Modelo lingüístico-computacional da estrutura valencial de adjetivos do português do Brasil

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

[1]Faculdade de Ciências e Letras – UNESP
Caixa Postal 174 – 14.800-901 – Araraquara – SP – Brasil
arianidf@uol.com.br, bento@fclar.unesp.br

Abstract. *In this paper an attempt is made to investigate the linguistic and computational representation issues of the valency adjectives. After a brief review of the properties of the valency adjectives and the "features structures" formalism, a machine-tractable lexical entry template of such adjective is presented.*

Keywords. *valency adjectives; formalism; lexical entry; natural language processing.*

Resumo. *Este artigo examina a questão da representação lingüístico-computacional dos adjetivos valenciais do português. Após uma breve revisão das principais propriedades dos adjetivos valenciais e do modelo formal "estruturas de traços", apresenta-se um modelo de entrada lexical computacionalmente tratável para esse tipo de adjetivo.*

Palavras-chave. *adjetivos valenciais; formalismo; entrada lexical, processamento automático das línguas naturais.*

1. Introdução

Todos os sistemas computacionais que processam língua natural pressupõem um tipo de "arquivo" em que são armazenadas as unidades lexicais que serão manipuladas pelo sistema durante a interpretação e/ou produção do código lingüístico. Esse "arquivo" é concebido como uma base de dados em que são especificadas, para cada unidade nela contida, informações de natureza lexical, morfológica, sintática, semântica e, até mesmo, pragmático-discursiva, dependendo das especificidades do sistema de processamento de língua natural (sistema de PLN) para o qual essa base foi desenvolvida (Palmer, 2001). Do ponto de vista computacional, esse tipo de "mega arquivo" é definido como sendo o "léxico" do sistema. Atualmente, devido às aplicações reais para as quais os sistemas de PLN são escritos, é premente a compilação de léxicos (monolíngües e/ou multilíngües) que sejam (Handke, 1995; Viegas, Raskin, 1998): (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 qualidade das informações associadas às entradas. Dessa forma, a construção de léxicos para fins do PLN requer, sobretudo, a investigação de questões como: os diferentes formalismos de representação da informação lexical e as propriedades do léxico e dos itens lexicais. Neste trabalho, investiga-se uma pequena parte dessas questões, relacionada à representação lingüístico-computacional dos

adjetivos valenciais. Ressalta-se que por meio dos adjetivos são codificados aspectos subjetivos do uso lingüístico (Bakhtin, 1986; Faulstich, 1990; Borba, 1996). Dessa forma, para melhor emular a análise e síntese dos enunciados lingüísticos, os sistemas de PLN não podem prescindir da manipulação das unidades que integram essa classe (Raskin, Niremburg, 1995).

Em (2), definem-se, do ponto de (psico)lingüístico, o conceito de item lexical e os tipos de informação lexical lingüisticamente relevantes para o processamento automático das línguas naturais. Em (3), descrevem-se as principais propriedades dos adjetivos valenciais. Na seção (4), apresenta-se uma breve descrição do modelo formal *estrutura de traços* (do inglês, *"features structures"*), o qual fora escolhido para representar as propriedades descritas na seção (3). Conclui-se, em (5), com o esboço de um modelo de entrada lexical (psico)lingüístico-computacional, baseado nas *estruturas de traços*.

2. Dos pressupostos (psico)lingüísticos

2.1. Do conceito de *item lexical*

Como bem salientam Langacker (1972) e Basílio (1999), a questão da delimitação das unidades abstratas que pertencentes ao léxico é antiga e tem sido discutida sob diferentes perspectivas. Os especialistas (seja na Lingüística, na Psicolingüística ou no PLN) divergem principalmente quanto a se a representação lexical se faz por *palavras* ou por *morfemas* (radicais ou raízes), Neste trabalho, optou-se por considerar as formas pertencentes ao paradigma flexional como realizações discursivas (p.ex.: formas *embalar, embalou, embalando*) de um item lexical canônico EMBALAR e as formas pertencentes ao paradigma derivacional (p.ex.: *embalar, embaladeira, embalado*) como itens lexicais distintos e, conseqüentemente, com entradas lexicais também distintas (Handke, 1995). Dessa forma, é possível estabelecer que embalado, embalados, embalada e embaladas têm como entrada lexical canônica[1] EMBALADO. Ainda sobre a questão da delimitação das unidades pertencentes ao léxico, ressalta-se que, neste trabalho, foram tomados como base os pressupostos do projeto WordNet (FELLBAUM, 1999). De acordo com esses pressupostos, é possível estabelecer que *embalado* (=ninado), em (1) O bebê dormiu embalado por um barulhinho bom, *embalado* (=embrulhado), em (2) Pacote embalado a vácuo, e *embalado* (= animado), em (3) O Real Madrid é o time mais embalado do campeonato, são itens lexicais distintos, isto é, *embalado*1, *embalado*2 e *embalado*3. O critério empregado para a identificação de itens distintos é o da substituição por sinônimos. De fato, *embalado*1, *embalado*2 e *embalado*3 podem ser substituídos, sem que haja alteração substancial do significado das sentenças em que ocorrem, por seus respectivos sinônimos: p.ex.: *ninado, embrulhado* e *animado*, respectivamente.

