Inclusão de Informação Semântica dos Adjetivos na Base da Rede Wordnet para o Português do Brasil

Ariani Di Felippo[1], Bento Carlos Dias-da-Silva[1]

[1] Faculdade de Ciências e Letras, Universidade Estadual Paulista (UNESP-Ar)
Centro de Estudos Lingüísticos e Computacionais da Linguagem (CELiC)
Rodovia Araraquara-Jau, Km 1, Caixa Postal 174,
14.800-901, Araraquara, São Paulo, Brazil.
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Av. do Trabalhador São-Carlense, 400, Caixa Postal 668,
13560-970, São Carlos, São Paulo, Brazil.
{arianidf@uol.com.br, bento@fclar.unesp.br}

Abstract. This paper proposes the formal inclusion of the valence of the adjectives in Wordnet.Br lexical database. To achieve this purpose, we present an overall of the actual Wordnet.Br structure and describe the systematization of the adjectival valence information. This proposal of extension attempts to refine Wordnet.Br lexical database, which is an important linguistic resource for the development of Brazilian Portuguese processing applications.

Keywords. Adjectives; Valence; Argument Structure; Wordnet.Br.

Resumo. Neste trabalho, propomos a inclusão formal da valência dos adjetivos na base da rede Wordnet.Br. Para tanto, apresentamos a estrutura atual dessa base e descrevemos a sistematização do conjunto de informações sobre a valência dos adjetivos. Essa proposta de extensão busca refinar a base da rede Wordnet.Br, que é um importante recurso lingüístico para o desenvolvimento de aplicações de processamento automático do Português do Brasil.

Palavras-chave. Adjetivos; Valência; Estrutura de Argumentos; Wordnet.Br.

1 Introdução

Com o desenvolvimento da rede WordNet[1] para a língua inglesa pela Universidade de Princeton [1] (EUA), vários países construíram ou estão construindo suas próprias *wordnets*, dada a importância desse tipo de base lexical na compilação de parcelas de léxicos para o desenvolvimento de diversos sistemas de processamento automático de línguas naturais (PLN).

A rede WordNet é, na verdade, uma base relacional, em que unidades lexicais do inglês, pertencentes às categorias dos substantivos, verbos, adjetivos e advérbios,

[1] O nome da rede americana é grafado com "N" maiúsculo para diferenciá-la das demais, caracterizando-a, como diz Fellbaum [1], como "a mãe de todas as Wordnets".

estão organizadas em termos de conjuntos de sinônimos (isto é, os *synsets*), os quais expressam conceitos lexicalizados. Tais conjuntos relacionam-se entre si em função das cinco relações de sentido: antonímia, hiponímia, meronímia, acarretamento e causa. Além disso, a rede WordNet registra informações periféricas, associadas a cada sentido armazenado. São elas: as frases-exemplo e as glosas e (isto é, definições informais) [1].

A base da rede *wordnet* brasileira (doravante, Wordnet.Br), que está sendo desenvolvida a partir do aplicativo *Thesaurus Eletrônico para o Português do Brasil – TeP*, apresenta, atualmente, um total de 17.416 substantivos, 15.073 adjetivos, 11.078 verbos e 1.139 advérbios, estruturados em função das relações de *sinonímia* e *antonímia* [2, 3].

A importância desse tipo de base para os pesquisadores do PLN deve-se à possibilidade, por exemplo, de geração de parcelas de léxicos especiais (munidos de conhecimento léxico-semântico) imprescindíveis para o desenvolvimento de diversos sistemas de PLN, como: sistemas de tradução automática, de sumarização automática, de recuperação de informação, entre outros [4, 5, 6 e 7].

Ao usuário da língua portuguesa, por sua vez, a base da Wordnet.Br, acoplada a ferramentas computacionais de auxílio à escrita, poderá oferecer a opção de seleção *on line* de palavras sinônimas e antônimas que, por motivos de estilo, precisão, adequação comunicativa, correção ou aprendizagem, o usuário queira substituir [8].

Na fase atual de desenvolvimento, os lingüistas têm realizado[2]: (a) a análise da consistência semântica dos *synsets* (ou conjunto de sinônimos); (b) a coleta e seleção das frases-exemplo, extraídas de córpus.

