *e chutou mais contra o gol de Velloso, mas não conseguiu ir além do empate por 1 a 1, que deu o bicampeonato ao Palmeiras, ontem no Pacaembu .*

• *O trio, porém, não conseguiu* ***criar chances*** *.*

O mesmo foi verificado em relação a ***criar um ambiente*** e ***criar um cenário***:

• *Então foi* ***criado um ambiente.***

• *No fundo, você está* ***criando um cenário****, criando uma falsa idéia .*

* * *

Em suma, o que consideramos especialmente relevante nesta abordagem com base em córpus, é que não fazemos conjecturas daquilo que ocorre e não ocorre em uma língua, pois, como já demonstramos no capítulo 2, uma perspectiva exclusivamente intuitiva pode ser muitas vezes contra-argumentada pelo córpus. Nosso olhar eminentemente empírico é capaz de detectar preferências de usos ao invés de intuir aquilo que pode ou não ocorrer em um córpus, com base em testes de aceitabilidade, também já questionados no capítulo 2.

As listas referentes aos TESTES 1 a 11 estão apresentadas em ordem decrescente de freqüência de cada verbo com o padrão V+(det)+N procurado. As pseudo-CMs estão indicadas por parênteses, seguidas pela sigla do ruído.

Na verdade, a listagem de cada verbo testado segue até que se chegue a CMs menos freqüentes (com uma única ocorrência). Optamos por expor as 100 CMs mais recorrentes encabeçadas por cada um dos dez verbos por razões metodológicas.

Nas 10 listas, cada CM é acompanhada pelo seu ranqueamento e freqüência, respectivamente. Tomemos como exemplo o primeiro bigrama do TESTE 1, [*fazer campanha*]: o número que segue imediatamente a CM (1) diz respeito à sua posição no córpus em relação às outras CMs encabeçadas pelo mesmo verbo. Já o segundo número (117) se refere ao número de ocorrências da CM [*fazer campanha*] e de suas diversas flexões no córpus. Note-se que as CMs estão agrupadas na forma canônica do verbo, mas são identificadas em todas as suas formas flexionadas.

Abreviaturas:

(ETQ): etiquetagem equivocada no córpus
(JAN): janela sintática menor do que a expressão representa
(DAT): CMs claramente datadas
(MET): metatexto do córpus
(COE): Recursos coesivos
(MAN):manchetes de jornal

Teste 1: FAZER +(DET)+N

fazer campanha,1,117
fazer parte,2,106,
fazer sucesso,3,96
fazer sentido,4,94
fazer compras,5,73
fazer falta,6,57
fazer perguntas,7,51
fazer alguma coisa,8,44
fazer política,9,44
fazer as contas,10,42(COE)
fazer gols,11,40
(fazer mal),12,39 (JAN)
fazer o gol,13,38 (COE)
fazer greve,14,36
fazer sexo,15,36
fazer filmes,16,35
fazer ginástica,17,34
fazer cinema,18,34
fazer amor,19,32
fazer um filme,20,30
fazer a conversão,21,30 (COE)
fazer comício,22,30
(fazer muito=tempo),23,29 (JAN)
fazer diferença,24,29
fazer água, 25,28
fazer um teste,26,27
fazer o pedido,27,26
fazer as coisas,28,26 (COE)
fazer exercícios,29,25
fazer a festa,30,25
fazer o teste,31,24 (COE)
fazer boca-de-urna,32,24
fazer a revisão,33,24 (COE)
(fazer coisas),34,24 (JAN)
fazer shows,35,24
fazer tempo,36,23
fazer a mesma coisa,37,23 (COE)
(fazer uma coisa),38,23 (JAN)
fazer teatro,39,22
fazer o filme,40,21 (COE)
fazer exames,41,21
fazer concessões,42,21
fazer comentários,43,21
fazer escola,44,20
fazer teste,45,20
fazer outra=coisa,46,20 (COE)
fazer palestra,47,20
fazer gol,48,20
fazer testes,49,20
fazer contas,50,19
fazer dinheiro,51,19
fazer a denúncia,52,19 (COE)
fazer propaganda,53,19
fazer a sua parte,54,19,1
fazer carreata,55,19,1
fazer acordo,56,18,1
fazer uma ressalva,57,18,1
fazer música,58,18,1
fazer a diferença,59,18,1
fazer milagres,60,18,1
fazer coro,61,18,1
fazer barulho,62,18
fazer críticas,63,18
fazer a prova,64,17(COE)
fazer o exame,65,17(COE)
fazer ameaças,66,17
fazer um acordo,67,17
fazer justiça,68,16
fazer doações,69,16
fazer as malas,70,16
fazer a defesa,71,15 (COE)
fazer sua estréia,72,15 (COE)
fazer o trabalho,73,15 (COE)
fazer show,74,15
fazer assembléia,75,15
fazer acordos,76,15
fazer dieta,77,14
fazer oposição,78,14
fazer a reforma=agrária,79,14 (COE)
fazer reféns,80,14
fazer amigos,81,14
fazer uma curva,82,14
fazer pesquisas,83,13
fazer a barba,84,13

fazer seu sucessor,85,13 (COE)
fazer previsões,86,13
fazer reservas,87,13
fazer o pagamento,88,13 (COE)
fazer essas declarações,89,13 (COE)
fazer a coisa,90,12 (COE)
fazer sol,91,12
(fazer um trabalho),92,12 (JAN)
fazer promoções,93,12
(fazer tanto sucesso),94,12 (JAN)
fazer mudanças,95,12
fazer aborto,96,12
fazer análise,97,12
fazer mais gols,98,11 (COE)
fazer alianças,99,11
fazer turismo,100,11
fazer muito sucesso,101,11
fazer uma pesquisa,102,11
fazer o serviço,103,11(COE)
fazer um discurso,104,11
fazer o projeto,105,11(COE)
fazer festa,106,11

Teste 2: TER+(DET)+N

ter razão,1,161
ter problemas,2,135
(ter direito),3,119 (JAN)
ter medo,4,114
ter uma idéia,5,107
(ter acesso),6,94 (JAN)
ter filhos,7,92
(ter certeza),8,75 (JAN)
ter dinheiro,9,68
ter sucesso,10,67
ter dúvidas,11,61
ter início,12,60
ter tempo,13,58
ter alta,14,54
ter jeito,15,53
ter paciência,16,50
ter sorte,17,49
ter um filho,18,48
ter problema,19,45
(ter nome),20,40 (JAN)
ter importância,21,40
ter dificuldades,22,32
ter limites,23,32
ter experiência,24,30
ter valor,25,30
ter dúvida,26,30
ter validade,27,29
ter candidato,28,29
ter uma visão,29,29
ter notícia,30,29
ter fim,31,28
ter condições,32,28
ter êxito,33,27
ter preço,34,26
ter fundamento,35,25
ter chance,36,25
ter coragem,37,24
ter força,38,24
ter uma filha,39,24
ter prejuízo,40,24
ter solução,41,23
ter telefone,42,23
ter culpa,43,23
ter chances,44,22
ter pressa,45,22
ter qualidade,46,22
(ter interesse),47,20 (JAN)
ter fôlego,48,20
ter queda,49,20
ter espaço,50,20
ter maioria,51,20
ter provas,52,20
ter escrúpulos,53,20
(ter conhecimento),54,19 (JAN)
ter uma fazenda,55,19
ter idéias,56,19
ter talento,57,19
ter poder,58,19
ter a mesma opinião,59,18(COE)
ter futuro,60,18
ter a doença,61,18(COE)
ter esperança,62,18
ter mais chances,63,17
ter uma explicação,64,17
(ter a impressão),65,17 (JAN)
ter defeitos,66,17
ter lugar,67,17
ter cura,68,17
ter prioridade,69,17
ter muito=tempo,70,16
ter muito dinheiro,71,16
ter música,72,16
ter saída,73,16
ter efeito,74,16
ter credibilidade,75,16
ter vez,76,16
(ter um caráter),77,15 (JAN)
ter dono,78,15
(ter preços),79,15(JAN)
(ter seu preço),80,15 (JAN)
ter história,81,15
(ter gente),82,15(JAN)

ter um problema,83,15
ter uma relação,84,14
ter memória,85,14
ter lucro,86,14
ter mais votos,87,14
ter aumento,88,14
ter continuidade,89,14
(ter recursos),90,14 (JAN)
ter câncer,91,14
ter amigos,92,14
ter liberdade,93,14
(ter vontade),94,14 (JAN)
ter inflação,95,13
ter transmissão,96,13
(ter um papel),97,13 (JAN)
(ter um desempenho),98,13 (JAN)
ter a bola,99,13
ter resultados,100,13
ter preferência,101,13
ter um estilo,102,13
(ter idéia),103,13 (JAN)
ter um jogo,104,13
(ter coisa),105,12(JAN)
ter fé,106,12
ter computador,107,12
ter erro,108,12
ter efeitos,109,12 (JAN)
ter caráter,110,12
ter água,111,12
ter um impacto,112,12
ter um efeito,113,12
ter cabimento, 114,12
ter piscina,115,12
ter um preço,116,12
(ter mais dinheiro),117,11(JAN)
ter um custo,118,11
ter uma reunião,119,11

Teste 3: DAR+(DET)+N

dar entrevistas,1,115
dar entrevista,2,67
dar resultado,3,46
(dar a volta),4,42 (JAN)
(dar declarações),5,42 (JAN)
dar lucro,6,38
dar sorte,7,37
dar aulas,8,37
dar um exemplo,9,36
dar autógrafos,10,35
dar início,11,29
(dar conta),12,28 (JAN)
(dar explicações),13,27 (JAN)
dar tempo,14,26
dar o troco,15,25
dar dinheiro,16,24
dar resultados,17,23
dar prejuízo,18,22
dar um tempo,19,22
dar um salto,20,22
dar trabalho,21,22
dar o tom,22,21
dar um passo,23,20
dar as caras,24,20
dar risada,25,19
dar detalhes,26,19
(dar o nome),27,18 (JAN)
dar o exemplo,28,18 (COE)
dar força,29,17
dar apoio,30,17
dar espetáculo,31,16
(dar origem),32,16 (JAN)
(dar momento),33,16 (ETQ)
dar a resposta,34,15 (COE)
(dar comida),35,14(JAN)
dar atenção,36,14
(dar sustentação),37,13 (JAN)
dar medo,38,13
dar azar,39,13
dar descontos,40,13
(dar continuidade),41,13 (JAN)
dar frutos,42,13
(dar lugar),43,12 (JAN)
(dar informações),44,12(JAN)
dar aula,45,12
dar as cartas,46,12
dar muito trabalho,47,12
dar uma entrevista,48,11
dar vexame,49,11
(dar uma idéia),50,11 (JAN)
(dar espaço),51,11 (JAN)
dar as mãos,52,11
dar nomes,53,11
(dar bola),54,11 (JAN)
dar outras providências,55,11
dar um jeito,56,11
dar porrada,57,10
(dar ouvidos),58,10 (JAN)
dar palestras,59,10
(dar opinião),60,10 (JAN)
(dar razão),61,9 (JAN)
dar a partida,62,9
dar prazer,63,9
dar respostas,64,9
dar voto,65,9
(dar a impressão),66,9 (JAN)
dar conselhos,67,9
dar um conselho,68,9
dar retorno,69,9
dar uma força,70,9
(dar a notícia),71,9 (COE)
dar entrada,72,8
dar exemplos,73,8
dar palestra,74,8
dar orientações,75,8
dar problema,76,8
dar uma guinada,77,8
(dar a vitória),78,8 (JAN)
(dar um carro),79,7 (JAN)
(dar alegria),80,7 (JAN)
dar palpite,81,7
dar pena,82,7
dar bandeira,83,7
dar dicas,84,7
dar futuro,85,7

