Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Identificação das Estruturas Argumentais dos Adjetivos: uma Abordagem Semi-automática Baseada em Córpus

Ariani Di Felippo
Thiago Alexandre Salgueiro Pardo
Sandra Maria Aluísio

NILC-TR-05-14

Outubro, 2005

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Devido às aplicações para as quais os sistemas de Processamento Automático das Línguas Naturais (PLN) são feitos, é premente a construção de léxicos que armazenam informações semânticas. Neste trabalho, propõe-se uma abordagem semi-automática baseada em córpus para a identificação da estrutura de argumentos (ou valências) dos adjetivos do português do Brasil. Essas estruturas são uma propriedade semântica essencial dos predicadores e podem ser utilizadas para o enriquecimento de léxicos para fins do PLN.

Este trabalho conta com o suporte das agências FAPESP, CAPES, CNPq e Comissão Fulbright.

ÍNDICE

1. Introdução

Atualmente, em função das aplicações reais para as quais os sistemas que processam língua natural são escritos, é premente, na construção de um tradutor automático, por exemplo, a compilação de léxicos (monolíngües e/ou multilíngües) que sejam: (i) *manipuláveis* pelo sistema do qual fazem parte, isto é, léxicos cujas informações sejam explicitamente especificadas por meio de um esquema de representação formal (ou formalismo); (ii) *lingüisticamente motivados,* tanto do ponto de vista da robustez (isto é, léxicos que contenham uma quantidade de unidades compatível com o léxico de uma língua natural) quanto da pertinência das informações associadas às entradas lexicais [Handke 1995; Viegas e Raskin 1998; Palmer 2001]. Dessa forma, a construção de léxicos para fins do Processamento Automática das Línguas Naturais (PLN) ou da Tecnologia da Linguagem Humana (TLH) requer, sobretudo, a investigação das propriedades dos itens lexicais.

Na década de 1990, as "recomendações" do projeto EAGLES [Sanfilippo 1996], e, mais recentemente, a formação do ISLE *Computational Lexicon Working Group* [Calzolari et al. 2001] evidenciam a urgente necessidade de se armazenar, em léxicos computacionais, informações de natureza semântica que sejam lingüisticamente relevantes para o processamento automático das línguas naturais. Diante dessa necessidade, vários estudos que focam a descrição das propriedades léxico-semânticas começaram a ser desenvolvidos. De um modo geral, tais estudos têm a classe dos verbos como foco de pesquisa, ou seja, a classe dos verbos tem merecido uma atenção maior, por parte dos investigadores, que classes como a dos adjetivos e advérbios [Pustejovsky 1996]. A tarefa de descrição dos argumentos dos verbos, seus papéis temáticos e comportamentos lingüísticos preferenciais [Levin 1993] representa uma parcela significativa dos objetivos desses estudos.

Reagindo, assim, à tendência apontada, privilegia-se, neste trabalho, a classe dos adjetivos (Adjs), sendo que o objetivo é o de propor uma abordagem semi-automática baseada em córpus (seguindo a abordagem proposta por Pardo et al. 2005) para a identificação da *estrutura de argumentos* (lexicalizadas e generalizadas) dos adjetivos de valência 1 (V_1) (ou seja, aqueles que projetam apenas um argumento), em posição adnominal (Padn), do português do Brasil. Essa especificação, aliás, pode contribuir para o enriquecimento semântico de léxicos para fins do PLN.

Na seção seguinte (Seção 2), uma breve revisão da literatura sobre os adjetivos é apresentada com o objetivo de delimitar o objeto de estudo deste trabalho e alguns trabalhos correlatos feitos para o inglês são pontuados. Na Seção 3, descrevem-se as etapas que compõem a abordagem ora proposta e explicitam-se os recursos lingüístico-computacionais utilizados em cada uma das etapas. Na Seção 4, são feitas algumas observações a respeito dos resultados obtidos. Por fim, na Seção 5, discutem-se as potencialidades e limitações da abordagem e os trabalhos futuros.

2. Dos Adjetivos Valenciais do Português e Trabalhos Correlatos

2.1. Dos adjetivos predicadores ou valenciais

Para a realização deste trabalho, partiu-se do princípio de que o adjetivo é elemento de natureza abstrata e que é próprio do adjetivo não incidir sobre si mesmo, mas sobre um suporte. Todo adjetivo, portanto, comporta uma incidência sobre outra coisa que não ele mesmo, sendo que essa "coisa" é, na maioria das vezes, um nome (N) ou substantivo [Borba 1996]. Além disso, considerou-se que os adjetivos do português podem ser divididos em duas classes: *classificadores* (CLs) e *qualificadores* (Q Ls) [Borba 1996; Neves 2000][1]. Esses dois tipos de adjetivos estabelecem relações sintático-semânticas distintas com os nomes sobre os quais incidem. Considerem-se os exemplos (1-3) retirados de Borba (1996) e Neves (2000):

(1) Mas o pessoal do Levita tem de investigar a infiltração *comunista* nessa festa.

