5

O substantivo-suporte no corpus

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong. (Richard P. Feynman)

5.1

Introdução

A Lexicografia Computacional iniciou-se como um conjunto de métodos automáticos de utilização de versões digitais de dicionários convencionais como fonte de informações para sistemas de PLN (Ooi 1998). Com a ampliação dos recursos disponíveis - evolução das tecnologias e a disponibilidade de vastas coleções de textos - e o aumento das exigências sobre os recursos lexicais utilizados por sistemas de PLN cada vez mais sofisticados, a Lexicografia Computacional passou a contribuir com questões relevantes da Lexicologia e da Lexicografia Lingüística.

Dentro desse contexto, o trabalho experimental de delimitação das expressões com substantivo-suporte consiste na análise de contextos para caracterizar os elementos do significado que se organizam dentro da expressão considerada, e como se comporta semanticamente essa expressão dentro de contextos maiores.

Na seção 5.2 consideramos a primeira parte do problema, ou seja o processo de composição de informações semânticas provenientes de duas palavras em uma expressão.

Na seção 5.3 consideramos a segunda: a expressão em contexto.

5.2

Composicionalidade semântica

O termo *composicionalidade* foi utilizado por Katz & Fodor (1963) para descrever um aspecto da competência semântica do falante de uma língua. Para os autores, composicionalidade é um modelo de interpretação semântica e produção baseado em um cálculo: cada átomo ou unidade lingüística, como a palavra, contribui para o significado global do enunciado, sistematicamente.

A principal idéia é a existência de um mecanismo pelo qual o significado de uma cadeia lingüística é decodificada a partir do que o falante sabe sobre cada unidade componente e sobre a organização dessas unidades. Cada vez que uma estrutura ou enunciado ocorre, seu significado é composicionalmente calculado. Segundo Searle, a idéia de composicionalidade de um enunciado está associada ao seu significado literal:

“The literal meaning of a sentence is entirely determined by the meanings of its component words (or morphemes) and the syntactical rules according to which these elements are combined”. (Searle 1979, p. 117)

Assim, o autor garante a existência de pelo menos uma interpretação do enunciado por meio desse cálculo. O otimismo com relação à possibilidade de seleção desse conjunto de pares ⟨palavra, significado literal⟩, para cada enunciado, vem sendo abandonado desde então.

Cruse (1986) apresenta a semântica de uma sentença como o resultado da combinação de sentidos mais simples, ou seja, contribuições semânticas de constituintes cada vez mais simples. Assim, o autor caracteriza um **constituente semântico** (p.25) como um par ⟨forma, significado⟩. Cruse pressupõe a associação estável entre um significado e um constituinte sentencial que se repete a cada ocorrência do constituinte. Fica patente a noção de composicionalidade subjacente ao conceito. Nesse contexto, a não-composicionalidade de constituintes semânticos atípicos, referidos por Cruse como “peripheral types of semantic constituents” (p. 29) seria devida à “unicidade colocacional”, que explicaria a discrepância do cálculo composicional.

Fillmore (1979), por outro lado, é particularmente crítico ao modelo de cálculo, a que chama de “a segunda idealização” da Lingüística, em referência ao construto do falante ideal do modelo chomskyiano.

Ao argumentar que o substantivo, ou o adjetivo, é o centro do significado em um enunciado, a noção de composicionalidade é utilizada implicitamente. Será necessário então adotar o modelo do cálculo para postular a existência das expressões com substantivo-suporte? Utilizo (Cruse 2004) para me posicionar com relação a esta questão. O autor enuncia o **princípio da composicionalidade** como:

“The meaning of a grammatically complex form is a compositional function of the meaning of its grammatical constituents.” (Cruse 2004, p.65)

Sendo esta a versão “forte” do princípio, Cruse separa três proposições subsumidas:

1. “The meaning of a complex expression is completely *determined* by the meaning of its constituents.”
2. “The meaning of a complex expression is completely *predictable* by general rules from the meaning of its constituents.”
3. “Every grammatical constituent has a meaning which contributes to the meaning of the whole” (*Ibid.*)

Das três, a terceira é a menos ambiciosa, exigindo apenas uma participação sistemática dos constituintes na interpretação do todo da expressão. Essa é a única possibilidade de compromisso entre a metodologia apresentada a seguir e o conceito tradicional de composicionalidade.