dar manchete,86,7
dar seu parecer,87,7(COE)
dar uma festa,88,7
dar chance,89,7
(dar sinais),90,7 (JAN)
dar tiros,91,7
dar pênalti,92,7
dar banho,93,7
dar pé,94,7
dar samba,95,7
dar problemas,96,7
dar resposta,97,6
dar as informações,98,6 (COE)
(dar sequência),99,6 (JAN)
dar mais trabalho,100,6
(dar esclarecimentos),101,6 (JAN)
(dar valor),102,6 (JAN)
dar prestígio,103,6
(dar prosseguimento),104,6 (JAN)
dar voltas,105,6
dar troco,106,6
dar votos,107,6
(dar destaque),108,6 (JAN)
dar um show,109,6
dar um toque,110,6
(dar membro),111,6 (ETQ)
dar show,112,6
dar a receita,113,6
(dar espaços),114,5 (JAN)
dar um abraço,115,5
dar o dinheiro,116,5
(dar um caráter),117,5 (JAN)
dar jeito,118,5
dar os nomes,119,5 (COE)
dar confiança,120,5
dar alguns exemplos,121,5
dar o recado,122,5 (COE)
dar sua opinião,123,5 (COE)
dar queixa,124,5
dar a informação,125,5 (COE)
dar sugestões,126,5
dar um parecer,127,5
dar fama,128,5
dar licença,129,5
(dar cobertura),130,5 (JAN)
dar esmola,131,5

Teste 4: PERDER+(DET)+N

perder tempo,1,54
perder a eleição,2,46
perder o controle,3,44
perder dinheiro,4,41
perder o emprego,5,38
perder peso,6,28
perder a bola,7,25
perder o mandato,8,23
perder força,9,20
perder a direção,10,19
perder o fôlego,11,17
perder a cabeça,12,15
perder a validade,13,14
perder o sentido,14,14 (COE)
perder as eleições,15,13
perder o título,16,13 (COE)
perder espaço,17,13
perder o rumo,18,13
perder a pose,19,12
perder o ritmo,20,11
perder o registro,21,11
perder a mãe,22,11
perder o jogo,23,11(COE)
perder a graça,24,10
perder um pênalti,25,10
perder a paciência,26,10
perder votos,27,10
perder valor,28,10
perder a razão,29,9
perder a copa,30,9
perder a virgindade,31,9
perder a oportunidade,32,9
perder o equilíbrio,33,9
perder o cargo,34,9
perder sentido,35,9
perder terreno,36,8
perder o interesse,37,8 (COE)
perder pontos,38,8
perder o pai,39,8
perder seus empregos,40,8(COE)
perder a calma,41,8
perder a força,42,8 (COE)
perder fôlego,43,8
perder competitividade,44,8
perder a esperança,45,8
perder validade,46,8
perder a guerra,47,7 (COE)
perder o pique,48,7
perder emprego,49,7
perder credibilidade,50,7
perder a majestade,51,7
perder móveis,52,7
perder mercado,53,7
perder o medo,54,6
perder o bom=humor,55,6
perder gols,56,6
perder o costume,57,6
perder o valor,58,6 (COE)
perder a qualidade,59,6
perder a mulher,60,6
perder a memória,61,6
perder a partida,62,6(COE)
perder as esperanças,63,6
perder a vaga,64,5
perder suas casas,65,5 (COE)
perder a posição,66,5
perder o lugar,67,5
perder seu mandato,68,5(COE)
perder o gol,69,5
perder a chance,70,5(COE)
perder eleições,71,5
perder os dedos,72,5
perder o marido,73,5
perder clientes,74,5
perder eficácia,75,5
perder a voz,76,5
perder o rebolado,77,5
perder eficiência,78,5
perder muito=tempo,79,5,
perder a ação,80,5 (COE)
perder a liderança,81,5
perder o filho,82,5 (COE)
perder a concentração,83,5
perder a causa,84,5 (COE)
perder chance,85,4
perder seus mandatos,86,4,(COE)

perder importância,87,4
perder o foco,88,4
perder várias chances,89,4
perder pênalti,90,4(COE)
perder um jogador,91,4
perder o apetite,92,4
perder a conta,93,4
perder a vez,94,4
perder os pontos,95,4
perder a hora,96,4
perder os empregos,97,4(COE)
perder o bonde,98,4
perder a identidade,99,4
perder a perna,100,4

Teste 5: USAR+(DET)+N

usar camisinha,1,43
usar o cinto,2,36 (COE)
usar drogas,3,36
usar cinto=de=segurança,4,19
usar cinto,5,15
usar óculos,6,14
(usar a urv),7,12 (DAT)
usar a camisinha,8,12 (COE)
usar computador,9,12
usar preservativo,10,12
usar o computador,11,11 (COE)
usar o equipamento,12,10 (COE)
usar o micro,13,10
usar o serviço,14,10 (COE)
usar a força,15,9
usar terno,16,9
usar brinco,17,9
usar o cinto=de=segurança,18,8 (COE)
usar armas,19,8
(usar a máquina),20,8 (JAN)
usar calcinha,21,8
usar óculos=escuros,22,7
usar a cabeça,23,7
(usar palavras),24,7 (JAN)
usar o sistema,25,7
usar cocaína,26,7
usar computadores,27,6
usar o telefone,28,6
usar camiseta,29,6
usar a droga,30,6,1(COE)
usar capacetes,31,6,1
usar chapéu,32,6,1
usar o dinheiro,33,6,1(COE)
(usar roupas),34,6,1(JAN)
usar a rede,35,5,1
usar as mãos,36,5,1
usar helicópteros,37,5,1
usar música,38,5,1
usar a piscina,39,5,1
usar o programa,40,5,1(COE)
usar luvas,41,5,1
usar a igreja,42,5,1(COE)
usar batom,43,5,1
usar água,44,4,1
usar armas=de=fogo,45,4,1
usar gravata,46,4,1
usar os equipamentos,47,4,1
usar a violência,48,4,1
usar roupa,49,4,1
usar a palavra,50,4,1
usar meias,51,4,1
usar a gráfica,52,4,1 (COE)
usar o carro,53,4,1 (COE)
usar o dólar,54,4,1
usar uniforme,55,3,1
usar a técnica,56,3,1(COE)
usar uma faca,57,3,1
(usar termos),58,3,1(JAN)
(usar o jargão),59,3 (JAN)
usar máscaras,60,3
usar um fone,61,3
usar uma camisinha,62,3
(usar expressões),63,3(JAN)
usar essa palavra,64,3(COE)
(usar uma imagem),65,3 (JAN)
usar máscara,66,3
usar fax,67,3
usar crachás,68,3
usar o líbero,69,3 (COE)
usar véu,70,3
usar barba,71,3
(usar o livro),72,3 (JAN)
usar seu cartão,73,3 (COE)
(usar faixas),74,3 (JAN)
usar bonés,75,3
usar o silêncio,76,3
usar cheque,77,3
usar maquiagem,78,3
usar esse recurso,79,3 (COE)
(usar o local),80,3 (JAN)
usar o mesmo argumento,81,3 (COE)
(usar as importações),82,3 (JAN)
(usar o corpo),83,3 (JAN)
(usar instrumentos),84,3 (JAN)
usar lentes,85,3
usar o preservativo,86,3 (COE)

usar ternos,87,3
usar cueca,88,3
usar o bom=senso,89,3
(usar imagens),90,3 (JAN)
usar pseudônimo,91,3
usar droga,92,3
(usar tecnologia),93,3 (JAN)
usar os programas,94,3 (COE)
usar passes,95,3
usar bigode,96,3
(usar o púlpito),97,3 (JAN)
usar cores,98,3
usar crediário,99,3
usar a tecnologia,100,2
usar sabão,101,2
usar pagers,102,2
(usar urv),103,2 (DAT)
usar a história,104,2
usar um computador,105,2
usar uma peruca,106,2
usar esse instrumento,107,2 (COE)
usar microcomputadores,108,2
usar brincos,109,2
usar disfarces,110,2
usar carro,111,2
usar o logotipo,112,2 (COE)
usar o sol,113,2
usar as palavras,114,2
usar o cartão=de=crédito,115,2
usar líbero,116,2
(usar amortecedores),117,2 (JAN)
usar outros índices,118,2 (COE)
(usar o hotel),119,2 (JAN)
usar flash,120,2
(usar a música),121,2 (JAN)
(usar envelopes),122,2 (JAN)
(usar a religiosidade),123,2 (JAN)
(usar bonecos),124,2 (JAN)
usar as mesmas armas,125,2 (COE)

Teste 6: RECEBER+(DET)+N

receber o dinheiro,1,41(COE)
receber alta,2,37
receber ameaças,3,27
receber a bola,4,26
receber dinheiro,5,25
receber propinas,6,19
receber a notificação,7,16(COE)
receber visitas,8,15
receber o carro,9,15 (COE)
receber salário,10,14
receber informações,11,12
receber prêmios,12,11
receber o benefício,13,10 (COE)
receber elogios,14,10
receber o produto,15,9 (COE)
receber água,16,9
receber salários,17,9
receber resposta,18,9
receber os bônus,19,9 (COE)
receber o salário,20,9
receber doações,21,8
receber denúncias,22,8
receber o pagamento,23,8
receber o processo,24,7 (COE)
receber a notícia,25,7 (COE)
receber o prêmio,26,7 (COE)
receber seu dinheiro,27,7 (COE)
receber um telefonema,28,7
receber tratamento,29,7
receber a carta,30,7(COE)
receber seu salário,31,7 (COE)
receber o mesmo tratamento,32,7 (COE)
receber ajuda,33,7
receber críticas,34,7
receber a indenização,35,6 (COE)
receber fax,36,6
receber uma ameaça,37,6
receber armas,38,6
receber os salários,39,6
receber a diferença,40,6
receber a folha,41,6 (COE)
receber apoio,42,6
receber recursos,43,6
receber o troféu,44,5
(receber este ano),45,5 (JAN)
receber a imprensa,46,5
receber treinamento,47,5
receber o pedido,48,5
receber a encomenda,49,5 (COE)
receber a denúncia,50,5 (COE)
receber pensão,51,5
receber benefícios,52,5
receber o documento,53,5 (COE)
receber comida,54,4
receber a restituição,55,4 (COE)
receber seus salários,56,4 (COE)
receber telefonemas,57,4
receber a vacina,58,4 (COE)
receber amigos,59,4
receber propina,60,4
receber os amigos,61,4
receber salário=mínimo,62,4
(receber um tratamento),63,4 (JAN)
receber uma carta,64,4
receber oxigênio,65,4
receber um prêmio,66,4
receber propostas,67,4
receber cartas,68,4
receber inscrições,69,4
(receber livre),70,4 (ETQ)
receber dividendos,71,4
receber muitas cartas,72,4
receber prêmio,73,4 (COE)
receber a pensão,74,4 (COE)
receber o visto,75,4
receber a aposentadoria,76,4
receber o troco,77,3
receber a revista,78,3 (COE)
receber um mínimo,79,3
receber esse dinheiro,80,3 (COE)
receber reais,81,3
receber presentes,82,3
receber alimentos,83,3

receber a informação,84,3 (COE)
receber reclamações,85,3
receber lançamento,86,3
receber horas=extras,87,3
receber tiros,88,3
receber os cumprimentos,89,3
receber o telefonema,90,3 (COE)
receber os dólares,91,3 (COE)
receber mais passes,92,3 (COE)
receber esse tratamento,93,3 (COE)
receber orientação,94,3
receber uma vaga,95,3
receber mensagens,96,3
receber a ligação,97,3(COE)
receber a conta,98,3 (COE)
receber os passes,99,3 (COE)
receber ameaça,100,3
receber alimentação,101,3
receber a fatura,102,3 (COE)
receber juros,103,3