Neste trabalho, propomos a inserção, na base da Wordnet.Br, das seguintes informações: (i) o *tipo valencial* e a *valência semântica* dos adjetivos. Vale lembrar que os *synsets* apontam para um conceito ou sentido. Sendo predicadores, os conceitos apontados pelos adjetivos qualificadores podem ser "traduzidos" em termos de uma estrutura de argumentos, refinando, assim, os dados referentes a esses adjetivos dessa base. A escolha dos adjetivos justifica-se pelo fato de que, apesar de serem predicadores da língua por excelência (ao lado dos verbos), eles têm recebido atenção menor, por parte dos pesquisadores do PLN, que classes como as dos verbos e substantivos [9].

Nas seções subseqüentes, delineamos a estrutura atual da base da Wordner.Br e sistematizamos as informações referentes à valência dos adjetivos do Português. Por fim, esquematizamos a extensão resultante do acréscimo do *tipo valencial* e da *valência semântica* dos adjetivos.

2 Da estrutura atual da Wordnet.Br

A elaboração da base da Wordnet.Br teve como ponto de partida o TeP, que foi elaborado segundo os princípios da WordNet. Da metodologia proposta por Miller e Fellbaum [10], foram utilizadas três noções básicas no desenvolvimento do Tep [3].

[2] Projeto financiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Processo: no 552057/01-0.

(i) o *método diferencial*: pressupõe o princípio de ativação de conceitos por meio de um conjunto de formas lexicais relacionadas pela sinonímia, o que elimina a necessidade de especificação do valor semântico para o sentido de uma entrada lexical;
(ii) a noção constitutiva de *synset*: conjunto de sinônimos;
(iii) a noção de *matriz lexical*: postula uma correspondência biunívoca entre sentido e *synset*.

Com essa metodologia, a relação de sinonímia passa a ser representada formalmente pela relação lógica de pertença (x é sinônimo de y $\leftrightarrow$ x $\wedge$ y $\in$ A, em que A é um *synset*). A antonímia, por sua vez, é representada por uma relação entre conjuntos (x é antônimo de y $\leftrightarrow$ x $\in$ A e y $\in$ B, A e B são *synsets* e A e B estão relacionados pela relação de antonímia). O esquema presente na Figura 1 ilustra a estrutura da base do TeP e, conseqüentemente, da Wordent.Br:

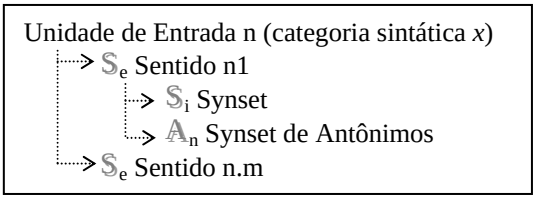

Fig. 1. Estrutura típica de um verbete da base da rede Wordnet.Br

Nesse esquema, "n" é o número de identificação da unidade de entrada, "x" é uma variável que representa uma das quatro categorias gramaticais (substantivo, verbo, adjetivo ou advérbio), "Se" é uma sigla para sentido, "n.1...n.m" são números que identificam cada sentido da unidade de entrada n, "Si" é uma sigla para *synset* e, por fim, "An" é uma sigla para antônimo.

3 Dos adjetivos no projeto Wordnet.Br

Para a compilação dos conjuntos de sinônimos e antônimos de adjetivos do TeP, partimos do princípio de que os adjetivos do português, assim como os do inglês e do espanhol, podem ser divididos em duas classes semânticas: os qualificadores (QLs) e os classificadores (CLs) [11, 12, 13 e 14].

Os qualificadores indicam o valor de uma propriedade ou atributo do substantivo com o qual se liga. Dessa forma, dizer "X QL" ou "X é QL" pressupõe um atributo A, tal que A(x) = QL. Por exemplo: dizer "torre alta" ou "a torre é alta", pressupõe um atributo ALTURA, tal que ALTURA(torre) = alta [15]. Já os classificadores colocam o substantivo com o qual ocorrem numa subclasse, nomeando-a, p.ex.: *cambial* em a "reforma cambial". Observe-se que a paráfrase "do câmbio" sinaliza que *cambial* liga a entidade "reforma" a outra, exterior a ela: o "câmbio" [12 e 16]. Enquanto os adjetivos QLs expressam "qualidades" ou "valores de atributos" dos

substantivos, os CLs são comumente definidos, em obras lexicográficas, por meio de paráfrases como “de ou pertencente/ relativo a X” [10].

Para a inclusão das informações referentes à valência dos adjetivos, partimos do princípio de que, do ponto de vista sintático-semântico, os adjetivos qualificadores são verdadeiros predicadores (Ps), tanto em posição adnominal (1a) como em posição predicativa (1b), e os classificadores são, na verdade, ou argumentos dos substantivos (2a) ou meros circunstanciais (2b).