2.2. Da *microestrutura* do léxico: as entradas lexicais

Partindo-se do modelo de processamento cognitivo da linguagem de Bierwisch e Schreuder (1992), Levelt (1992; 1993) e Bock (1982), concebe-se uma "entrada lexical canônica" (E) como uma estrutura de dados, no sentido computacional desse termo, composta por quatro componentes: *forma gráfica, traços gramaticais, estrutura de argumentos* e *forma semântica*. No componente **FG**, especificam-se a forma gráfica (ou ortográfica) e as estruturas morfológica e fonético-fonológica de E; em **TG**, especificam-se as propriedades

sintáticas de E e dos constituintes de hierarquia superior dos quais E é núcleo; em **EA,** especifica-se o número e o tipo semântico dos argumentos exigidos por E; em **FS**, especifica-se o conteúdo proposicional da expressão contendo E, restringindo-a. Neste trabalho, cabe ressaltar que: (i) o modelo de entrada lexical a ser proposto (cf. seção (5)) engloba apenas os componentes FG, TG e EA; (ii) o componente FG especifica apenas a forma gráfica de E (as informações fonético-fonológicas e morfológicas dos adjetivos valenciais não foram representadas nesse modelo); (ii) o componente EA é considerado uma estrutura de interface entre TG e FS; (iii) as informações referentes ao componente FS dos adjetivos valenciais não foram investigadas; na verdade, o componente EA especifica "parte" das informações semânticas referentes ao componente FS[2].

3. Das propriedades dos adjetivos valenciais: uma síntese

Na conceituação da categoria "adjetivo", pressupôs-se que é próprio desse elemento não incidir sobre si mesmo, mas sobre um suporte. Desse suporte, inclusive, o adjetivo "herda" (por meio de regras de concordância) as "categorias gramaticais secundárias" de gênero e de número.

Na conceituação da subclasse "valencial", pressupôs-se que os adjetivos em posição predicativa (p.ex.: (4) O rapaz era *descendente* de portugueses.) são verdadeiros predicadores, expressando, assim, um *"estado-de-coisas"* (predicação). Sendo predicador, o adjetivo tem a propriedade de poder ligar-se a um certo número de elementos exigidos pela sua semântica, *os argumentos* (As). P.ex.: em (4), *descendente* estabelece relação com dois As: *o rapaz* (A1) e *de portugueses* (A2). A essa propriedade, é dado o nome de **valência** (Borba, 1996). Essa propriedade poder ser descrita em três níveis. O nível da **valência lógico-semântica** diz respeito ao número de As exigidos pela semântica do predicador. Os adjetivos do Português podem ser de valência 1 (p.ex: (5) João (A1) é *bonito*), valência 2 (p.ex.: (4) O rapaz (A1) era *descendente* de portugueses (A2)), valência 3 (p.ex.: (6) O homem (A1) era *doador* de órgão (A2) para transplantes (A3)) e valência 4 (p.ex.: (7) A carga (A2) era *transportável* do navio (A3) para o cais (A4) pelos guindastes (A1)). O nível da **valência sintática** (ou subcategorização) diz respeito à realização sintática dos As lógico-semânticos. Todo adjetivo expressa um A sintaticamente obrigatório, o *sujeito* (p.ex.: *João*, em (5)); os demais As são opcionais ou facultativos (ou seja, ocorrem ou não na sintaxe). Além disso, a categoria sintagmática dos As é determinada pelo adjetivo predicador (p.ex.: em (4), o A1 realiza-se na forma de sintagma nominal (SN), *o rapaz*; o A2, na forma de sintagma preposicional (SPrep), *de portugueses*). O nível da **valência semântica** diz respeito às relações semânticas estabelecidas entre o predicador e seus argumentos; tais relações podem ser representadas por meio de *papéis temáticos*, p.ex.: em (4), o A1 tem a função semântica de TEMA e o A2, de ORIGEM.

4. Da representação lingüístico-computacional dos adjetivos valenciais

De acordo com Dias-da-Silva (1996), a representação das informações lexicais precisa ser necessariamente explícita, consistente e, principalmente, não-ambígua, para que possa ser transformada em programas computacionais. Para isso, recorre-se a uma série de formalismos, dentre os quais: *as estruturas de traços*, *a lógica de predicados* e *os frames* (Rosner, Johnson, 1992; Handke, 1995). Desses formalismos, optou-se pelas *estruturas de traços* principalmente por causa da expressividade de tais estruturas ao descrever formalmente as informações lingüísticas, uma vez que essas informações podem ser

descritas como uma *estrutura de dados*, no sentido computacional do termo (Shieber, 1986; Pollard, Sag, 1987; Keonig, 1999). Grosso modo, as *estruturas de traços* consistem na "quebra" de um dado lingüístico em partes menores, os chamados *atributos* ou *traços*. A cada traço ou atributo, são associados *valores*. P. ex: assume-se que o item <u>*carro*</u> apresenta dois traços ou atributos: NÚM(ero) e GÊN(ero), sendo que os respectivos valores desses atributos são: *sg* (singular) e *masc* (masculino), como ilustrado no esquema (1) (Shieber, 1986).