Teste 7: DEIXAR (DET) N

deixar o cargo,1,98
deixar o país,2,77
deixar o governo,3,69
deixar o local,4,40
deixar filhos,5,40
deixar a cidade,6,26
deixar o ministério,7,24
deixar o poder,8,22
deixar o partido,9,20
deixar dúvidas,10,18
deixar o time,11,17
deixar a equipe,12,16
deixar o clube,13,16
deixar a fazenda,14,15 (COE)
deixar o campo,15,15
deixar a prisão,16,13
deixar vítimas,17,12
deixar o carro,18,12
deixar a favela,19,12
deixar a prefeitura,20,11
deixar suas casas,21,9 (COE)
deixar o futebol,22,9
deixar a casa,23,9 (COE)
deixar a empresa,24,9
deixar o hotel,25,9
deixar a seleção,26,9
deixar sua casa,27,8 (COE)
deixar filha,28,8
deixar o hospital,29,8
deixar marcas,30,8
deixar a sala,31,8
deixar a presidência,32,8
deixar o emprego,33,7
(deixar as coisas),34,7 (JAN)
deixar cargo,35,7
deixar sua marca,36,7(COE)
deixar a área,37,6
deixar seu país,38,6 (COE)
deixar o palmeiras,39,6
deixar o prédio,40,6 (COE)
deixar vestígios,41,6
deixar o apartamento,42,6
deixar a escola,43,5
deixar saudade,44,5
deixar o morro,45,5
deixar o gabinete,46,5
deixar as drogas,47,5
deixar seus cargos,48,5 (COE)
deixar dúvida,49,5
deixar o plenário,50,4
(deixar patente),51,4 (ETQ)
deixar a delegacia,52,4
deixar seqüelas,53,4
deixar o mercado,54,4 (COE)
deixar os cargos,55,4 (COE)
deixar o basquete,56,4
deixar a mulher,57,4
deixar barato,58,4
deixar as quadras,59,4
deixar o palco,60,4
deixar a quadra,61,4
deixar o aeroporto,62,4
deixar o gramado,63,4
deixar a universidade,64,3
deixar o estádio,65,3
deixar o fluminense,66,3
deixar o quarto,67,3
deixar essa posição,68,3 (COE)
deixar o vôlei,69,3
(deixar pistas),70,3 (JAN)
deixar o grupo,71,3 (COE)
deixar recados,72,3
deixar a emissora,73,3 (COE)
deixar o pai,74,3
deixar o jogo,75,3
deixar a política,76,3
deixar a clínica,77,3 (COE)
deixar seu voto,78,3 (COE)
deixar o senado,79,3 (COE)
deixar a embarcação,80,3
deixar a profissão,81,3
deixar a receita,82,3 (COE)
deixar Gaza,83,3
(deixar as pessoas),84,3 (JAN)
deixar rastros,85,3
deixar suas funções,86,3 (COE)

deixar os filhos,87,3
(deixar abertas),88,3 (ETQ)
deixar o estado,89,3
deixar o palanque,90,3 (COE)
deixar a propriedade,91,3 (COE)
deixar a capital,92,3
deixar a copa,93,3 (COE)
(deixar os jogadores),94,3 (JAN)
deixar o banco,95,3
deixar a defesa,96,3
deixar o presídio,97,3
deixar testemunhas,98,3
deixar o exército,99,3
deixar um bilhete,100,3
deixar o dinheiro,101,3 (COE)
deixar o caso,102,2
deixar os campos,103,2
(deixar o resto),104,2 (JAN)
deixar a legenda,105,2
deixar o bairro,106,2

Teste 8: TOMAR (DET) N

tomar posse,1,164
tomar banho,2,58
tomar decisões,3,52
tomar café,4,43
tomar uma decisão,5,33
tomar conhecimento,6,29
tomar providências,7,26
tomar cuidado,8,26
tomar o poder,9,22
tomar sol,10,22
tomar a decisão,11,20 (COE)
tomar a iniciativa,12,15 (COE)
tomar um banho,13,13
tomar gols,14,12
tomar conta,15,12
tomar a bola,16,12
(tomar medidas),17,11(JAN)
tomar chá,18,11
tomar cerveja,19,11
tomar partido,20,11
tomar tempo,21,10
tomar forma,22,9
tomar uma atitude,23,9
tomar drogas,24,9
(tomar posição),25,9 (JAN)
tomar alguns cuidados,26,9
tomar café=da=manhã,27,8
tomar remédios,28,8
tomar uma providência,29,8
(tomar as providências),30,8 (JAN)
tomar sorvete,31,8
tomar coragem,32,8
tomar a cidade,33,7 (COE)
tomar uma posição,34,7
tomar essa decisão,35,7 (COE)
tomar fôlego,36,6
tomar o remédio,37,6 (COE)
tomar água,38,6
tomar esta decisão,39,6 (COE)
tomar algumas precauções,40,6
tomar um táxi,41,6
tomar um ônibus,42,6
(tomar qualquer providência) 43, 6 (JAN)
tomar sua decisão,44,6 (COE)
tomar a mesma decisão,45,6 (COE)
tomar jeito,46,6
(tomar dinheiro),47,5 (JAN)
tomar um café,48,5
tomar vitaminas,49,5
tomar vinho,50,5
tomar a droga,51,5 (COE)
tomar corpo,52,5
(tomar parte),53,5 (JAN)
tomar um chope,54,5
tomar o gol,55,5 (COE)
tomar muito cuidado,56,5
tomar remédio,57,5
(tomar um rumo),58,4 (JAN)
tomar café-da-manhã,59,4
(tomar qualquer decisão),60,4 (JAN)
tomar um susto,61,4
tomar a palavra,62,4
tomar pílula,63,4
tomar gol,64,4
(tomar qualquer atitude),65,4 (JAN)
tomar a vacina,66,4 (COE)
tomar ônibus,67,4
tomar iniciativa,68,4
tomar champanhe,69,4
tomar álcool,70,4
tomar tal decisão,71,4 (COE)
tomar nota,72,4
tomar todas=as decisões,73,4
tomar nenhuma providência,74,3 (JAN)
tomar uma injeção,75,3
tomar mamadeira,76,3
tomar os medicamentos,77,3 (COE)
tomar um cafezinho,78,3
tomar alguma providência,79,3
tomar esta atitude,80,3 (COE)
tomar precauções,81,3
tomar providência,82,3
tomar chope,83,3

tomar um lanche,84,3
tomar juízo,85,3
tomar a dianteira,86,3
tomar pílulas,87,3
tomar notas,88,3
tomar o seu lugar,89,3 (COE)
tomar vergonha,90,3
tomar muito=tempo,91,3
tomar antidepressivos,92,3
tomar aldeia,93,3 (JAN) (MAN)
tomar comprimidos,94,3
tomar sopa,95,3
tomar um porre,96,3
tomar a pílula,97,3 (COE)
(tomar o azt),98,3 (DAT)
tomar o microfone,99,3
tomar veneno,100,3
tomar empréstimo,101,3
tomar leite,102,3
tomar o café=da=manhã,103,3
tomar o comprimido,104,2 (COE)
tomar calmantes,105,2
tomar essas providências,106,2 (COE)
tomar um drinque,107,2
tomar aspirina,108,2
tomar alguma atitude,109,2
tomar a frente,110,2
tomar os cartões,111,2 (COE)

Teste 9: GANHAR (DET) N

ganhar dinheiro,1,101
ganhar a eleição,2,90
ganhar a copa,3,56
ganhar as eleições,4,42
ganhar espaço,5,35
ganhar tempo,6,34
ganhar o jogo,7,32
ganhar força,8,32
ganhar a vida,9,28
ganhar terreno,10,16
ganhar jogo,11,16
ganhar muito dinheiro,12,15
ganhar votos,13,15
ganhar importância,14,15
ganhar destaque,15,12
ganhar o título,16,12
ganhar eleição,17,12
(ganhar real),18,11 (ETQ)
ganhar salário=mínimo,19,11
ganhar mercado,20,11
ganhar corpo,21,11
ganhar peso,22,10
ganhar fama,23,10
ganhar títulos,24,9
ganhar experiência,25,9
ganhar a partida,26,9 (COE)
ganhar um oscar,27,9
ganhar prêmios,28,8
ganhar uma copa,29,8
ganhar confiança,30,8
ganhar impulso,31,8
ganhar algum dinheiro,32,7
ganhar o oscar,33,7
ganhar mais dinheiro,34,7
ganhar o mundo,35,7
ganhar a causa,36,6
ganhar a guerra,37,6 (COE)
ganhar credibilidade,38,6
ganhar pontos,39,6
ganhar fôlego,40,6
ganhar velocidade,41,6
ganhar mais força,42,5
ganhar o campeonato,43,5
(ganhar versão),44,5 (JAN)
ganhar o prêmio,45,5 (COE)
ganhar a presidência,46,5
ganhar alguma coisa,47,5
ganhar medalhas,48,5
ganhar uma eleição,49,5
ganhar autonomia,50,5
ganhar presentes,51,5
ganhar a concorrência,52,5
(ganhar líquido),53,5 (ETQ)
(ganhar financeiro),54,5 (ETQ)
(ganhar contornos),55,4 (JAN)
ganhar adeptos,56,4
ganhar o consumidor,57,4
ganhar o torneio,58,4
ganhar status,59,4
ganhar eleições,60,4
ganhar a posição,61,4
ganhar eficiência,62,4
ganhar manchetes,63,4
ganhar essa eleição,64,4 (COE)
ganhar o brasileiro,65,4(COE)
ganhar intensidade,66,4
ganhar um título,67,4
(ganhar linhas),68,4 (JAN)
ganhar ritmo,69,4
ganhar a licitação,70,4 (COE)
ganhar agilidade,71,4
ganhar vida,72,4
ganhar popularidade,73,4
ganhar liberdade,74,3
ganhar a viagem,75,3
ganhar a taça,76,3
ganhar vários títulos,77,3
ganhar volume,78,3
ganhar todas=as partidas,79,3
ganhar produtividade,80,3
ganhar a liberdade,81,3
ganhar salário,82,3

ganhar nome,83,3
ganhar a mesma coisa,84,3 (COE)
ganhar qualidade,85,3
ganhar espaços,86,3
ganhar corridas,87,3
(ganhar motor),88,3 (JAN)
ganhar salários,89,3
ganhar troféus,90,3
ganhar roupas,91,3
ganhar respeito,92,3
ganhar a parada,93,3
ganhar a ação,94,3 (COE)
ganhar um prêmio,95,3
ganhar pouco dinheiro,96,3
ganhar a prova,97,3
ganhar alguns quilos,98,3
ganhar copa,99,3 (MAN)
(ganhar seu nome),100,3 (JAN)
ganhar o nobel,101,3
ganhar prêmio,102,3
ganhar animação,103,3
ganhar forma,104,3
(ganhar um número),105,2 (JAN)
ganhar muita grana,106,2
ganhar mais tempo,107,2
(ganhar outra caravana),108,2 (JAN)
ganhar uma partida,109,2
ganhar a liderança,110,2