(1) a. O rosto dele era *pálido* e os seus olhos eram muito *tristes*.
b. Olha para o rapaz *pálido* que se encontra a sua frente.
(2) a. Ao longo da mesa *presidencial* **[do presidente]**, havia um vaivém continuado.
b. O jogador *profissional* **[joga por profissão]** deve ter um contrato assinado.

Sendo P, o adjetivo QL tem a propriedade de poder ligar-se a um certo número de elementos exigidos pela sua semântica, os *argumentos* (As). A essa propriedade é dada a denominação *valência* [12]. Em outras palavras, um P designa um “estado-de-coisas (isto é, algo que pode ocorrer em algum mundo real ou mental), o qual é “projetado” na forma da expressão por meio das relações que se estabelecem entre P e seus As [17]. A valência de um P pode ser descrita em três níveis: valência lógico-semântica; sintática e semântica. Neste trabalho, focalizamos os níveis lógico-semântico e semântico.

4 Da sistematização das informações sobre os adjetivos

4.1 Da valência lógico-semântica

O nível lógico-semântico diz respeito ao número de argumentos projetados pela semântica de um P. No caso dos adjetivos, há duas interpretações possíveis. Na primeira, consideram-se apenas os constituintes diretamente dependentes dos adjetivos; na segunda, considera-se como argumento adjetival o constituinte em função de sujeito [18]. Neste trabalho, optou-se pela segunda interpretação.

Dessa forma, os adjetivos Ps do Português podem ser de quatro tipos valenciais, como descrito na Tabela 1.

Table 1. Tipologia da valência lógico-semântica dos adjetivos

Tipologia	Descrição e Exemplificação
Valência 1 (V1)	(3) João (A1) era *bonito*.
Valência 2 (V2)	(4) O homem (A1) era *descendente* de portugueses (A2).
Valência 3 (V3)	(5) O rapaz (A1) era *doador* de órgão (A2) para transplantes (A3)
Valência 4 (V4)	(6) A carga (A2) era *transportável* do navio (A3) para o cais (A4) pelos guindastes (A1)

4.2 Da valência semântica dos adjetivos

O nível semântico diz respeito à combinatória de traços semânticos entre P e seus argumentos e às relações semânticas que se instauram a partir dessa combinatória.

Ressaltamos que o "sentido" ou "valor semântico" de um adjetivo P (ou valencial) é determinado pela combinatória estabelecida entre ele e os traços semânticos dos argumentos com os quais aceita ocorrer [9]. Por exemplo:

(7) a. O garoto (A1) é *esperto*.
b. A água (A1) estava *esperta*.
(8) a. O valor (A1) *deduzido* da conta-corrente (A2) foi alto.
b. As conclusões (A1), *deduzidas* das palavras do presidente (A2), foram otimistas.

Nas sentenças em (7), o adjetivo valencial *esperto* projeta apenas um argumento, o A1, já que a semântica desse item implica "*x* é *esperto*". Na sentença em (7a), o adjetivo tem sentido de "astuto, finório", o qual é gerado pela combinatória de *esperto* com um argumento de traço semântico *+humano* ("o garoto"). Já na sentença em (7b) *esperto* tem sentido de "quase quente", o qual é gerado pela combinatória desse adjetivo com um argumento de tipo semântico *–animado* ("a água"). Na relação P(A), os argumentos têm funções semânticas específicas. No caso de *esperto*, em (7), o A1 é a entidade sobre a qual se verifica uma situação.

A valência semântica dos Ps tem sido amplamente representada por meio de um construto formal denominado *estrutura de argumentos*. Essa representação é composta por *papéis temáticos* (rótulos abstratos que representam as funções semânticas dos argumentos) e *restrições selecionais* (rótulos abstratos que restringem a semântica dos argumentos) [19, 20 e 21]. Por exemplo:

(9) a. **esperto1**: <(A1)Tema[+*hum*]>
b. **esperto2**: <(A1)Tema[-*anim*]>

Em (9), ilustramos a estrutura de argumentos de *esperto* em (7a) e (7b). Em (9a), o A1 de *esperto* está associado ao papel temático Tema e à restrição selecional *+humano*. Em (9b), o A1 está associado ao mesmo papel temático, mas a restrição selcional é *–animado*. Dessa forma, a diferença de sentido observada em (7) fica representada por meio das *restrições selecionais* impostas pelo adjetivo *esperto* ao argumento que projeta.