5.3 O Método

A identificação das ocorrências de substantivos-suporte no corpus pode ser vista como uma instância de Desambiguação do Sentido Lexical. A questão central é como o falante distingue o sentido de uma palavra ambígua ou vaga quando se depara com um enunciado onde ela ocorre. Trata-se de um problema muito relevante em Linguística Computacional visto que o computador, a princípio, não conta com esta capacidade de discernir, uma função essencial do sistema lingüístico humano, de extrema eficiência. Uma formulação simplificadora do problema seria separar em agrupamentos as diversas ocorrências de uma palavra ambígua, de acordo com seu significado na ocorrência.

O método desenvolvido para este trabalho tem como motivação a questão colocada por Kilgarriff (1997) quando discute o que é o sentido de uma palavra. A partir de um modelo computacional em que o sentido da palavra (*word sense*) ou a unidade lexical não são unidades básicas, mas sim as ocorrências de uma palavra em contexto, operacionalizadas como extratos de um corpus, ele elabora:

“To know what a word sense s_1 is, is to know which uses of the word are part of s_1 and which are not, probably because they are part of s_i , where $i \neq 1$. If we are to know what word senses are, we need operational criteria for distinguishing them” (Kilgarriff 1997, p.97)

Nosso método é, portanto, um critério operacional para distinguir substantivos-suporte e suas ocorrências dentro de um corpus.

Em um trabalho que analisa as expressões verbais da forma (Verbo + Sintagma Nominal), Garrão (Garrão 2006) discute esta questão ao propor que, na construção de léxicos eletrônicos, a delimitação das unidades lexicais complexas, as chamadas Expressões Multi Vocabulares (EMVs), deva ser compreendida com base em uma análise de corpus. A abordagem experimental utilizada em (Garrão 2006) e introduzida em (Garrão *et al.* 2006) é bem próxima à empregada aqui.

5.3.1

O Modelo do Espaço Vetorial

Schütze & Pederson (1995) introduziram o Modelo do Espaço Vetorial em uma abordagem bastante original para o problema de Desambiguação de Sentidos Lexicais. A idéia é utilizar vetores de grandes dimensões na representação do contexto de cada ocorrência de uma palavra em foco (w), e agrupar esses contextos por uma medida de similaridade entre os vetores. Os agrupamentos obtidos seriam avaliados de acordo com sua coesão: os mais coesos corresponderiam a sentidos diferentes da palavra w . Para identificar o sentido de uma nova ocorrência de w , essa ocorrência é representada por um vetor que será comparado aos centróides dos agrupamentos previamente construídos. Os sentidos eleitos serão os representados pelos agrupamentos mais próximos ao vetor da nova ocorrência.

O Modelo do Espaço Vetorial, descrito com grande clareza em (Manning & Schütze 1999), é o mais utilizado entre os modelos de recuperação de documentos, devido à sua simplicidade conceitual e a clareza da metáfora da proximidade espacial entre documentos, representados como vetores de palavras. A representação de um documento nesse modelo é feita considerando cada palavra como uma dimensão no espaço vetorial; o tamanho do vocabulário contido na coleção de documentos determina a dimensão total do espaço. A proximidade entre documentos é medida pelo ângulo entre os respectivos vetores.