Teste 10: VER (DET) N metatexto do córpus

ver quadro,1,150 (MET)
ver texto,2,65 (MET)
ver o filme,3,39
ver o jogo,4,30
ver tv,5,29
ver computador,6,25
ver televisão,7,15
ver o crime,8,12
ver tabela,9,11 (MET)
ver a cena,10,11
ver a copa,11,11
ver as coisas,12,10
ver o mundo,13,10
ver o acidente,14,10
ver mapa,15,9 (MET)
ver navios,16,8 (JAN)
ver foto,17,7 (MET)
ver correio,18,7
ver modem,19,6
ver os policiais,20,6
ver o show,21,6
ver um filme,22,6
ver o espetáculo,23,6
ver as fotos,24,6 (MET)
ver fotos,25,5 (MET)
ver a moda,26,5
ver o quadro,27,5 (MET)
ver o lance,28,5
ver engenharia,29,5
ver a lista,30,5
ver desenho,31,5 (MET)
ver o carro,32,5
ver esse filme,33,5
ver coisas,34,5
ver o sol,35,4
ver um jogo,36,4
ver o rosto,37,4
ver a jardinagem,38,4
ver ilustração,39,4 (MET)
ver novelas,40,4
ver o filho,41,4
ver parque,42,4
ver estrelas,43,4
ver a vida,44,4
ver textos,45,4 (MET)
ver a cidade,46,4
ver o fenômeno,47,4
ver problemas,48,4
ver o resultado,49,4 (MET)
ver crediário,50,4
ver gente,51,4
ver o país,52,4
ver os ladrões,53,4
ver essas coisas,54,4
ver filmes,55,4
ver os corpos,56,4
ver o marido,57,4
ver reportagem,58,4 (MET)
ver futebol,59,3
ver a luz,60,3
ver as imagens,61,3
ver o cinema,62,3
ver o time,63,3
ver a coisa,64,3
ver o discurso,65,3
ver o trailer,66,3
ver o papa,67,3
ver sangue,68,3
ver o roubo,69,3
ver arte,70,3
ver alguma coisa,71,3
ver a página,72,3 (MET)
ver o gol,73,3
ver uma coisa,74,3
ver livro,75,3
ver endereços,76,3
ver os preços,77,3
ver os problemas,78,3
ver as pessoas,79,3
ver uma cena,80,3
ver a seleção,81,3

ver o coração,82,3
ver seus filmes,83,3
ver preços,84,3 (MET)
ver a tv,85,3
ver os dados,86,3 (MET)
ver a uva,87,3 (JAN)
ver o evento,88,3
ver importação,89,3
ver pessoas,90,3
ver o problema,91,3
ver relação,92,3 (MET) (JAN)
ver tanta gente,93,3
ver um homem,94,3
ver a noite,95,3
ver dificuldades,96,3
ver a mulher,97,3
ver a televisão,98,2
ver o contrato,99,2
ver pó,100,2

Teste 11: CRIAR (DET) N

criar a urv,1,173 (DAT)
criar o real,2,64
criar empregos,3,50
criar problemas,4,18
criar jogadas,5,9
criar polêmica,6,8
(criar condições),7,8 (JAN)
criar mais empregos,8,7
criar um clima,9,7
criar chances,10,6
criar um estilo,11,6
criar distorções,12,5
criar o indexador,13,5
(criar um fato),14,5(JAN)
criar jurisprudência,15,5
criar produtos,16,5
criar várias chances,17,4
criar o plano,18,4 (COE)
(criar ambientes),19,4 (JAN)
(criar este ano),20,4 (JAN)
criar cenários,21,4
criar meus filhos,22,4
criar histórias,23,4
criar obstáculos,24,4
criar gado,25,4
criar dependência,26,4
criar coragem,27,4
criar uma linguagem,28,4
criar inflação,29,4
criar personagens,30,4
criar asas,31,4
criar situações,32,4
criar o fundo,33,3 (COE)
criar uma expectativa,34,3
criar o psdb,35,3
criar confusões,36,3
criar a peça,37,3 (COE)
criar fama,38,3
criar moeda,39,3 (MAN)
criar desenhos,40,3
criar a moeda,41,3 (COE)
criar regras,42,3
criar outro problema,43,3 (COE)
criar os filhos,44,3
criar atritos,45,3
(criar um mundo),46,3 (JAN)
(criar mensagens),47,3 (JAN)
criar leis,48,3
criar espaços,49,3
(criar uma relação),50,3 (JAN)
(criar santos),51,3 (JAN)
criar expectativas,52,3
(criar obras),53,3 (JAN)
(criar tensões),54,3 (JAN)
criar postos=de=trabalho,55,3
criar um personagem,56,3
criar alternativas,57,3
criar ilusões,58,3
(criar um espaço),59,3 (JAN)
criar o serviço,60,2 (COE)
criar movimentos,61,2
criar desemprego,62,2
criar oportunidades,63,2
criar sindicatos,64,2
criar as jogadas,65,2
criar raízes,66,2
(criar a mulher),67,2(JAN)
criar um nosso estilo,68,2 (COE)
criar uma imagem,69,2
criar o universo,70,2
criar outros fundos,71,2
criar um gênero,72,2
criar drinques,73,2
criar outras chances,74,2
criar imagens,75,2
criar ovelhas,76,2
criar impostos,77,2
criar um boletim,78,2
(criar um sujeito),79,2 (JAN)
criar uma jurisprudência,80,2
(criar valores),81,2 (JAN)
criar a x-girl,82,2
criar o futuro,83,2
criar os anúncios,84,2 (COE)
criar dificuldades,85,2
(criar receita),86,2 (JAN)
criar punições,87,2

criar um ambiente,93,2
criar a coreografia,101,2 (COE)
criar a cena,102,2 (COE)
criar roupas,103,2
criar pânico,104,2
criar tais juizados,105,2 (COE)
criar confusão,106,2
(criar uma sociedade),107,2 (JAN)
criar o clima,108,2 (COE)
criar um desequilíbrio,109,2
(criar colônias),110,2 (JAN)
criar uma cultura,111,2
criar despesas,88,2
criar indisposições,89,2
criar pôsteres,90,2
criar galinhas,91,2
criar cargos,92,2
criar um cenário,94,2
criar museus,95,2
(criar mecanismos),96,2 (JAN)
criar tumulto,97,2
criar direitos,98,2
criar ruído,99,2
criar algumas jogadas,100,2
criar uma forma,112,2
criar palavras,113,2
criar uma apresentação,114,2
(criar alguma peça),115,2 (JAN)
criar um debate,116,2
criar o problema,117,2 (COE)
criar tributos,118,2
criar animais,119,2
criar os modelos,120,2 (COE)
criar ansiedade,121,2

4
Composicionalidade com base em córpus

> *Queremos estabelecer uma ordem no nosso conhecimento da linguagem: uma ordem para uma finalidade determinada; uma ordem dentre as muitas possíveis, não a ordem.*
>
> Wittgenstein, *Investigações Filosóficas*

Neste capítulo iremos focar a última etapa de implementação computacional do nosso estudo para aplicação de uma medida de composicionalidade semântica em relação às CMs detectadas no capítulo 3. Trata-se de uma medida de similaridade entre os microcontextos (parágrafos do córpus) em que as CMs ocorrem (cf. Garrão, Oliveira, Freitas & Dias, 2006). É importante adiantar que nosso critério de aferição do grau de transparência semântica se baseia em uma técnica utilizada no domínio computacional de Recuperação de Informação (RI), em detrimento de uma aferição intuitiva, com base nos testes já criticados na seção 2.3.1.

Na verdade, muito se especula sobre a importância da aplicação de teorias semânticas já existentes na lingüística, como aquelas apresentadas e questionadas no capítulo 2, para fins de PLN. Contrariamente, pretendemos demonstrar neste capítulo a importância de PLN e do córpus para avaliar semanticamente as CMs em questão. Através de uma medida de similaridade entre os microcontextos em que uma dada CM aparece e os microcontextos em que detectamos apenas o SN que compõe a CM, pretendemos avaliar empiricamente o que é dito ser "transparência e opacidade semântica".

Nossa proposta é aferir o grau de transparência/opacidade semântica de uma CM pelos contrastes entre os contextos de uso da CM propriamente dita (como por exemplo, *fazer campanha*) e os contextos de uso do SN que compõe a CM (*campanha*) em detrimento de uma avaliação semântica apriorística dos itens que compõem a CM. Em outras palavras, é o que está fora da CM que vai determinar o seu grau de composicionalidade semântica, como demonstraremos na seção 4.2.

Conforme apresentados na seção 2.3.1, os critérios tradicionalmente utilizados para caracterizar um segmento lingüístico como uma CM são:

i) não-composicionalidade – o significado do todo não corresponde à soma das partes, como o exemplo já criticado no capítulo 2, *bater as botas*. (Guenthner e Blanco, 2004; Neves, 1999, entre outros). Um dos problemas desta definição é que outros autores argumentam também que é possível caracterizar algumas CMs pela possibilidade de seus componentes contribuírem para a semântica do composto (formando uma *colocação*) em oposição ao conceito de *expressões idiomáticas*, como por exemplo, a distinção entre a colocação *pagar as contas* e a expressão idiomática *pagar mico* (Cruse, 1986)**.** Segundo o autor, a composicionalidade seria o melhor critério para distinguir uma colocação de uma expressão idiomática, embora admita que a distinção seja difícil em muitos casos.

ii) não-substituição ou arbitrariedade - não é possível substituir palavras que compõem uma CM mantendo a integridade da expressão, ainda que a palavra substituída seja sinônima da original (Tagnin, 1999; Mannning e Schütze, 1999). Este critério pode ser contra-argumentado até pela expressão exaustivamente utilizada para definir a noção de opacidade semântica: *bater as botas – bater a caçuleta*. Tal critério também se fia em uma questão problemática na Semântica: como podemos determinar que uma palavra é sinônima da outra? Pode-se dizer, por exemplo, que, em determinados contextos, *dar uma festa* e *fazer uma festa* são CMs sinônimas; por outro lado, será que uma visão semântica tradicional consideraria *dar* e *fazer* verbos sinônimos?

iii) não-modificação - muitas CMs não podem ser livremente modificadas pela adição de informação lexical ou de transformações gramaticais (Guenthner e Blanco, 2004). Contrariando este último critério, Cruse (1986) afirma que colocações variam quanto ao número de palavras envolvidas, quanto às relações sintáticas entre as palavras, e quanto ao grau de rigidez com que os itens são combinados ("*tomar uma decisão*"/ "*uma decisão foi tomada*"/ "a *tomada de decisões*"). Este terceiro critério fica ainda mais complicado quando a CM não é nominal, como *alto falante* ou *criado mudo,* mas verbal, com várias possibilidades aspectuais: como *receber tratamento, receber o mesmo tratamento; tomar decisão; tomar várias decisões.* Pudemos verificar em Garrão (2001) que até mesmo as CMs consideradas altamente opacas como *bater perna, fazer questão, pagar mico* podem sofrer modificação através de inserção de um marcador aspectual: *bater muita perna, fazer muita questão, pagar o maior mico*.