(2) O rapaz é trabalhador *profissional*.

(3) a. Angela conseguiu um abatimento *impressionante* na compra. (Padn posposta)
b. Em seu lugar, ficou *a nebulosa* Luela. (Padn anteposta)
c. O negrão é *grande*, mas não é dois. (posição predicativa)

Na sentença descrita em (1), o adjetivo CL "comunista" estabelece uma relação nominal com o nome "infiltração" (núcleo do sintagma "a infiltração"). Em outras palavras, diz-se que o adjetivo tem função na estrutura argumental desse nome, isto é, equivale a um sintagma preposicional que expressa o que seria o complemento (nominal) do nome (infiltração *comunista* < *de comunistas*). Em (2), o adjetivo CL "profissional" equivale a um sintagma preposicional de valor adverbial e/ou a um advérbio em –mente: trabalhador profissional < X trabalha com profissionalismo/ profissionalmente. Nas sentenças em (3), os adjetivos QLs "impressionante", "nebulosa" e "grande" estabelecem relação adjetival com os nomes "abatimento", "Luela" e "negrão", respectivamente. Diz-se relação "adjetival" porque esses elementos são, tanto em Padn posposta (3a) ou anteposto (3b) como em posição predicativa (3c), verdadeiros atribuidores de propriedade aos nomes e, por serem semanticamente incompletos, precisam ligar-se a outros elementos para adquirir um valor semântico pleno. *Valência* é, então, o número de elementos implicados por um determinado item lexical[2].

A um item lexical valencial, é dado o nome *predicador* (P) e, aos elementos que dão completude de sentido a ele, é dado o nome *argumentos* (As). Em um nível lógico-semântico, os Ps determinam o número de As com os quais podem ocorrer.

[1] Ressalta-se que, em dependência do nome com o qual ocorre e da posição no sintagma, pode haver permeação entre essas classes. Por exemplo, certos adjetivos que são em princípio QLs podem passar a CLs em sintagmas cristalizados (p.ex.: *água* doce). Dessa forma, a classificação em QL ou CL depende do emprego do adjetivo na sentença.

[2] Os adjetivos do português podem ser de valência 1 (V_1) (p.ex.: garota *simpática*), valência 2 (V_2) (p.ex.: operações *úteis* ao governo), valência 3 (V_3) (p.ex.: réu *condenável* à pena máxima pelo juiz) e valência 4 (V_4) (p.ex.: livros *transferíveis* da diretoria para a biblioteca pelo funcionário) [Borba 1996].

Assim, chega-se à fórmula básica $P(A_n)$[3]. Por exemplo, na sentença (4) *O garoto é esperto*, "esperto" requer apenas um A, "garoto". Além do número de As, um P também impõe restrições sobre esses As. O adjetivo "esperto" (4), por exemplo, impõe sobre seu A o traço [+ animado] (restrição de seleção). Dessa forma, construções do tipo *"a* mesa *é esperta"* ou *"a* mesa *esperta"* não são semanticamente bem-formados, pois "mesa" tem o traço [+ não-animado]. A relação P(A) também estabelece funções semânticas específicas para os As. Tais funções são comumente denominadas *temáticas* e são representadas por meio de rótulos denominados *papéis temáticos* ou *casos* (ex.: *agente, paciente,* entre outros) [Raposo 1992]. No caso de "esperto", em (4), o A projetado recebe o papel temático *Tema* (T)[4].

À representação unificada dessas informações (P, número de As, restrição de seleção e papéis temáticos), dá-se o nome *estrutura de argumentos*, que pode ser assim representada: *esperto*($Tema_{[+animado]}$) [Grinshaw 1992][5]. Essas estruturas nada mais são do que codificações de informações conceituais subjacentes ao uso dos predicadores, indicando quantos e quais os argumentos que um P requer quando a predicação indicada por ele é realizada superficialmente.

Neste trabalho, em especial, focalizam-se os adjetivos V_1 (portanto, qualificadores) somente em Padn, pospostos e antepostos, como em (3a) e (3b), respectivamente.