Como exemplo, vamos considerar as seguintes expressões como os dois únicos textos de uma coleção:

Doc1: a dog and a cat

Doc2: and a frog and a cat

O espaço vetorial correspondente seria composto por cinco dimensões

$(a, dog, and, cat, frog)$

e conteriam dois vetores. Considerando a frequência bruta da palavra como o tamanho do vetor na dimensão dessa palavra, os vetores seriam:

$$(2, 1, 1, 1, 0)$$

$$(2, 0, 2, 1, 1)$$

Qualquer que seja a língua em questão, é natural que nem todas as palavras que ocorrem no documento sejam importantes na avaliação de seu tópico. Em sistemas manuais de indexação de documentos, os **termos** são determinados manualmente por um especialista, que selecionam palavras ou expressões que descrevam o conteúdo do documento, tendo em vista as palavras-chaves que o usuário do sistema irá fornecer para encontrar o documento indexado. Já em um sistema automático, principalmente os que lidam com documentos de domínios diversos, o primeiro passo para a seleção dos termos para a indexação é a eliminação das palavras gramaticais, normalmente listadas em um conjunto denominado **stopwords**. Muitas vezes o conjunto de stopwords inclui também palavras de altíssima frequência, já que isso indicaria baixo poder discriminatório. Não utilizei esse tipo de stopword pois certamente teria que incluir alguns substantivos-suporte.

O segundo passo no tratamento das palavras que serão indexadas é **pesagem dos termos**. Em Recuperação de Informações, esta pesagem é uma maneira de capturar a noção de **poder discriminatório**, promovendo com pesos altos as palavras de maior poder. Dessa maneira, existem três fatores principais na pesagem dos termos: **frequência do termo** (número de ocorrências de um termo em um documento), **frequência documental** (número de documentos onde ocorre o termo) e **coeficiente de normalização** em função do tamanho do documento. Esses fatores são combinados para definir o esquema de pesagem dos termos.

A frequência do termo quantifica a participação do termo no documento, portanto conta positivamente na pesagem. A frequência documental, sendo alta, tende a indicar que o termo não é muito informativo, pois palavras de maior especificidade tendem a se concentrar em pequenos conjuntos de documentos, enquanto que palavras de significado mais geral tendem a ocorrer em toda a coleção. Portanto o fator frequência documental conta negativamente. Finalmente, o coeficiente de normalização é utilizado para compensar o fato de que documentos maiores tendem a conter um maior número de termos distintos, o que faz com que sejam mais facilmente recuperados que os menores. Dentre os muitos esquemas de pesagem avaliados na literatura, o esquema que utilizei, que multiplica a frequência do termo pelo inverso da frequência documental, junto com a normalização do tamanho do documento, apresenta um

desempenho equiparável aos melhores (Salton & Buckley 1988).

O cálculo das similaridades entre dois vetores de documentos pode ser feito de várias maneiras, utilizando medidas como a distância de Euclides, coeficiente de Jaccard, coeficiente de Dice e a medida do cosseno, para citar as mais usadas. Essa última baseia-se no ângulo entre dois vetores de documentos (Baeza-Yates & Neto 1999). O cosseno entre os vetores $\vec{u}$ e $\vec{v}$ é dado pelo coeficiente entre o produto escalar dos vetores por seus módulos:

$$\cos(u, v) = \frac{\sum(u_i \times v_i)}{\sqrt{\sum u_i^2 \times \sum v_i^2}}$$

Por exemplo, a medida do cosseno de similaridade entre “a dog and a cat”(2,1,1,1,0), e “a frog and a cat”(2,0,2,1,1) é dada por:

$$\frac{2 \times 2 + 1 \times 0 + 1 \times 2 + 1 \times 1 + 0 \times 1}{\sqrt{(2^2 + 1^2 + 1^2 + 1^2 + 0) \times (2^2 + 0 + 2^2 + 1^2 + 1^2)}} = 0.84$$

A matriz de similaridades dos n documentos é uma matriz simétrica $n \times n$, onde cada célula m_{ij} contém a medida de similaridade entre os documentos d_i e d_j .

5.3.2

Explorando o contexto

Nessa parte experimental do trabalho, definimos **contexto** de uma dada palavra ou expressão como o parágrafo do corpus onde ocorre. Visto que o corpus foi construído a partir de resultados de busca na Internet¹, ao invés de coletar uma só sentença ou um número arbitrário de sentenças, decidi coletar parágrafos inteiros, de tamanhos variados.

