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.