2.2. Dos trabalhos correlatos

Alguns grandes projetos visam a desenvolver repositórios de informação semântica para os itens lexicais da língua inglesa. Entre esses projetos, desenvolvidos por meio de abordagens manuais, destaca-se o *FrameNet* [Baker et al. 1998], posto que a base lexical construída no projeto armazena informações sobre a estrutura de argumentos dos adjetivos. Especificamente, a base *FrameNet* armazena 8.900 unidades lexicais do inglês (verbos, adjetivos e nomes), sendo que 6.100 estão classificadas em *frames* (p.ex.: "congestion", "transportation", entre outros). Um *frame* pode ser visto como o "quadro" semântico ao qual um item lexical pertence. Cada *frame,* aliás, especifica os *elementos-frame* que dele participam (p.ex.: "experiencer", "buyer", "degree", entre outros). Além dos elementos-frame, a *FrameNet* fornece frases-exemplo anotadas com esses elementos. Na Figura 1, exemplifica-se a estrutura de argumentos (ou melhor, o *frame* e os *elementos-frame*) associada ao adjetivo "afraid" na *FrameNet*.

[3] No caso deste trabalho, P é o adjetivo (Adj) e A é o seu argumento (N).

[4] O papel temático *Tema* indica o caso semanticamente mais neutro, é a entidade sobre a qual se verifica uma situação.

[5] Ressalta-se que a concepção de *estrutura de argumentos* não é questão consensual nos estudos lingüísticos. Há divergências no que diz respeito à concepção dos papéis temáticos e aos rótulos empregados para representá-los. Neste trabalho, busca-se fornecer uma visão geral dessa estrutura.

Frame
Experiencer_subj

Elementos-frame
EXPERIENCER is **afraid** CONTENT.

Frase-exemplo
You (Experiencer) are AFRAID of heights (Content)?

Figura 1. Entrada lexical do adjetivo *"afraid"* na *FrameNet*.

Por meio da Figura 1, observa-se que o adjetivo do inglês "afraid" tipicamente pertence ao *frame "experiencer_subj"* ("sujeito experienciador"), que engloba os *elementos-frame "experiencer"* ("experienciador") e *"content"* ("conteúdo": do que se tem medo). Diz-se tipicamente porque a FrameNet também fornece informações sobre os elementos-frame que opcionalmente podem aparecer vinculados aos itens lexicais. O adjetivo "afraid", por exemplo, pode aparecer ligado ao *elemento-frame* opcional *"degree"* ("grau"), como em: *Robbie was very much (Degree) afraid that she knew the answer*. Quanto aos pressupostos teóricos da *FrameNet*, vale salientar que os próprios autores afirmam que a representação da semântica dos predicadores por meio de *frames* (e *elementos-frame*) é bastante similar à representação por meio de papéis temáticos [Baker et al 1998]. A principal diferença entre essas abordagens, segundo Baker at al. (1998), é o fato de que, na FrameNet, o rótulo do argumento (ou elemento-frame) é, muitas vezes, particular a um *frame* e, conseqüentemente, ao conjunto de itens lexicais que a ele pertence. Por exemplo, para os itens do *frame ingestion* ("ingestão"), o argumento que receberia o papel temático *Agente* é anotado com o rótulo específico daquele *frame*: *ingestor* ("ingestor").

Especificamente quanto à metodologia empregada na construção da *FrameNet*, ressalta-se que essa é composta pelas seguintes etapas manuais de especificação das informações semânticas: (i) geração inicial de *frames* e *elementos-frame*; (ii) identificação dos itens lexicais pertencentes a cada frame; (iii) extração, de córpus, das frases-exemplo e (iv) anotação das frases-exemplo com os *elementos-frame*. A principal desvantagem desse tipo de abordagem pauta-se no fato de que é preciso especificar previamente, no caso da *FrameNet*, os *frames* e *elementos-frame* de cada item lexical para, posteriormente, anotar as frases-exemplo. As tarefas (i-iv) demandam tempo e equipe de pesquisadores especializados.

Diante desse tipo de fator complicador, algumas pesquisas têm investigado a determinação ou identificação automática [Brent 1991; Resnik 1992; Grishman e Sterling 1992; Manning 1993; Framis 1994; Briscoe e Carroll, 1997] e semi-automática [Green et al. 2004; Gomez 2004] das estruturas argumentais dos itens lexicais, em especial, dos verbos. Tais abordagens, na maioria das vezes, baseiam-se em *parsers* (isto é, analisadores sintáticos automáticos) e/ou dicionários de *subcategorização* (ou padrões de realização sintagmática dos argumentos) para identificar os argumentos de um item predicador em uma sentença ou assumem como conhecidos o número, o tipo e a ordem de seus argumentos. O principal objetivo dessas pesquisas é identificar os itens lexicais mais prováveis para serem argumentos dos predicadores ou realizar operações de generalização sobre as estruturas aprendidas [Grishman e Sterling 1994; Framis 1994; Gomez 2004],

calculando a similaridade entre argumentos de estruturas semelhantes ou usando repositórios de informação lexical como a *WordNet* [Fellbaum 1998].