Tais opiniões contraditórias sobre a detecção de CMs — baseadas naquilo que rotulamos no capítulo 2 como semântica do cálculo e naquilo que Fillmore (1979) rotulou como semântica da inocência —, nos impulsionou a procurar uma via alternativa para caracterizar o grau de composicionalidade de uma CM. Segundo Aranha, Freitas, Dias & Passos (2004), "palavras com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes".

Tomando como pressuposto essa idéia, consideramos possível fazer avaliações sobre o grau de similaridade entre microcontextos contendo uma CM e microcontextos contendo somente o SN que compõe a CM. A hipótese, portanto, é a de que o grau de transparência semântica da CM é proporcional ao aumento do grau de similaridade entre os parágrafos contendo uma certa CM e os parágrafos contendo somente o SN presente na CM. Em outras palavras, se os parágrafos do córpus que contêm todas as CMs *fazer campanha* forem similares aos parágrafos do córpus que contêm o SN *campanha*, a probabilidade de se encontrar *fazer campanha* nos mesmos microcontextos de *campanha* é muito grande. Isso indicaria que a CM é transparente.

Além de eliminar o risco da aferição intuitiva de composicionalidade semântica das CMs, este recurso se revelou também vocacionado a detectar o grau de polissemia dos SNs quando não pertencem à CM, como também do grau de polissemia das CMs propriamente ditas, como veremos na seção 4.2.

4.1
Passo-a-passo do método

Um dos métodos mais utilizados para medir o grau de similaridade entre documentos é baseado em um *Modelo de Espaço Vetorial* (Baeza-Yates & Ribeiro-Neto, 1999). Esse modelo representa os documentos através de todas as palavras neles contidas. Por meio dele, pode-se estabelecer o grau de similaridade entre os documentos. Cada documento, ou parágrafo (que chamamos aqui de microcontexto) é dividido em uma tabela de freqüência de palavras. As tabelas são chamadas de vetores. Um vocabulário é construído a partir de todos os

microcontextos. Cada microcontexto é representado como um vetor em relação ao vocabulário total dos microcontextos.

Exemplo simplificado:

Documento A:

Um cachorro e um gato.

um	cachorro	e	gato
2	1	1	1

Documento B

Um sapo

um	sapo
1	1

O vocabulário é a soma de todas as palavras utilizadas, isto é, de todos os documentos:

um, cachorro, e, gato, sapo

Portanto:

Documento A:

Um cachorro e um gato.

um	cachorro	e	gato	sapo
2	1	1	1	0

Vetor: (2,1,1,1,0)

Documento B

Um sapo

um	cachorro	e	gato	sapo
1	0	0	0	1

Vetor: (1, 0, 0, 0, 1)

Este exemplo bastante simplificado demonstra como podemos estabelecer entre documentos medidas de similaridades, sendo que, numa aferição real, as

palavras funcionais (*um, e*, no exemplo acima) são descartadas (chamadas de "*stop words"*). Essas medidas são altamente relevantes para aplicações computacionais que lidam com a Teoria da Informação (TI), como sistemas de Recuperação de Informação (como, por exemplo, o *Google*™). Quanto mais palavras os documentos tiverem em comum, mais similares serão entre si, maior a relação entre eles. É essa técnica que facilita a pesquisa do usuário em um sistema de busca; por meio dela, é possível estabelecer o grau de semelhança entre inúmeros documentos disponíveis na rede.

Neste estudo, portanto, nos apropriamos desta mesma idéia de similaridade entre documentos para detectar a relação entre os microcontextos que contenham uma certa CM e os microcontextos que contenham somente o SN fora da CM. Tomemos como exemplo a aferição do grau de similaridade entre os microcontextos de *fazer campanha* e *campanha*: primeiramente, aferimos o grau de similaridade entre todos os parágrafos do córpus CETENFolha que contenham as CMs *fazer campanha* e calculamos a sua similaridade média. Posteriormente, aferimos o grau de similaridade entre todos os parágrafos do córpus CETENFolha que contenham os SNs *campanha* (sem estarem precedidos pelo verbo "fazer" ou por qualquer outro verbo) e calculamos a similaridade média entre eles. Por fim, calculamos a similaridade média entre os parágrafos que contêm a CM e aqueles que contêm o SN.

Quanto maiores forem os números expostos nos vetores de cada microcontexto em relação ao vocabulário total dos microcontextos, maior a relação entre os microcontextos das CMs e dos SNs e, conseqüentemente, maior o grau de composicionalidade semântica da CM. Quanto menores forem esses números, menor a relação entre os microcontextos, e maior o grau de opacidade semântica da CM.

4.2
Aferição do grau de composicionalidade das CMs

Depois da identificação das CMs mais freqüentes do córpus, nós observamos os contrastes entre os microcontextos em que aparecem, através do

Modelo de Espaço Vetorial. As duas expressões cujos graus de similaridade pretendemos aferir (como, por exemplo, *fazer campanha* e *campanha*) são representadas como vetores em um espaço multidimensional. O cosseno entre esses dois vetores indica as palavras que eles têm em comum e, por essa razão, o método pode ser considerado como uma medida de similaridade entre dados.

Para cada CM *w* nós realizamos as seguintes etapas:

i) extração de todos os parágrafos contendo *w* (conjunto *P*1; por exemplo *fazer campanha);*

ii) extração de todos os parágrafos contendo o substantivo em *w* que não ocorre em *P*1 (conjunto *P*2; *campanha);*

iii) indexação de *P*1 e *P*2 no Modelo de Espaço Vetorial;

iv) cálculo das matrizes de similaridades entre parágrafos em *P*1 e obtenção da média dos seus valores;

v) cálculo das matrizes de similaridades entre parágrafos em *P*2 e obtenção da média de seus valores;

vi) cálculo das matrizes de similaridades entre os parágrafos em *P*1 e *P*2 e obtenção da média de seus valores.

Portanto, o aumento das similaridades entre os parágrafos em P1 e P2 é proporcional ao aumento do grau de composicionalidade da CM em questão. Para avaliar tal hipótese, é calculada a similaridade intra-*P*1 (entre todos os microcontextos que contêm *fazer campanha*) , e em seguida intra-*P*2 (entre todos os microcontextos que contêm *campanha*). Finalmente, é avaliada a similaridade entre *P*1 e *P*2. É esta última etapa que nos dará o grau de composicionalidade semântica da CM.

É importante mencionar que escolhemos pelo menos 30 ocorrências tanto de P1 quanto de P2 para a medida de avaliação proposta. As CMs cujo grau de composicionalidade pretendemos aferir foram extraídas das 10 listas apresentadas no capítulo 3 e ranqueadas em ordem crescente de transparência semântica. As Tabelas abaixo (Tabelas de 4 a 13) estão organizadas da seguinte forma:

– cada uma delas contém CMs retiradas de cada uma das 10 listas de CMs apresentadas no capítulo 3: *fazer+SN, ter+SN, dar+SN, perder+SN, usar+SN, receber+SN, deixar+SN , tomar+SN, ganhar+SN, criar+SN.*

– a coluna da extrema esquerda contém a lista de SNs na estrutura V+SN;

– *SM1* é a similaridade média intra-*P*1, ou seja, entre as CMs;

– *SM2* é a similaridade média intra-*P*2; ou seja, entre os SNs;

– *SM3* é a similaridade média entre *P*1 e *P*2*; Var* são as variâncias correspondentes.

Portanto, quanto maior for o valor de *SM3*, mais similares são os parágrafos contendo a CM e os parágrafos contendo o SN fora da CM. Conseqüentemente, mais composicional ou transparente será a CM. Além disso, quanto maior o grau de similaridade entre as CMs (SM1), menos polissêmica é a CM. Quanto maior o grau de similaridade entre os SNs (SM2), menos polissêmico é o SN.

É importante ressaltar, nesta etapa do estudo, que para os casos em que as ocorrências de CMs não eram suficientes para uma aplicação confiável do método (menos de 30 ocorrências), nós adicionamos um córpus através da ferramenta *Google*™ através de buscas contendo as CMs de mesmo padrão tanto na forma canônica quanto na forma flexionada[20].

FAZER	SM1	Var	SM2	Var	**SM3**	Var
falta	1,85	0,01	5,59	0,10	**0,26**	0,0003
a festa	0,77	0,004	1,95	0,004	**0,33**	0,0003
sentido	5,33	0,07	3,8	0,04	**0,33**	0,0005
água	3,13	0,04	7,81	0,02	**0,35**	0,0005
dinheiro	0,64	0,001	2,98	0,007	**0,36**	0,0005
amigos	0,62	0,0008	3,5	0,02	**0,36**	0,0007
compras	1,44	0,013	2,03	0,007	**0,44**	0,002
parte	0,46	0,0004	2,72	0,09	**0,5**	0,003
sucesso	0,84	0,006	2,52	0,04	**0,5**	0,0007
campanha	0,60	0,003	2,45	0,016	**0,54**	0,0004

Tabela 4: resultados com verbo *fazer*

[20] As buscas eram feitas através de escolhas aleatórias das flexões dos verbos.