Para cada expressão $S + Adj$ a ser examinada, coletei três conjuntos de contextos:

conj1 - contextos contendo $S Adj$

conj2 - contextos contendo S mas não $S Adj$

conj3 - contextos contendo Adj mas não $S Adj$

Os seguintes contextos são exemplos dos conjuntos coletados para a análise da expressão NATUREZA CULTURAL.

ex. 5.1

¹O corpus descrito na seção 4.2.6 mostrou-se insuficiente para prover o número de exemplos necessários para o nosso método. Busquei coletar os parágrafos na Internet de maneira sistemática, utilizando o Google™, processando as páginas de resultados de 3 em 3, para cobrir em extensão as páginas fornecidas pelo buscador.

conj1 - *Os Estados-Partes estimularão a adoção de medidas que facilitem o trânsito de agentes culturais, vinculados à execução dos projetos de NATUREZA CULTURAL.*

conj2 - *A modéstia deveria ser cultivada da mesma maneira pelos dois sexos [...] mas há homens que vão continuar repetindo que a mulher deve ter mais modéstia do que o homem [...] eles são incapazes de apreciar o prazer que provoca a virtude da renúncia [no entanto:] As mulheres, tanto quanto o homem, deveriam ter as paixões e os desejos próprios à sua NATUREZA; a bestialidade só aparece quando a razão não os controla, mas a obrigação de controlar paixões e desejos é o dever de toda a humanidade e não o dever de um sexo só.*

conj3 - *Com seu orçamento invariavelmente contingenciado, o Ministério da Cultura bem que gostaria, mas tem optado por não mexer na lei de incentivo por temer um colapso na produção CULTURAL brasileira que nestes dez anos se tornou absolutamente dependente deste tipo de patrocínio - que, na verdade, não é real, porque está baseado em isenção fiscal. Esta é a justificativa apresentada, por exemplo, para que o decreto que traz as primeiras mudanças na Lei Rouanet na gestão de Gilberto Gil esteja juntando pó na Casa Civil.*

5.3.3

Medindo a composicionalidade

No modelo do espaço vetorial, a composicionalidade de uma expressão $[X Y]$ pode ser heurísticamente medida usando as similaridades entre os contextos de $[X Y]$, X e Y . Assume-se que, se $[X Y]$ é não composicional, os contextos de $[X Y]$ são muito distantes dos de X e de Y . Zhai (1997) chama essa medida de “Context Similarity”, e a utiliza na delimitação de expressões multi-vocabulares (ou “átomos lexicais”).

Em (Garrão *et al.* 2006), um método semelhante é empregado para verificar quando a expressão $V SN$ é não composicional e, de uma perspectiva lexicográfica, propor que sejam armazenadas em um léxico computacional. Os exemplos 5.2 e 5.3 mostram contextos daquele trabalho: em 5.2, USAR CAMISINHA é considerado composicional e em 5.3 TOMAR PARTIDO é considerado não composicional. Garrão *et al.* também consideram alguns casos limítrofes, onde o método é inconclusivo.

ex. 5.2

USAR CAMISINHA

Se estão com Aids, eles informam o freguês ou companheiro e pedem para USAR CAMISINHA.

Você não ia ser louca de ter vida sexual sem USAR CAMISINHA, não é verdade?

É consenso entre os jovens de ambas nacionalidades a necessidade de se USAR CAMISINHA durante a transa.

CAMISINHA

Nesse sentido, pode ser inconveniente para muitas mulheres chamar a atenção para a CAMISINHA e estabelecer negociações sobre sexo seguro antes que tenham intimidade com seu parceiro.

O pai de André, o médico epidemiologista Glacus de Souza Brito, 36, um dia levou para casa uma CAMISINHA, desenrolou-a e mostrou ao filho.