Neste trabalho, em especial, propõe-se uma abordagem semi-automática baseada em córpus para a identificação da *estrutura de argumentos* dos adjetivos V_1, em Padn, do português do Brasil. Mais especificamente, objetiva-se identificar os itens lexicais que podem ocorrer como As desses adjetivos e os conceitos subjacentes a esses As (e não os papéis temáticos ou traços semânticos). Por meio da especificação dos As em termos de conceitos, é possível obter estruturas generalizadas. Para essa generalização, utiliza-se a classificação hierárquica dos conceitos nominais da WordNet [Fellbaum 1998]. Como a Wordnet.Br [Dias-da-Silva 2002], em fase de desenvolvimento, ainda não armazena a organização hierárquica dos conceitos, decidiu-se utilizar a WordNet do inglês americano. Para isso, foi necessário mapear os itens do português (os As) para o inglês. Esse mapeamento foi feito por meio de um dicionário bilíngüe Português-UNL [Dias-da-Silva et al. 1998] (cf. 3.2.2).

A seguir, são descritas (i) as etapas que compõem a abordagem ora proposta e (ii) os recursos (e ferramenta) lingüístico-computacionais utilizados em cada etapa.

3. Da Abordagem Semi-automática Baseada em Córpus

3.1. Da preparação dos dados

3.1.1. Da seleção dos adjetivos e compilação das ocorrências Adj(N) e (N)Adj

Com base nos trabalhos de Borba (1996) e Neves (2000), foram selecionados dois adjetivos que são em princípio QLs e V1: “bom” e “novo”. Como mencionado, o objetivo é o de identificar a estrutura de argumentos desses adjetivos em Padn, pospostos e antepostos. Para tanto, foram compiladas as ocorrências dos pares Adj(N) e (N)Adj (para os dois adjetivos selecionados) no córpus *Mac-Morpho* [Aluísio et al. 2004] por meio do pacote de ferramentas *WordSmith Tools* (versão 3) [Scott 1999], mais especificamente, da ferramenta *Concord*.

O *Mac-Morpho*, integrante do portal *Lacio-Web*[6], é um corpus jornalístico de 1,1 milhão de palavras do português do Brasil, anotado automaticamente e revisado manualmente. Segundo suas características, pode-se dizer que o *Mac-Morpho* é um córpus anotado, fechado, de treinamento e teste (para ferramentas de PLN) e de suporte para pesquisas lingüísticas e lingüístico-computacionais. A escolha do córpus Mac-Morpho pautou-se basicamente no fato de que, por meio das etiquetas morfossintáticas, é possível identificar ocorrências de padrões como Adj(N) e (N)Adj em textos reais.

A ferramenta *Concord*, que integra o pacote de ferramentas de análise de córpus *WordSmith Tools*, é um concordanceador, ou seja, uma ferramenta que produz concordâncias ou listagens das ocorrências de um item específico (chamado palavra de busca ou nódulo) acompanhado do texto ao seu redor (o co-texto). O *Concord* produz concordâncias do tipo “Key Word in Context” (KWIC) (“palavra-chave no contexto”). Nesse tipo de concordância, a palavra de busca aparece

[6] Mais informações no endereço http://www.nilc.icmc.usp.br/lacioweb/

centralizada e ladeada por porções contínuas do texto de origem. Na Figura 2, ilustram-se as ocorrências compiladas do córpus pelo *Concord*.

```
1                Um_ART bom_ADJ exemplo_N de_PREP|+ isso_PROS
2    reagiu_V com_PREP bom_ADJ humor_N a_PREP|+ o_ART
3    refletindo_V o_ART bom_ADJ desempenho_N de_PREP|+ a_ART
4     de_PREP um_ART bom_ADJ pensador_N
```

Figura 2. Exemplos de ocorrências do par "bom(N)".

Na Tabela 1, apresenta-se o número de ocorrências do par Adj-N para cada adjetivo. Ressalta-se que o número de ocorrências do par Adj-N diz respeito apenas àquelas ocorrências em que o adjetivo V1 aparece em esquema adnominal, posposto e anteposto.