Nas sentenças em (8), *deduzido* projeto 2 argumentos, A1 e A2, já que sua a semântica implica "*x deduzido* de *y*". Em (8a), *deduzido* tem sentido de "subtração/retirada" e, em (8b), de "inferência". Tanto em (8a) como em (8b), as relações semânticas estabelecidas entre esse *deduzido* e seus dois argumentos são idênticas. No que diz respeito à representação dessas relações, o A1 associa-se ao papel temático Tema e o A2, ao de Origem. A distinção entre os sentidos de "subtração" e "inferência", então, pode ser traduzida em termos das restrições de seleção que o *deduzido* impõe ao A2. Enquanto que em (8a) o A2 "conta-corrente" conforma-se ao traço semântico [+(origem da) *subtração*], em (8b) o A2 "palavras"

conforma-se ao traço semântico [+(origem da) *inferência*]. Assim, os sentidos "subtração" e "inferência" de *deduzido* podem ser representados formalmente pelas estruturas de argumentos ilustradas em (10).

(10) a. **deduzido1**: <(A1)Ob, (A2)Or[*subtração*]>
b. **deduzido2**: <(A1)Ob, (A2)Or[*inferência*]>

Observamos, então, que o(s) sentido(s) dos adjetivos predicadores ou valenciais (os QLs) são gerados pela combinatória destes com seus argumentos e que, formalmente, pode(m) ser representado(s) em *estruturas de argumentos*. Assim, os sentidos dos adjetivos qualificadores, indiretamente armazenados na base da Wordnet.Br por meio dos *synsets*, podem ser "traduzidos" em estruturas de argumentos. Em outras palavras, essas estruturas seriam uma espécie de "comentário", no sentido computacional desse termo, dos sentidos. É nessa direção que apresentamos, a seguir, a inclusão das informações referentes à valência dos adjetivos qualificadores da Wordnet.Br.

5 Da extensão da base da rede Wordnet.Br

Paralelamente ao esquema apresentada na Figura 1, propomos que a base da rede Wordnet.Br, além de conter os conjuntos de sinônimos e antônimos, deverá conter informações sobre a valência dos adjetivos. A classificação esboçada em (4.1) permite estender a informação original (Fig.1) relativa aos adjetivos QLs (os valenciais) com os seguintes tipos: (P) = {V1} + {V2} + {V3} + {V4}. Assim, propomos que aos QLs da base da Wordnet.Br sejam associadas informações sobre sua tipologia valencial ou valência lógico-semântica e valência semântica.

O esquema descrito na Figura 2 exemplifica essa extensão.

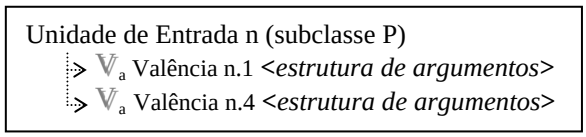

Fig. 2. Estrutura adicional para os verbetes da Wordnet.Br

Nesse esquema, "n" é o número de identificação da unidade de entrada; "P" indica que o adjetivo é o do tipo predicador; "Va" é uma sigla para valência, "n.1...n.4" indicam qual o subtipo valencial (V1, V2, V3 e V4), ao qual é associada a estrutura de argumentos (papéis temáticos + restrições selecionais) do adjetivo propriamente dito.

Salienta-se que indexações apropriadas deverão permitir, quando pertinente, o relacionamento entre as entradas estruturadas em termos das relações léxico-semânticas (sinônima e antonímia) e as entradas estruturadas em termos das informações semânticas aqui propostas. Mais especificamente, indexações deverão

permitir o relacionamento entre *valência* e *sentido*, este indiretamente armazenado na base da Wordnet.Br em função dos conjuntos de sinônimos. Na Figura 3, ilustramos essa possível indexação.

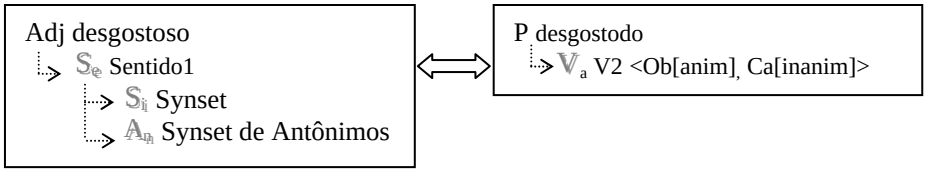

Fig. 3. Associação da estrutura atual à adicional dos verbetes da Wordnet.Br

Vale ressaltar, por fim, que, uma vez especificada a estrutura de argumentos que "traduz" o(s) sentido(s) de um adjetivo QL, poderemos generalizá-la para os demais membros do *synset*. Assim, os adjetivos {anojado; desagradado; descontente; dissaborido; dissaboroso; malcontente; penalizado; triste}, que compõem o *synset* associado ao Sentido1 de *desgostoso*, poderão herdar a valência semântica especificada na Figura 3.