TER	SM1	Var	SM2	Var	**SM3**	Var
fôlego	1,066	0,01	2,17	0,005	**0,34**	0,0005
acesso	0,86	0,007	3,04	0,01	**0,36**	0,0002
uma idéia	0,54	0,008	1,94	0,009	**0,36**	0,0004
razão	1,00	0,007	5,29	0,12	**0,42**	0,001
sucesso	1,47	0,03	4,11	0,04	**0,43**	0,0007
força	5,98	0,13	4,06	0,006	**0,44**	0,0004
problema	0,88	0,002	2,51	0,02	**0,72**	0,001
medo	1,28	0,017	5,42	0,07	**0,84**	0,006

Tabela 5: resultados com o verbo *ter*

DAR	SM1	Var	SM2	Var	**SM3**	Var
bandeira	1,53	0,0290	1,69	0,002	**0,19**	0,0001
frutos	2,54	0,0570	3,13	0,010	**0,28**	0,0003
tempo	2,16	0,0600	6,15	0,270	**0,28**	0,0004
sorte	0,80	0,0010	4,40	0,060	**0,62**	0,0030
entrevistas	0,86	0,0030	12,10	0,400	**0,66**	0,0006
resultado	0,87	0,0070	4,33	0,120	**0,69**	0,0005
lucro	5,20	0,1500	5,07	0,100	**0,94**	0,0010
declarações	1,01	0,0009	18,67	0,880	**1,07**	0,0020

Tabela 6: resultados com o verbo *dar*

PERDER	SM1	Var	SM2	Var	**SM3**	Var
a cabeça	0,79	0,004	1,24	0,005	**0,18**	0,0002
o bonde	1,12	0,02	1,91	0,003	**0,37**	0,0005
peso	3,08	0,02	5,18	0,01	**0,48**	0,0008
dinheiro	0,57	0,0007	7,72	0,12	**0,57**	0,001
tempo	0,39	0,0004	5,31	0,05	**0,58**	0,002
a eleição	1,39	0,002	5,67	0,07	**0,74**	0,0009
o emprego	1,25	0,01	3,15	0,01	**1,04**	0,02

Tabela 7: resultados com o verbo *perder*

USAR	SM1	Var	SM2	Var	**SM3**	Var
a cabeça	0,91	0,005	2,35	0,005	**0,259**	0,0003
a força	7,63	0,18	2,26	0,005	**0,37**	0,0006
o cinto	19,71	2,09	5,59	0,08	**0,55**	0,0008
computador	1,14	0,002	4,52	0,013	**0,63**	0,0003
drogas	2,35	0,007	3,98	0,015	**0,65**	0,0007
camisinha	2,09	0,03	6,12	0,17	**0,76**	0,001

Tabela 8: resultados com o verbo *usar*

RECEBER	SM1	Var	SM2	Var	**SM3**	Var
alta	2,43	0,008	17,0	0,3	**0,41**	0,0003
visita	1,15	0,0006	10,8	0,032	**0,6**	0,0004
bola	1,94	0,05	8,07	0,16	**0,70**	0,0006
dinheiro	0,69	0,0007	4,93	0,067	**0,75**	0,005
benefício	2,15	0,008	4,04	0,1	**1,13**	0,001
propina	1,91	0,004	12,40	0,24	**1,45**	0,001

Tabela 9: resultados com o verbo *receber*

DEIXAR	SM1	Var	SM2	Var	**SM3**	Var
marcas	1,129	0,005	2,82	0,008	**0,31**	0,0003
o país	0,67	0,0007	2,49	0,02	**0,5**	0,003
o cargo	0,68	0,003	2,32	0,02	**0,51**	0,002
o governo	0,65	0,0005	3,67	0,05	**0,55**	0,0005
o local	0,64	0,05	3,67	0,54	**0,55**	0,0005
a cidade	0,67	0,0004	4,02	0,03	**0,60**	0,0004
os filhos	1,18	0,01	6,04	0,04	**0,65**	0,001
vestígio	8,33	0,33	10,61	0,2	**0,7**	0,0004

Tabela 10: resultados com o verbo *deixar*

TOMAR	SM1	Var	SM2	Var	**SM3**	Var
partido	0,96	0,001	3,18	0,02	**0,3**	0,0003
iniciativa	1,76	0,004	12,9	0,60	**0,31**	0,0002
conhecimento	1,0	0,007	2,13	0,006	**0,33**	0,006
café	1,8	0,02	27,8	3,53	**0,4**	0,0005
o poder	0,80	0,007	2,41	0,003	**0,47**	0,007
posse	1,14	0,009	3,41	0,04	**0,52**	0,0004
uma decisão	0,45	0,0005	1,77	0,012	**0,54**	0,002
banho	1,32	0,01	3,0	0,03	**0,8**	0,01
providência	1,11	0,01	3,15	0,03	**0,94**	0,01
cuidado	0,9	0,01	6,45	0,08	**0,99**	0,01

Tabela 11: resultados com o verbo *tomar*

GANHAR	SM1	Var	SM2	Var	**SM3**	Var
espaço	4,64	0,12	2,51	0,01	**0,25**	0,0002
terreno	1,13	0,006	2,29	0,07	**0,28**	0,0002
força	7,76	0,18	3,48	0,01	**0,31**	0,0003
dinheiro	0,5	0,001	3,28	0,06	**0,31**	0,001
tempo	0,54	0,001	2,56	0,02	**0,48**	0,0009
a vida	0,71	0,002	8,59	0,15	**0,62**	0,001
o jogo	1,05	0,001	6,23	0,08	**0,67**	0,001
a eleição	1,05	0,05	6,3	0,13	**0,69**	0,0006

Tabela 12: resultados com o verbo *ganhar*

CRIAR	SM1	Var	SM2	Var	**SM3**	Var
raízes	2,43	0,008	17,0	0,3	**0,41**	0,0003
atritos	1,15	0,0006	10,8	0,032	**0,6**	0,0004
polêmica	1,94	0,05	8,07	0,16	**0,70**	0,0006
obstáculos	0,69	0,0007	4,93	0,067	**0,75**	0,005
ameaças	2,15	0,008	4,04	0,1	**1,13**	0,001
os filhos	1,91	0,004	12,40	0,24	**1,45**	0,001

Tabela 13: resultados com o verbo *criar*

4.3
Avaliação dos resultados

A Tabela 14 abaixo apresenta os resultados qualitativos dos testes aplicados com cada verbo. Na coluna da extrema esquerda figuram as CMs consideradas como as mais composicionais ou transparentes em relação às outras encabeçadas pelo mesmo verbo. Isto é, aquelas que mais apareceram em microcontextos similares aos microcontextos dos SNs que as compõem. A coluna da extrema direita apresenta as CMs menos composicionais, ou mais semanticamente opacas. Ou seja, aquelas que menos ocorreram em microcontextos similares aos microcontextos dos SNs que as compõem. Já na coluna do meio figuram os casos não-extremos ou os meio-termos.

Lema	***+ composicional***	***meio-termo***	***- composicional***
FAZER	fazer sucesso fazer campanha	fazer compras	fazer falta fazer sentido
TER	ter medo ter problema	ter força	ter fôlego ter uma idéia
DAR	dar lucro dar declarações	dar sorte	dar bandeira dar frutos
PERDER	perder emprego perder a eleição	perder tempo	perder a cabeça perder o bonde
USAR	usar camisinha usar drogas	usar o cinto	usar a cabeça usar a força
RECEBER	receber propina receber benefício	receber visita	receber alta
DEIXAR	deixar vestígio deixar filhos	deixar o país	deixar marcas
TOMAR	tomar cuidado tomar providência	tomar decisão	tomar partido tomar iniciativa
GANHAR	ganhar a eleição ganhar o jogo	ganhar tempo	ganhar espaço ganhar terreno
CRIAR	criar os filhos	criar obstáculos	criar raízes

Tabela 14: resumo qualitativo dos resultados

Os resultados parecem ser, em alguma medida, consistentes com nossas inevitáveis previsões, fornecendo base empírica a nossas parcas intuições sobre os padrões de composicionalidade de CMs do PB. De fato, nós previámos que *tomar partido* e *partido* assim como *dar bandeira* e *bandeira* iriam figurar em microcontextos pouco relacionados entre si; ou seja, microcontextos com um baixo índice de palavras idênticas. Seguem alguns poucos exemplos retirados de *P1* e *P2*, respectivamente, que demonstram esta opacidade semântica entre os microcontextos acima:

Fragmento de P1 *(tomar partido)*
"*Se queremos agir, temos de* ***tomar partido****, «sujar as mãos», e não só no sangue, que é nobre, mas também «na merda» nas alianças sujas, na mentira"* .
" Era uma coisa complicada, que minha mãe não aceitava de vez em quando explodia, a gente via as consequências, tinha que ***tomar partido****, isso ao longo de anos."*
"Em vez de ***tomar partido****, a mente independente de Einstein preferiu encarar os paradoxos entre as duas correntes e tentar unificá-las."*.

Fragmento de P2 *(partido*)
*"O **partido** vai presidir as Comissões de Trabalho, Administração e Serviço Público, a de Seguridade Social e Família e, por fim, a Comissão de Defesa Nacional".*
*"«Quero ver o Vernon Reid dar um» break» no rock e tocar maracatu e **partido** alto aí no Brasil», disse ontem Naná Vasconcelos, por telefone, de Nova York."*
*"Depois de ter se assumido como bom **partido**, solteiro, em busca de namorada e posado até de cuecas na revista «Caras», o deputado Robson Tuma mais conhecido como Tuminha não tem do que reclamar ."*

Quadro 1: microcontextos de *partido / tomar partido*

Fragmento de P1 (*dar bandeira*)
*"Tem gente que anda com vidrinhos de álcool no carro ou, como a mulher está viajando, leva outra camisa e troca para não **dar bandeira**. "*
*"Dar pala: sinônimo de **dar bandeira** ."*
*"Entrou na Faap, fez cinema, publicidade, teatro, **deu bandeira** ."*
*"Progresso -- Não falei, Ordem, pra você não ficar **dando bandeira** ?"*

Fragmento de P2 *(bandeira)*
*"O verde e amarelo do Brasil ganham o vermelho da **bandeira** do Olodum, que apresentará seu filhote Dança Olodum! , um desdobramento do Bando de Teatro do grupo baiano" .*
*"A **bandeira** 2 está autorizada a partir das 6h de hoje.".*
*"Ainda em maio, o Fashion Mall lança um cartão de crédito internacional de afinidade com a **bandeira** Visa."*
*" Irracionalmente, queimaram a **bandeira** do patrocinador, como se este tipo de pressão trouxesse resultados*".

Quadro 2: microcontextos de *bandeira / dar bandeira*

Por outro lado, também suspeitávamos que *tomar banho* e *banho* assim como *usar camisinha* e *camisinha* ocorreriam em parágrafos similares; isto é, que compartilhassem um grau elevado de palavras iguais. Sobre o primeiro par, estávamos parcialmente certas, uma vez que a CM também é utilizada como uma forma de insulto:

Fragmento de P1 *(tomar banho)*
"Segundo os pesquisadores, os pacientes podem ter sido infectados com microorganismos ao ***tomar banho*** *ou beber água ."*
"Quanto à expressão «ele que vá ***tomar banho****», embora não registrada em fita, foi pronunciada diante de testemunhas que também a interpretaram como sendo dirigida ao ministro Ricupero, e não à reportagem da Folha."*
"Elas podem dormir, ***tomar banho*** *e se alimentar entre 20h30 e 8h. Mas não temos estrutura para atender durante o dia», diz Maria Cecília ."*
"Dentro dos grupos eram combinados esquemas de revezamento, para todos poderem ir para casa almoçar e ***tomar banho****, por exemplo."*

Fragmento de P2 *(banho)*
"O corpo de Betty, na cena do ***banho,*** *está um luxo, e a produção conseguiu um charme a mais: uma paisagem urbana que pode ser de qualquer cidade do planeta".*
" Sua rotina inclui nutrição, ***banho,*** *medicação e fisioterapia."*
" Enquanto Hargreaves e sua irmã Ruth, assessora especial do presidente, vestiam roupas de ***banho,*** *Itamar trajava calça escura e camisa de manga comprida ."*

Quadro 3: alguns microcontextos de *banho/ tomar banho*

Já o segundo par parece estar intimamente relacionado; ou seja, a CM *usar camisinha* aparece em microcontextos que compartilham um grau elevado de palavras iguais. Portanto, dentro da nossa proposta, essa CM parece ser bastante composicional. Além disso, o SN *camisinha* parece apresentar um grau baixíssimo de polissemia, uma vez que SM2, ou seja, a similaridade média entre todos os parágrafos que contêm o SN é altíssimo (6,12).