Tabela 1. Número de ocorrências dos pares Adj(N) e (N)Adj.

Adjetivo	Estrutura P(A1)	No. ocorrências
bom	Adj(N) (bom/bons/boa/boas-N)	417
	(N)Adj (N-bom/bons/boa/boas)	20
novo	Adj(N) (novo/novos/nova/novas-N)	1063
	(N)Adj (N-novo/novos/nova/novas)	101

3.1.2. Da limpeza manual das ocorrências

As concordâncias apresentadas na Figura 2 passaram por um processo de "limpeza" que consistiu na exclusão manual (i) das porções de co-textos consideradas excedentes e (ii) das etiquetas dos adjetivos e nomes. Ao final desse processo, foram obtidos apenas os pares Adj(N) e (N)Adj, como os ilustrados na Tabela 2.

Tabela 2. Exemplos dos padrões Adj(N) e (N)Adj "limpos".

Adjetivo	Estrutura P(A1)	Estrutura "limpa"
bom	Adj(N)	bom(exemplo) boa(performance)
	(N)Adj	(plano)bom (eficiência)boa
novo	Adj(N)	novo(modelo) novo(centros)
	(N)Adj	(capítulo)novo (doença)nova

3.1.3. Da redução automática dos predicadores e argumentos à forma canônica

Após a obtenção dos dados exemplificados na Tabela 2, prossegui-se com a redução automática dos adjetivos e dos nomes às suas formas canônicas (masculino e singular). Essa redução foi necessária porque o Dicionário Português-UNL contém apenas as formas canônicas dos itens do português (cf. 3.2.2). Por exemplo, o adjetivo do par *boa*(performance) (cf. Tabela 2) foi reduzido ou transformado para ***bom***(performance) e o nome do par *novo*(centros) (cf. Tabela 2) foi reduzido para *novo*(**centro**).

O processo de redução à forma canônica foi feito por meio do léxico do *ReGra* (Revisor Gramatical da Línguas Portuguesa) [Nunes 1996]. Esse léxico foi desenvolvido pelo NILC[7] e, atualmente, é composto de mais de um milhão e meio de formas lexicais válidas para o português contemporâneo, acrescidas de informações morfossintáticas (classe gramatical, número, gênero, grau e regência) e de formas canônicas [Rino et al. 2002].

3.1.4. Do mapeamento automático dos itens do português (os As) para o inglês

Após a redução dos adjetivos e nomes à forma canônica, seguiu-se com o mapeamento automático dos nomes da língua portuguesa para a inglesa. Esse mapeamento para a língua inglesa, como mencionado, permitiu, na etapa seguinte, recuperar os conceitos subjacentes aos argumentos (nomes) dos adjetivos na WordNet americana.

Para o mapeamento, em especial, foi utilizado o Dicionário Português-UNL [Dias-da-Silva et al. 1998] (doravante Dic Port-UNL), criado durante o desenvolvimento do módulo de decodificação da UNL[8] para o português do Brasil [Oliveira et al., 2001]. O Dic Port-UNL faz a correspondência entre os itens lexicais da língua portuguesa (formas canônicas) e os itens lexicais da interlíngua UNL, que são especificados na língua inglesa [Uchida 2000]. Um item lexical ou lexema da língua portuguesa pode remeter a vários lexemas UNL, os quais correspondem às diferentes acepções do item. Dessa forma, ressalta-se que todos os lexemas da UNL vinculados aos argumentos (isto, aos nomes) dos adjetivos foram selecionados, pois, assim, garante-se que todos os sentidos dos mesmos sejam considerados, já que o sentido apropriado do argumento é desconhecido. Os itens lexicais da interlíngua UNL são denominados **U**niversal **W**ords (UWs). Após o mapeamento automático dos argumentos em UWs, obtêm-se estruturas P(UW). Na Figura 3, ilustram-se essas estruturas.

P (UW)
bom (example)
bom (instance)
bom (cross)
bom (disposition)

Figure 3. Exemplos de estruturas P(UW)

3.1.5. Da recuperação automática do hiperônimo corresponde aos As na WordNet

A partir das estruturas P(UW), já é possível proceder ao aprendizado automático dos argumentos dos adjetivos. Para que o processo de aprendizado seja eficiente e reflita mais fielmente o conhecimento humano, deseja-se, ainda, generalizar as UWs das estruturas. Por exemplo, em vez de estruturas especializadas como *bom*(car) e *bom*(bike), ter-se-ia a estrutura *bom*(motor_vehicle), que abrange as anteriores (pois

[7] Mais informações no endereço www.nilc.icmc.usp.br.
[8] A UNL (*Universal Networking Language*) é uma linguagem artificial para representação semântica do conteúdo ideacional dos enunciados lingüísticos [Uchida 2001].