6 Considerações finais

Neste artigo, apresentamos uma proposta de inclusão, na base da rede Wordnet.Br, das informações referências à valência dos adjetivos. Com essa proposta, objetivamos refinar os dados dos Wordnet.Brs adjetivos. Assim, com esse acréscimo, a manipulação da base da rede Wordnet.Br poderá gerar listas de unidade lexicais, para a compilação de léxicos monolíngües, que, além de fornecerem as relações léxico-semânticas que se instauram entre os adjetivos, fornecerão também sua tipologia valencial e valência semântica.

Na fase atual de desenvolvimento desta proposta, buscamos identificar na literatura pressupostos teórico-metodológicos pertinentes para a descrição da valência semântica dos adjetivos e estudamos a possibilidade de extensão desta proposta para as classes dos verbos e substantivos. Além disso, por meio de trabalho colaborativo entre lingüistas e cientistas da computação, buscamos desenvolver uma interface computacional gráfica para a montagem dos verbetes estruturados em função da valência.

Por fim, gostaríamos de ressaltar que, até o momento da feitura deste artigo, os autores não têm informações a respeito de proposta similar para a inclusão, nas redes Wordnets para as demais línguas, de informações referentes à valência dos itens lexicais. No entanto, contribuições são sempre bem-vindas.

Referências

1. Fellbaum, C. (ed.): Wordnet: an electronic lexical database. The MIT Press, Cambridge (1999)
2. Dias-da-Silva, B.C., Moraes, H.R.: A construção de um thesaurus eletrônico para o português do Brasil. Alfa, Vol.47, n.2, (2003) 101 – 115
3. Dias-da-Silva, B.C., Oliveira, M.F. e Moraes, H.R.: Groundwork for the development of the Brazilian Portuguese Wordnet. Advances in natural language processing. Berlin, Springer-Verlag (2002) 189-196
4. Briscoe, E.J., Boguraev, B. (eds).: Computational lexicography for natural language processing. Longman, London/New York (1989)
5. Saint-Dizier, P., Viega, E.: Computational lexical semantics. Cambridge University Press, Cambridge (1995)
6. Dias-da-Silva, B.C.: Bridging the gap between linguistic theory and natural language processing. In: Proceedings of the 16th international congress of linguistics. Elsevier Sciences n. 16, Oxford (1998) 1-10
7. Palmer, M.: Miltilingual resources – Chapter 1. In: Hovy, E., et al. (eds.). Linguistica Computazionale, Vol.14-15 (2001)
8. Ilari, R. , Geraldi, J.W.: Semântica. Ed. Ática, São Paulo (1985)
9. Pustejovsky, J.: The Generative Lexicon. Cambridge, MIT Press (1996)
10. Miller, G.A., Fellbaum, C.: Semantic networks of English. Cognition,Vol.41, n.1-3 (1991) 197-229
11. Quirk, R. et al.: A Comprehensive Grammar of the English Language. Longman, London (1991)
12. Borba, F.S.: Uma Gramática de Valências para o Português. Ed. Ática, São Paulo (1996)
13. Demonte, V.: Semántica composicional y gramatica: los adjetivos en la interficie léxico-sintaxis. Revista Española de Lingüística. 29, Vol 2 (1999) 283-316
14. Neves, M.H.M.: Gramática de Usos do Português. Editora UNESP, São Paulo (2000)
15. Gross, D., Fischer, U., Miller, A.: The organization of the adjectival meaning. Journal of Memory and Language. 28 (1995) 92-106
16. Basílio, M., Gamarski, L.: Adjetivos denominais no português falado. In: Castilho, A. T. (org.): Gramática do Português Falado, Vol. IV. UNICAMP, Campinas (2002) 629-650
17. Dik, S.C.: The Theory of Functional Grammar. Mouton de Gruyter, Berlin New York (1997)
18. Busse, W., Vilela, M.: Gramática de Valências. Almedina, Coimbra (1986)
19. Grimshaw, J.: Argument structure. The MIT Press, Cambridge (1992)
20. Fillmore, C.J.: The case for case. In: Bach, E., Harms, R. T. (eds.). Universals in Linguistic Theory. Holt, Rinehart and Winston, Inc. (1968) 1-88
21. Palmer, F.R.: Grammatical Roles and Relations. Cambridge University Press, Cambridge (1994)