Como psicólogo, ele acha que, para os meninos menores, a escolha da CAMISINHA como objeto de coleção tem a ver com o tamanho do pênis.

ex. 5.3

TOMAR PARTIDO

Mas é fácil, como dizia Swift, reconhecer o gênio: todos os imbecis da época TOMAR PARTIDO contra ele.

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.

Os sérvios acusam a ONU e a Otan de terem TOMAR PARTIDO dos muçulmanos e croatas na guerra.

PARTIDO

A busca de coligações será um trabalho feito pela direção do PARTIDO.

As pressões virão de líderes antiquercistas do PARTIDO que ainda não conseguiram articular uma ação conjunta contra ele.

Petistas tentam mobilizar militância do PARTIDO para reverter a queda contínua nas pesquisas de opinião.

Essa medida é adequada para os testes de identificação do meu objetivo, considerando-se que o padrão de composicionalidade esperado em uma expressão com substantivo-suporte *S Adj* será de tal modo que *S* contribui muito pouco para o significado, em comparação com a contribuição de *Adj*, uma

espécie de composicionalidade desequilibrada. Nos resultados das experiências, as evidências para essa hipótese são dadas por altos níveis de similaridade entre os contextos da expressão *S Adj* e os contextos da palavra que mais contribui na expressão. Portanto,

Similaridade entre os contextos de *S Adj* (**conj1**) e *Adj* (**conj3**)
é maior que
Similaridade entre os contextos de *S Adj* (**conj1**) e *S* (**conj2**).

Um resumo algorítmico do método proposto para computar o caráter de suporte do substantivo é apresentado na figura 5.1.

- 1 Selecionar n contextos com *S Adj*, *S* sem *Adj* e *Adj* sem *S*;
- 2 Remover as palavras *S* e *Adj* dos contextos
- 3 Indexar os contextos; e
- 4 Calcular a matriz de similaridades entre os três conjuntos de contextos.

Figura 5.1: Resumo do método de identificação de substantivos-suporte em contexto

O passo 3 foi incluído para que as palavras que constituem o padrão segundo o qual o contexto foi selecionado não sejam um fator de influência na similaridade entre contextos, especialmente porque os contextos são pequenos textos que poderiam ficar distorcidos pela presença de termos frequentes.

5.4 Experimentos

Um bom critério operacional para a caracterização do substantivo-suporte consiste na extração de expressões nominais *S Adj*, onde *Adj* é um adjetivo denominal. A lista resultante de substantivos pode ser analisada para determinar intercambialidade, isto é, co-ocorrência com o mesmo grupo de adjetivos mantendo o significado da expressão relativamente estável.

Utilizando o critério de intercambialidade, foi selecionada desse corpus uma lista preliminar de substantivos-suporte. Tomando como base essa lista, uma seleção de combinações *S Adj* foram identificadas, para servirem como dados empíricos para validar a hipótese subjacente ao método descrito na seção 5.3.3. A tabela 5.1 mostra as expressões selecionadas.

As tabelas 5.2, 5.3 e 5.4 mostram os resultados desses testes, multiplicados por um fator de 100.000 para que os valores sejam melhores visualizados.

<i>S Adj</i>	<i>S</i>	<i>Adj</i>
conj1	conj2	conj3
fator biológico	fator	biológico
papel biológico	papel	biológico
natureza cultural	natureza	cultural
natureza econômica	natureza	econômica
cunho cultural	cunho	cultural

Tabela 5.1: Expressões *S Adj* testadas

<i>S Adj</i>	<i>S</i>	<i>Adj</i>
fator biológico	19.27	43.42
papel biológico	8.38	48.86
natureza cultural	19.90	84.24
natureza econômica	13.68	99.13
cunho cultural	14.98	56.47

Tabela 5.2: Resultados com as expressões *S Adj*

A tabela 5.2 pode ser interpretada da seguinte forma. Tomando como exemplo a expressão na primeira linha, FATOR BIOLÓGICO, a similaridade entre os contextos de FATOR BIOLÓGICO e os contextos do substantivo FATOR é 19.27; a similaridade entre os contextos de FATOR BIOLÓGICO e os contextos do adjetivo BIOLÓGICO é 43.42. É importante observar que os valores de similaridade em cada linha são relativos somente aos contextos envolvidos na sintagma *S Adj*, portanto, somente os valores na mesma linha podem ser comparados. Cada grupo de três conjuntos de contextos (*S Adj*, *S* e *Adj*) foi indexado separadamente. Na prática, para cada conjunto, tem-se uma coleção independente, e valores de peso das palavras não se correlacionam de uma coleção para outra.