"vehicle" é a superclasse de "car" e "bike"), evitando-se, dessa forma, redundância na especificação das estruturas argumentais dos adjetivos.

Associou-se automaticamente a cada UW seu hiperônimo (isto é, conceito mais genérico) imediato da WordNet para seu primeiro sentido. Diz-se hiperônimo "imediato" (mais específico) pelo fato de a WordNet organizar hierarquicamente os conceitos nominais em vários níveis, do mais específico ao mais geral. Optou-se pela busca do primeiro sentido (em vez de todos) porque a ambigüidade de sentido já está representada pelas UWs (vale lembrar que todas as UWs possíveis para um item lexical são consideradas). Na Figura 4, ilustra-se como essa informação está armazenada na WordNet.

Bike
Sense 1
motorcycle, bike -- (a motor vehicle with two wheels and a strong frame)
⇨ motor vehicle, automotive vehicle -(a self-ropelled wheeled vehicle that does not run on rails)

Figure 4. Exemplo de entrada lexical no formato *wordnet*

Além do hiperônimo, mantém-se também a própria UW nas estruturas, que assumem, agora, a forma P(UW,H), em que H é o hiperônimo em questão. Decidiu-se por manter a UW por não se saber, *a priori*, o nível de generalidade mais apropriado para os argumentos dos adjetivos, por exemplo, se "motor_vehicle" (mais geral) é mais apropriado do que "car" (mais específico) como argumento de "bom". Espera-se, dessa forma, que, durante o aprendizado, isso seja determinado automaticamente. Na Figura 5, ilustram-se as estruturas P(UW,H) produzidas.

Adj [UW, H]
bom [example, information]
bom [instance, happening]
bom [cross, structure]
bom [disposition, nature]

Figure 5. Exemplo das estruturas P(UW,H).

3.2. Do Aprendizado das Estruturas P(A_1)

A partir das estruturas P(UW,H), prosseguiu-se com o aprendizado dos argumentos propriamente ditos, lexicalizados ou generalizados.

O aprendizado remete, basicamente, ao cálculo de freqüência dos possíveis argumentos (UWs ou Hs) de um adjetivo. Por exemplo, tendo-se as estruturas hipotéticas P(UW1,H1) e P(UW2,H1), aprender-se-ia que o adjetivo em questão co-ocorre (i) uma única vez com UW1 e UW2 e (ii) duas vezes com H1. Neste caso, ter-se-ia que o argumento generalizado H1 é mais provável do que UW1 e UW2. É interessante ressaltar que a decisão sobre o argumento do adjetivo ser lexicalizado ou generalizado é diretamente determinada pela freqüência de ocorrência deste na forma lexicalizada ou generalizada.

Na Tabela 3, mostram-se as estruturas ranqueadas por freqüência aprendidas para os adjetivos "bom" e "novo", em Padn (pospotos e antepostos). Por exemplo,

para o adjetivo “bom” anteposto, tem-se na primeira linha da tabela que o argumento mais provável (cujo número de ocorrências é 9 no córpus e cuja freqüência relativa é de 0,0082) é “message”.

Tabela 3. Estruturas de argumentos aprendidas.

Adj	No. oc.	Freq.	$P(A_1) \rightarrow$ Adj(N)	No. oc.	Freq.	$P(A_1) \rightarrow$ (N)Adj
Bom	9	0,0082	bom(message)	4	0,0238	(activity)bom
	8	0,0072	bom(feeling)	3	0,0178	(plan)bom
	8	0,0072	bom(activity)	3	0,0178	(time_period)bom
	8	0,0072	bom(person)	3	0,0178	(work)bom
	8	0,0072	bom(line)	2	0,0119	(object)bom
	7	0,0063	bom(time_period)	2	0,0119	(plain)bom
	6	0,0054	bom(belief)	2	0,0119	(line)bom
	6	0,0054	bom(happening)	2	0,0119	(time)bom
	5	0,0045	bom(state)	2	0,0119	(animal_group)bom
	5	0,0045	bom(work)	2	0,0119	(disturbance)bom
Novo	16	0,0073	novo(activity)	5	0,0091	(time_period)novo
	13	0,0059	novo(line)	5	0,0091	(activity)novo
	10	0,0046	novo(property)	4	0,0073	(property)novo
	9	0,0041	novo(happening)	4	0,0073	(message)novo
	9	0,0041	novo(person)	3	0,0055	(disorder)novo
	8	0,0036	novo(message)	3	0,0055	(feeling)novo
	8	0,0036	novo(device)	3	0,0055	(point)novo
	7	0,0032	novo(work)	3	0,0055	(component)novo
	7	0,0032	novo(area)	3	0,0055	(clothing)novo
	7	0,0032	novo(idea)	3	0,0055	(case)novo