Fragmento de P1 *(usar camisinha)*
"É preciso dizer ao adolescente que tem de ***usar camisinha****"*
"Se estão com Aids, eles informam o freguês e ou parceiro e pedem para ***usar camisinha..***"
"Você não ia ser louca de ter vida sexual sem ***usar camisinha****, não é verdade ?"*

Fragmento de P2 *(camisinha)*
"Bispos aceitam ***camisinha*** *no combate a Aids* ***."***
"Sem ***camisinha*** *não dá, afirma o analista de sistemas Jorge Luis, 24, que costuma sair com esses adolescentes* ***."***
"*A* ***camisinha*** *foi apontada como o melhor método anticoncepcional ."*

Quadro 4: alguns microcontextos de *usar / usar camisinha*

Por outro lado, suspeitávamos que o par "tomar café" e "café" seguiria o mesmo padrão composicional de "usar camisinha", o que não ocorreu, uma vez que o SN "café" em P2 ocorrem em ambientes lingüísticos não relacionados a P1, tais como economia, agricultura, arquitetura (como uma cor).

Fragmento de P1 *(tomar café)*
*"Ele parou em dois bares, para **tomar café** e água gelada."*
*"Não há mais o risco de sair para **tomar café** e descobrir, de última hora, que a máquina está quebrada ".*
*"Os dois deverão **tomar café** juntos a partir das 8h no Othon Palace Hotel, em frente à praia de Copacabana (zona sul)" .*

Fragmento de P2 *(café)*
*"Chegaram a brincar com os rivais - em sua maioria tensos, apesar do encontro ter acontecido na porta de um **café**, o Cabalas" .*
*Junte-se a todo esse clima o cardápio que inclui frutas tropicais, **café**, caldo de feijão e caipirinha .*
*«Nosso objetivo é impedir a saída clandestina do **café** que nós produzimos e desmantelar as quadrilhas de sonegadores .*
*"De manhã, depois da toilette e do **café**, sentava-se no divã da sala principal e lia os jornais ".*
*"Um conjunto de **café** de US$ 3.000 não é só para coleção ?"*
*"(...) Soldados marcham decididos no rumo da estação carregando acima dos bonés, em meio ao espinhal de canos com as bocas voltadas para o sol, (…) um pedaço de fumo de rolo no bolso inferior e coroando tudo o imenso chapéu cor de **café**, (…).*

Quadro 5: alguns microcontextos de *café/ tomar café*

Em suma, esses resultados nos deixaram extremamente confiantes com a aplicabilidade do método. Portanto, ao invés de atribuir à intuição do pesquisador o poder de aferição do grau de composicionalidade de uma CM, nós preferimos confiar naquilo que o córpus nos revela.

De fato, reconhecemos que os resultados aqui obtidos deveriam ser corroborados por um córpus ainda mais robusto do PB, pois alguém poderia refutar nossas conclusões caracterizando o córpus como tendencioso. Por ora, nós preferimos pensar, ao contrário, que nossa intuição é que tende a ser traiçoeira.

5
Discussão, aplicação e trabalhos futuros

Este estudo foi desenvolvido, essencialmente, em razão de mais de sete anos de investigação do comportamento de CMs verbais. Os recortes teóricos anteriormente escolhidos (Garrão, 2001; Garrão & Dias, 2001/2; Basílio, Oliveira & Garrão, 2003) serviram de base para os questionamentos aqui feitos. Dentre esses podemos destacar:

i) que o critério dedutivo se revelou pouco produtivo para dar conta de CMs verbais, através de considerações controversas de aceitabilidade. Este critério se mostra particularmente improdutivo para a caracterização de CMs mais freqüentes da língua, cuja detecção depende diretamente de uma abordagem empírica;

ii) que uma visão de composicionalidade forte ou representacionista (cf. Neves, 1999; Tagnin, 1999 e Vale, 2002) para a identificação de CMs tem sérias implicações, uma vez que propõe como medida de avaliação do fenômeno, a semântica do cálculo ou um ideal de falante baseado na inocência (cf. Fillmore, 1979), embora os "calculadores" e os falantes reais não o sejam;

iii) que uma visão de composicionalidade fraca ou neo-representacionista (cf. Lakoff, 1993; Gibbs, 1994) para lidar com as CMs também não está livre de problemas, uma vez que parte de uma visão de significado muito inclusiva. Embora seja elucidativa em muitos aspectos semânticos, tal perspectiva, por não ter uma ambição explicativa, não determina formalmente os limites de cada faceta da construção do significado.

iv) que a alegada possibilidade de separação entre semântica e pragmática, ou entre conhecimento lingüístico e enciclopédico é improvável. Compartilhamos com Wittgenstein (1979) e Harris (1996) uma visão pragmática radical, em que o uso lingüístico não é um dos componentes da linguagem, mas a única forma produtiva de se pensar os fenômenos lingüísticos, assim como também

concordamos com Kilgarriff (2000) que os significados só existem dependentes de propostas ou tarefas.

Em relação aos dados obtidos podemos afirmar com alguma segurança que:

i) uma abordagem não-representacionista com base em córpus é objetiva e pragmática, uma vez que utiliza como fonte de seus dados, o discurso do falante desavisado, porém nada inocente. Discurso este sem pretensões descritivas nem comprometimentos teóricos.

ii) a língua pode ser descrita como um fenômeno probabilístico, uma vez que há nitidamente padrões de combinações vocabulares recorrentes. De certa forma, esta perspectiva atenua a visão chomskiana da linguagem, focada na semântica do cálculo, e prioriza uma visão de língua inseparável da pragmática; isto é, enfatiza o teor eventivo do fenômeno lingüístico.

iii) os chamados verbos leves estão, de fato, entre os mais produtivos no domínio da combinação multivocabular, como argumenta Vale (2002); por outro lado, nosso método foi capaz de identificar outros verbos que superam suas ocorrências para o padrão procurado. Por exemplo, “perder” (6°) “usar” (7°) “deixar” (9°), “ganhar” (11°), “criar” (13°), dentre outros, superam a freqüência de verbos tradicionalmente rotulados por leves ou suporte, como “levar” e “tirar”, por exemplo, que estão em 17° e 27°, respectivamente.

iv) o método de Logaritmo de Verossimilhança (Banerjee & Pedersen,2003) se mostrou bastante adequado para a tarefa proposta, com precisão de 87, 2%. Isto é, gerou apenas 12.8% de pseudo-CMs. Some-se a isso, a rapidez da obtenção dos resultados – o que minimiza o trabalho do pesquisador – e a confiabilidade dos tipos de CMs obtidas – o que descarta qualquer critério dedutivo e moroso em relação à aceitabilidade e possibilidade de ocorrências.

v) o determinante/quantificador comumente presente na CM de padrão procurado (formando estruturas do tipo *fazer <u>uma</u> declaração*, *ganhar <u>mais</u> tempo*) também foi adequadamente captado pelo método de Logaritmo de Verossimilhança.

vi) o Modelo de Espaço Vetorial (Baeza-Yates & Ribeiro-Neto, 1999) também se revela bastante promissor para aferição do grau de composicionalidade proposta. Além de prescindir da semântica do cálculo, ele também demonstra, de uma forma geral, uma vocação para detectar o grau de polissemia dos SNs que eventualmente podem figurar numa CM.

vii) o córpus, além de servir como base de dados para detecção de CMs, também tem um papel preditivo ao fornecer os ambientes lingüísticos tipicamente relacionados às CMs.

Sobre as implicações do córpus utilizado, podemos apontar:

i) o fato de o córpus ter sido compilado há mais de dez anos (em 1994). Se considerarmos a relação que Wittgenstein estabelece entre "a língua" e "uma cidade", podemos dizer que encontramos algumas poucas "casas derrubadas", ou seja, algumas CMs não mais amplamente utilizadas (como, por exemplo, *criar a URV*, *usar AZT*);

ii) o teor informativo do córpus escolhido, o que, por um lado, pode ser considerado relativamente desejável do ponto de vista do uso da língua (para ensinamento de segunda língua e aplicações de PLN). Mas por outro lado, pode ter enfatizado de modo exagerado domínios como a política, economia e esportes em detrimento de outros assuntos.

Sobre as possíveis aplicações lexicográficas dos resultados obtidos pelo método de Logaritmo de Verossimilhança, podemos apontar:

a) construção de um dicionário das CMs mais utilizadas no PB, com exemplos de usos retirados do próprio córpus CETENFolha;

b) construção de um dicionário bilíngüe dessas CMs e seus prováveis pares em outras línguas, como, por exemplo, o inglês (relevante para aprendizes de português como segunda língua).

Sobre as possíveis aplicações do Modelo de Espaço Vetorial aplicado nesta pesquisa, podemos apontar:

a) um critério de aferição de transparência/opacidade semântica livre de deduções e intuições do pesquisador.

b) um critério de aferição de polissemia e ambigüidade com base em córpus em detrimento de uma avaliação intuitiva.

Sobre as possíveis aplicações dos resultados obtidos pelo método de Logaritmo de Verossimilhança em PLN, podemos apontar:

a) a construção de um dicionário eletrônico das CMs mais utilizadas no PB, com exemplos de usos retirados do próprio córpus CETENFolha;

b) otimização da etiquetagem de córpus computadorizados, incluindo o próprio córpus CETENFolha. Isto é, se possível, indexar as CMs como uma co-ocorrência lingüística motivada.

c) incremento de softwares de Tradução Automática como o Delta Translator® e o Globalink Power Translator Pro® para evitar erros de traduções.

d) incremento de dicionários bilíngües disponíveis na Internet, como o Babylon™ e o Google™, que são amplamente utilizados para traduzir periódicos de uma língua para outra.

Sobre as possíveis aplicações do Modelo de Espaço Vetorial utilizado nesta pesquisa, para fins de PLN, podemos apontar:

a) exclusão de CMs opacas como possíveis candidatas à busca do usuário. Ou seja, CMs menos composicionais ou mais opacas não estão relacionadas diretamente com os SNs que as compõem (como *bandeira* em *dar bandeira*, *partido* em *tomar partido* e *frutos* em *dar frutos*). Desta forma, se o usuário estiver fazendo uma busca sobre “bandeira da Dinamarca” ou “partido comunista” e “frutos silvestres”, o sistema de busca poderia excluir os documentos que contivessem CMs como essas do total dos documentos relevantes ao usuário para aumentar a precisão da busca.

b) detecção de termos relevantes e irrelevantes (pseudo-termos) para Recuperação de Informação. Em outras palavras, SNs com um baixo índice de polissemia, como *camisinha*, por exemplo, podem ser consideradas como termos de busca. Outras, como *decisão* e *idéia*, podem ser consideradas pseudo-termos, uma vez que são encontradas em documentos com assuntos difusos e não-relacionados.