Em cada linha da tabela 5.2, a similaridade entre *S Adj* e *Adj* é mais do que o dobro da similaridade entre *S Adj*, *S*. Assim, pode se concluir que o contexto em que o adjetivo aparece e os contextos nos quais o sintagma nominal aparece são muito mais próximos daqueles contendo o substantivo e a sintagma nominal, dando suporte à hipótese inicial.

De modo a promover ainda mais a hipótese de substantivos-suporte, foi explorada a idéia de que, se o substantivo nas expressões nominais *SAdj* for substantivo-suporte, satisfazendo assim o critério de intercambiabilidade, as similaridades entre expressões nominais S_1Adj e S_2Adj , com $S_1 \neq S_2$, devem ser altas. Portanto, foi realizado o seguinte teste: depois de selecionar um adjetivo denominal e dois substantivos-suporte, foram comparados os contextos das expressões resultantes.

Na tabela 5.3, os substantivos são FATOR e PAPEL e o adjetivo é BIOLÓGICO; na tabela 5.4, os substantivos são CUNHO e NATUREZA e o adjetivo

fator biológico	
vs biológico	21.94
papel biológico	
vs biológico	25.44
fator biológico	
vs papel biológico	192.30

Tabela 5.3: Resultados com *biológico*

cunho cultural	
vs cultural	27.06
natureza cultural	
vs cultural	39.60
cunho cultural	
vs natureza cultural	56.12

Tabela 5.4: Resultados com *cultural*

é CULTURAL. Os resultados demonstram que as expressões S_1Adj e S_2Adj são consideravelmente mais próximas do que S_1Adj e Adj ou S_2Adj e Adj , sustentando ainda mais a hipótese inicial.

Discussão

Muitas vezes, uma só palavra ou expressão é o suficiente para que o falante localize o discurso dentro de um assunto específico. Por exemplo, analisando os parágrafos do corpus que contêm o padrão S -DIDÁTICO, encontram-se 329 exemplos, a vasta maioria a respeito de ‘educação’ e ‘ensino’. A noção de **campo semântico**, grupo de palavras que co-ocorrem em um certo tipo de contexto, identificando uma relação de tópico (Stubbs 2002), é adequada para descrever o vocabulário encontrado no extrato do corpus, em que se destacam as frequências de algumas palavras como LIVRO, MATERIAL, RECURSO, TEXTO e AULA. Poucos parágrafos desse conjunto não se enquadravam nesse tópico, entre eles uma exceção digna de nota:

ex. 5.4 *As coisas, os acidentes de percurso, os loucos atores do pastelão DIDÁTICO: Collor, Alves dos Santos, PC, as CPIs, tudo.*

Essas expectativas com relação às palavras e seus contextos são obviamente individuais, de cada falante, resultantes de sua vivência lingüística e seu conhecimento de mundo. No entanto, um corpus suficientemente representativo da experiência lingüística e cultural de um grupo é um instrumento fidedigno de medição de expectativa.

Os adjetivos que compõem as expressões apresentadas nos estudos de casos foram escolhidos por gerarem esse tipo de expectativa quanto ao campo semântico em que se inserem. Busquei o contraste entre o substantivo-suporte, vago, e o adjetivo, específico. Pelas características discutidas na seção 3.4.2, os adjetivos denotativos se apresentam muitas vezes com esse poder discriminatório quanto ao tema do contexto em que ocorrem. Assim, os resultados tão categóricos que obtive têm relação com essas escolhas.