4. Da Discussão

A análise dos resultados obtidos evidencia que, por meio da abordagem proposta, foi possível identificar estruturas argumentais lexicalizadas e generalizadas. A importância de se identificar argumentos lexicalizados pauta-se no fato de que nem sempre a generalização é mais apropriada. Isso ocorre tipicamente em expressões da língua utilizadas em sentido figurado, para as quais não faz sentido realizar generalizações, por exemplo, a expressão “trem bom”, na qual o item lexical trem não remete necessariamente à generalização motor_vehicle. Outros casos consistem em itens que quase sempre ocorrem com o mesmo adjetivo, não variando e, portanto, não sendo necessário que se faça generalizações, por exemplo, a expressão “água ardente”.

Os dados obtidos também evidenciam uma característica típica dos adjetivos quando pospostos ou antepostos ao nome no interior do sintagma nominal. Em geral, quando pospostos, seus argumentos lexicalizam conceitos concretos, por exemplo, entre os 10 argumentos mais freqüentes aprendidos para o adjetivo bom, 6 o fazem (work, object, plain, line, animal_group e disturbance). Quando antepostos, os argumentos selecionados pelos adjetivos lexicalizam conceitos abstratos, por exemplo, entre os 10 argumentos mais freqüentes aprendidos para o adjetivo bom, 7 o fazem (message, feeling, activity, time_period, belief, happening e state).

5. Das Considerações Finais

A abordagem apresentada neste trabalho caracteriza-se por: (i) semi-automatizar o processo de identificação da estrutura de argumentos dos adjetivos V_1; (ii) gerar estruturas ordenadas de acordo com suas probabilidades, sendo possível, assim, identificar que determinada estrutura é mais provável que outra.

As informações sobre os adjetivos geradas neste trabalho podem ser empregadas no enriquecimento de léxicos para fins do PLN. Mais precisamente, de léxicos para sistemas de (i) geração de textos (p.ex.: sumarizadores, tradutores, entre outros), (ii) analisadores semânticos e sintáticos. A inserção desse tipo de informação em léxicos computacionais contribui para a verificação e garantia da boa-formação semântica de sintagmas nominais formados por N-Adj e Adj-N. Munido de informações sobre as estruturas de argumentos dos adjetivos V1, um sistema de PLN provavelmente consideraria o trecho "incolores *idéias* verdes" (da sentença: *Incolores* idéias *verdes* dormem furiosamente) má formado semanticamente, posto que os adjetivos *incolor* e *verde* não selecionam argumentos do tipo "idéia", ou seja, argumentos que lexicalizam conceitos do tipo [MENTAL OBJECT].

A limitação dessa abordagem consiste basicamente no fato de que somente é possível recuperar com precisão suficiente ocorrências no córpus de adjetivos em Padn. Para as ocorrências dos adjetivos valenciais em posição predicativa, não há padrões claros de como estes co-ocorrem com as palavras das sentenças. O uso de um analisador sintático poderia auxiliar nesta tarefa, entretanto, não há para o português do Brasil um analisador robusto o suficiente e livre para uso.

Como trabalho futuro, pensa-se em estender o aprendizado para adjetivos que projetam dois argumentos, por exemplo, "sensível" (X é sensível a Y). A busca pelas ocorrências desses adjetivos no corpus pode ser feita por meio do padrão N+ADJ+PREP.

Uma potencialidade real deste trabalho é a de se popular o léxico computacional do ReGra [Nunes et al. 1996] com informações sobre as estruturas argumentais dos adjetivos, permitindo que este sistema realize sua tarefa de forma mais informada.

Referências Bibliográficas

Aluísio, S.; Pinheiro, G.M.; Manfrim, A.M.P; Oliveira, L.H.M. de; GENOVES Jr., L.C. e Tagnin, S.E.O. (2004). The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation* (LREC). Lisboa, Portugal, p. 1779-1782.

Baker, C.F.; Fillmore, C.J. e Lowe, J.B. (1998). The Berkeley FrameNet project. In *Proceedings of COLING/ACL*, p. 86-90, Montreal.