* * *

Por mais motivador e promissor que esse caminho não-representacionista tenha se revelado, temos consciência de que há alguns ajustes a serem feitos. Os resultados poderiam ser ainda mais confiáveis se nós dispuséssemos de um córpus anotado mais robusto de textos no PB. Poderíamos também tornar os resultados mais precisos se considerássemos não somente um parágrafo, mas um escopo lingüístico maior para o propósito da medida de composicionalidade. Isto aumentaria ainda mais a credibilidade da nossa metodologia.

Por outro lado, podemos dizer que este método não somente nos alforria daquele com base na semântica da inocência, como também se revela altamente profícuo para a lexicografia e para PLN, uma vez que fornece os ambientes lingüísticos em que foram detectadas as CMs.

De uma forma ampla, concluímos que o objetivo primeiro deste estudo foi alcançado. Pretendíamos contribuir para uma apreciação da língua com o mínimo de comprometimento representacionista. Quisemos demonstrar que a língua talvez possa prescindir de tantos modelos teóricos e rótulos. Como Wittgenstein define em *Da certeza* (§559), "o jogo de linguagem é, por assim dizer, imprevisível. Quero dizer: não está fundamentado. Não é racional (ou irracional). Está aí - como a nossa vida". Portanto, como jogadores, talvez a atitude mais prudente seja a constatação e descrição de partes dos jogos, sem tentar alçar vôos teóricos mais ambiciosos. Por ora, tarefa cumprida.

6
Referências Bibliográficas

ARISTÓTELES. Categories. In BARNES (org.). **The complete works of Aristotle.** vol 1, Princeton University Press, 2000.

AIRES, R.V.X.; ALUÍSIO, S.M. Criação de um *corpus* com 1.000.000 de palavras etiquetado morfossintaticamente. **Série de Relatórios do NILC**, NILC-TR-01-8, 2001.

ARANHA, C; FREITAS, M.C.; DIAS, M.C.; PASSOS, E. **Um modelo de identificação e desambigüização de palavras e contextos.** Anais do TIL 2004, Salvador, 2004.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information Retrieval**. Addison Wesley Longman Publishing Co. http://www.dcc.ufmg.br/irbook/, 1999.

BASÍLIO, M. Questões Clássicas e Recentes na Delimitação de Unidades Lexicais. In BASÍLIO, M. (org.). **A delimitação das unidades lexicais**. *Palavra* n° 5. Rio de Janeiro, Departamento de Letras da PUC: Grypho, 1999. p. 9 -18.

BASÍLIO, M.; OLIVEIRA,C.; GARRÃO, M. A não-delimitação das unidades lexicais. In HENRIQUES, C. (org.) **Linguagem, Conhecimento e Aplicação: estudos de língua e lingüística.** Rio de Janeiro, Ed. Europa, 2003. p. 137-148.

BANERJEE, S.; PEDERSEN, T. The design, implementation, and use of the Ngram statistic package. In GELBUKH, A. **Computational Linguistics and Intelligent Text Processing:** Fourth International Conference on Intelligent Text Processing and Computational linguistics, City of Mexico: Springer Berlin/ Heidelberg 2003. p.370-381.

BIDERMAN, M.T. Conceito Lingüístico de Palavra. In BASÍLIO, M. (org.). **A delimitação das unidades lexicais**. *Palavra* n° 5. Rio de Janeiro, Departamento de Letras da PUC: Grypho, 1999. p. 81-97.

BOAS, F. **The Handbook of American Indian Languages.** Washington D.C.: Smithsonian Institute, 1911.

BOGURAEV, B.; PUSTEJOVSKY, J. Issues in Text-based Lexical Acquisition. In: BOGURAEV, B. & PUSTEJOVSKY, J (orgs.). **Corpus Processing for Lexical Acquistion.** Cambridge, Massachusetts: MIT Press. 1995

BORBA, F. S. **Circulação do léxico e direção das alterações semânticas.** Araraquara, UNESP, 2001.

CHOMSKY, N. **Syntactic Structures.** The Hague: Mouton & Co, 1957.

CRUSE, D. **Lexical Semantics**. Cambridge, Inglaterra: Cambridge University Press,1986.

FILLMORE, C.J. **Innocence: a second idealization for linguistics.** Berkeley Linguistics Society, 5, 1979. p. 63-76.

FIRTH, J. A Synopsis of Linguistic Theory 1930-1955. In PALMER, F. (ed.) **Selected Papers of J. R. Firth**. Longman, Harlow, 1968.

GAATONE, D. À quoi sert la notion d´ "expression figée"?, **Lexique, Sintaxe et Sémantique**. Mélanges offerts à Gaston Gross, Besançon: Pufc, Centre Lucien Tesniére, 1990. p. 265-308.

GARRÃO, M. **Um estudo de expressões cristalizadas e sua inclusão em um tradutor automático bilíngüe: o caso de *bater+SN***. Dissertação de Mestrado inédita, PUC-Rio, 2001.

GARRÃO, M; DIAS, M. C. **Um estudo de expressões cristalizadas e sua inclusão em um tradutor automático bilíngüe**. Cadernos de Tradução no. VIII, NUT- UFSC, 2001/2. p.165-182.

GARRÃO, M.; OLIVEIRA, C.; FREITAS, M.C.; DIAS, M.C. Corpus-based Compositionality. In VIEIRA et al. (Eds.). **Computational Processing of the Portuguese Language**. The 7th International Workshop. Springer Berlin/ Heidelberg , 2006. p. 268-271.

GIBBS, R. Jr. **The Poetics of Mind.** Cambridge University Press. EUA, 1995.

GLOCK, H.J. **Dicionário Wittgenstein.** Rio de Janeiro: J. Zahar, 1996.

GROSS, G. **Les expressions figées em français:** noms composés et autres locutions. Collection L´essentiel français, Ophrys, Paris, 1996.

GROSS, M. **Une Classification des phrases figées en français.** In Revue Québecoise de Linguistique, 11, 1982. p. 151-185.

GUENTHNER, F.; BLANCO, X. **Multi-lexemic expressions: an overview**. In *Lingüística Investigaciones Supplementa,* Amsterdam/Philadelphia: Benjamins, 2004. p. 201-218.

HARRIS, R. **The Language Connection**. Thoemmes Press. UK, 1996.

JACKENDOFF, R. **The Architecture of lhe Language Faculty.** Cambridge, Massachusetts: MIT Press, 1997.

KATZ, J.; FODOR, J. **The Structure of a Semantic Theory**. Language, 39. 1963. p. 170-210.

KEMPSON, R. **Semantic Theory.** Cambridge: Cambridge University Press. 1995.

KILGARRIFF, A. **I Don't Believe in Word Senses.** Computers and the Humanities, 31 (2), 1997. p.91-113.

LAKOFF, G. **Women, Fire and Dangerous Things:** what categories reveal about the mind. Chicago: University of Chicago Press, 1987.

_____ . The Contemporary Theory of Metaphor. In ORTONY, A. (ed.). **Metaphor and Thought**. Cambridge: Cambridge University Press, 1993.

LAKOFF, G.; JOHNSON, M. **Metaphors we live by.** Chicago, IL: University of Chicago Press, 1980.

LANGACKER, R. **Concept, Image and Symbol**. Berlin/New York: Mouton de Gruyter, 1991.

MANNING, C.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing.** Cambridge, Massachusetts: MIT Press, 1999.

MARTINS, H. **Metáfora e Polissemia no estudo das línguas do mundo:** uma aproximação não representacionista. Tese de Doutorado inédita, UFRJ, 1999.

_____. Três Caminhos na Filosofia da Linguagem. In MUSSALIM, F; BENTES, A.C. (orgs.). **Introdução à Lingüística.** Volume III, São Paulo: Cortez Editora, 2003. p. 439-474.

NEVES, M.H.M. A delimitação das unidades lexicais: o caso das construções com verbo-suporte. In BASÍLIO, M. (org.) **A delimitação das unidades lexicais**. *Palavra* n° 5. Rio de Janeiro, Departamento de Letras da PUC: Grypho, 1999. p. 98-114.

NIETZSCHE, F. **Sobre verdade e mentira no sentido extra-moral**. Coleção Os Pensadores. São Paulo: Abril Cultural, 1978 [1873].

NOGUEIRA, C. **Algoritmo para extração de Combinações do tipo V+(det)+N**. Programa feito em linguagem Java. IME, 2004.

PINKER, S. **The Language Instinct.** New York: Harper Perennial, 1995.

RANCHHOD, E. O Lugar das Expressões Fixas na Gramática do Português. In CASTRO, I.; DUARTE, I. (eds.). **Razões e Emoção. Miscelânea de estudos oferecida a Maria Helena Mira Mateus.** Lisboa: Imprensa Nacional - Casa da Moeda, 2003, pp. 239-254.

SALOMÃO, M. M. **Polysemy, Aspect and Modality in Brazilian Portuguese**: The Case for a Cognitive Explanation of Grammar. Tese de Doutorado, University of California at Berkeley, 1990.

SCHERER, M. **Uma questão de vocabulário**: considerações sobre o campo lexical no ensino de português para estrangeiros. Dissertação de Mestrado inédita. PUC-Rio, 2002.

SEARLE, J. **Literal Meaning.** Erkenntnis, 13, 1978. p. 207-224.

SINCLAIR, J. **Corpus, Concordante, Collocation.** Oxford University Press, Oxford, 1991.

SOUSA FILHO, D.M. **Iniciação à história da Filosofia.** Rio de Janeiro: Jorge Zahar, 1997.

STUBBS, M. **Text and corpus analysis:** computer-assisted studies of language and culture. Oxford: Blackwell, 1996.

SUMARES, M. **Sobre Da Certeza de Ludwig Wittgenstein.** Lisboa: Editora Afrontamento, 1994.

TAGNIN, S. **Um dicionário de colocações verbais. Para quê?** Anais do V PROPOR; Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. São Carlos, SP : ICMC/USP, 2000. p. 13-28.

VALE, O. **Expressões Cristalizadas no português do Brasil:** uma proposta de tipologia, Tese de Doutorado, Araraquara: UNESP, 2002.

VERLINDE, S; SELVA, J. **Corpus-based versus intuition-based lexicography**: defining a word list for a French learners disctionary. In Proceedings of the Corpus Linguistics 2001 Conference. Lancaster University, UK, 2001. p. 594-598.

WHORF, B.L. An American Indian Model of the Universe. In NYE, A. (org.) **Philosophy of Language:** The Big Questions. Malden: Blackwell, 1998.

WITTGENSTEIN, L. **Investigações Filosóficas.** Coleção Os Pensadores, São Paulo: Abril Cultural, 1979.

ZIPF, G. K. **Human Behavior and the principle of least effort** Cambridge, MA: Addison-Wesley Press, 1949.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo