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

PUC-Rio - Certificação Digital Nº 0210380/CA

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.

PUC-Rio - Certificação Digital Nº 0210380/CA

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.

PUC-Rio - Certificação Digital Nº 0210380/CA

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.

PUC-Rio - Certificação Digital Nº 0210380/CA

* * *

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.

PUC-Rio - Certificação Digital Nº 0210380/CA