Borba, F.S. (1996). *Uma gramática de valências para o português*. São Paulo: Editora Ática.

Briscoe, T. e Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ANLP Conference*, p. 356-363, Washington, D.C.

Calzolari, N.; Lenci, A.; Zampolli, A. (2001). International Standards for Multilingual Resource Sharing: The ISLE Computational Lexicon Working Group. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics* (ACL)/ Workshop on Sharing Tools and Resources. Toulouse, France.

Dias-da-Silva, B.C.; Sossolote, C.; Zavaglia, C.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N.; Aluísio, S.M. (1998). The design of the Brazilian Portuguese machine tractable dictionary for an interlíngua sentence generator. In *III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada.* PUCRS, Porto Alegre.

Dias-da-Silva, B.C.; Oliveira, M.F.; Moraes, H.R. (2002). Groundwork for the development of the Brazilian Portuguese Wordnet. *Advances in natural language processing*. Berlin: Springer-Verlag, p.189-196.

Fellbaum, C. (Ed.). (1998). *Wordnet: an electronic lexical database*. Cambridge: The MIT Press.

Framis, F.R. (1994). An experiment on learning appropriate selection restrictions from a parsed corpus. In *Proceedings of the International Conference on Computational Linguistics*, Japan.

Gomez, F. (2004). Building Verb Predicates: A Computational View. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* (ACL), p. 359-366, Barcelona, Spain.

Green, R.; Dorr, B.J. e Resnik, P. (2004). Inducing Frame Semantic Verb Classes from WordNet and LDOCE. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* (ACL), p. 375-382, Barcelona, Spain.

Grimshaw J. (1992). *Argument structure.* Cambridge: The MIT Press.

Grishman, R. e Sterling, J. (1992). Acquisition of selectional patterns. In *Proceedings of the International Conference on Computational Linguistics*, p. 658-664, Nantes, France.

Grishman, R. e Sterling, J. (1994). Generalizing Automatically Generated Selectional Patterns. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Handke, J. (1995). *The structure of the Lexicon: human versus machine*. Berlin: Mouton de Gruyter.

Levin, B. (1993). *Towards a lexical organization of English verbs*. Chicago University Press, Chicago.

Manning, C. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (ACL), p. 235-242, Columbus, Ohio.

Neves, M.H.M. (2000). *Gramática de usos do português*. São Paulo: Editora UNESP.

Nunes, M.G.V. et alli. A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos. *Relatórios Técnicos do ICMC-USP*, 42. Setembro 1996, 36p.

Rino, L.H.M.; Di Felippo, A.; Pinheiro, G.M.; Martins, R.T.; Filié, V.M.; Hasegawa, R.; Nunes, M.G.V. (2002). Aspectos da construção de um revisor gramatical automático para o português. *Estudos Lingüísticos*, SP, v. 31, ISSN 1413 0939. 1 CD-ROM.

Palmer, M. (1999). Multilingual resources – Chapter 1. In: Hovy, E. et al. (Eds.). *Multilingual Information Management: Current Levels and Future Abilities*. http://www.cs.cmu.edu/~ref/mlim/, Abril de 2005.

Pardo, T.A.S.; Marcu, D.; Nunes, M.G.V. (2005). Um Modelo Estatístico Gerativo para o Aprendizado Não Supervisionado das Estruturas Argumentais dos Verbos. In *Anais do III Workshop de Tecnologia da Informação e da Linguagem Humana* (*TIL*), 2005.

Raposo, E.P.(1992). *Teoria da gramática: a faculdade da linguagem*. Lisboa: Caminho.

Resnik, P. (1992). Wordnet and distributional analysis: a class-based approach to lexical discovery. In Proceedings of AAAI Workshop on Statistical Methods in NLP.

Sanfilippo, A.; Calzolari, N., Ananiadous, S., Gaizauskas, R., Saint-Dizier, P. e Vossen, P. (Eds.). (1999). Preliminary Recommendations on Lexical Semantics Encoding. *Relatório EAGLES LE3-4244*.

Scott, M. (1999). *Wordsmith Tools version 3*. Oxford: Oxford University Press. ISBN 0-19-459289-8.

Uchida, H. (2000). *Universal Networking Language: An Electronic Language for Communication, Understanding and Collaboration*. UNL Center, IAS/UNU, Tokyo.

Viegas, E. e Raskin, V. (1998). Computational semantic lexicon acquisition: methodology and guidelines. *Relat. Téc. CRL/NMSU*, MCCS-315.