(1) <u>carro</u> [(NUM sg), (GEN masc)]

5. Conclusão: proposta de representação lingüístico-computacional

Na Figura 1, esboça-se o modelo de entrada lexical canônica para os adjetivos valenciais. Por meio de uma *matriz de atributo-valor* (MAV) (do inglês, "*attribute-value matrix*"), forma em que comumente as estruturas de traços são representadas (Koenig, 1999), exemplifica-se a entrada lexical canônica do adjetivo valencial *temeroso*:

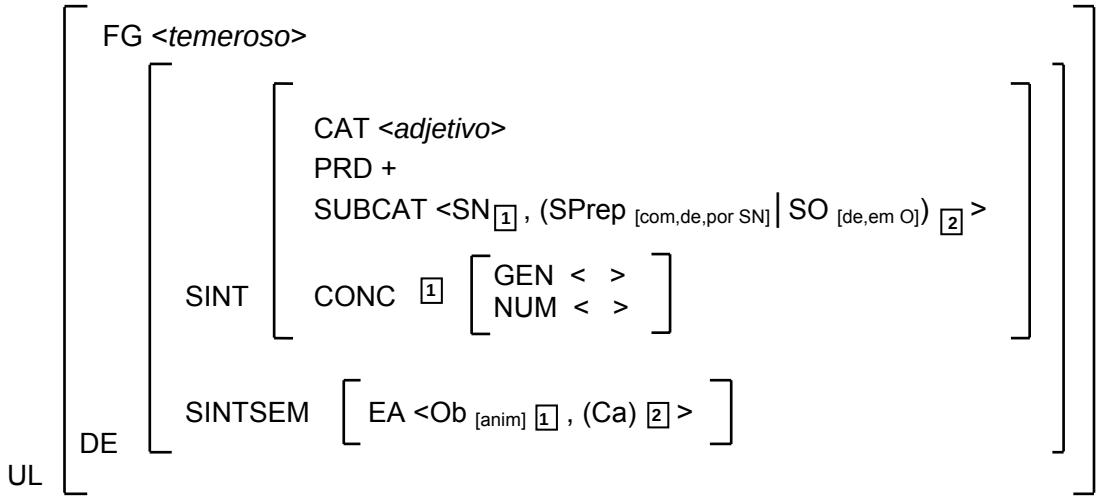

Figura 1. Modelo de entrada lexical canônica para os adjetivos valenciais

A entrada lexical canônica foi denominada UL ("unidade lexical"). Uma UL, segundo o modelo proposto, é uma estrutura de traços complexa, formada pelas estruturas FG ("forma gráfica") e DE ("descrição estrutural"). Na estrutura FG, especifica-se a forma gráfica do item lexical canônico, no caso, *temeroso*.

A estrutura de traços DE é formada por outras duas estruturas de traços complexas, as estruturas SINT(ática) e SINTSEM (sintático-semântica). A estrutura SINT abrange as informações previstas no componente TG do esquema de representação lexical (cf. Seção (2.2)). Mais especificamente, a estrutura SINT é formada por outras três estruturas CAT, PRED, SUBCAT e CONCD. O traço CAT(egoria) especifica a *categoria gramatical primária* do item, no caso, representada pelo rótulo *adjetivo*. O traço PRED, cujo valor é o sinal "+", indica que o adjetivo em questão exerce a função predicativa. O traço SUBCAT (=valência sintática) especifica a categoria sintagmática em que se realiza cada argumento do adjetivo. P.ex.: o A1 de *temeroso* realiza-se sob o forma de SN e o A2, sob a forma de SPrep ou SO (sintagma oracional). Para relacionar todas as categorias sintagmáticas nas quais o argumento A2 projetado pelo adjetivo em questão pode se realizar, foi inserida uma notação, marcada

pelo uso da barra vertical "|". Grosso modo, essa notação pode ser entendida como o *ou* lógico, que indica alternância de valores. Essa marcação, então, assume a interpretação: o argumento A2 é tal que tanto poderá se realizar na forma de SPrep [com/de/por SN] ou SO [de/em O], sendo que ambas as formas encontram-se em disjunção exclusiva, ou seja, uma delas pode ocorrer (SPrep ou SO), mas não ambas. No componente SUBCAT, representa-se também a opcionalidade e a obrigatoriedade dos argumentos. No modelo proposto, os As opcionais são especificamente representados entre parênteses. Na Figura 1, o A marcada com o índice 2, que pode realizar-se sob a forma SPrep ou SO, está representado entre parênteses. O traço CONCD, por sua vez, abrange as *categorias gramaticais secundárias* características do item lexical. As categorias secundárias de *gênero* (GEN) e de *número* (NUM), especificamente, são atribuídas aos adjetivos por meio de regras de concordância, ou seja, o *gênero* e o *número* atribuídos aos adjetivos são, na verdade, do substantivo com o qual está ligado. Dessa forma, os traços GEN e NUM dos adjetivos valenciais canônicos não são especificados (ou não-marcados) no léxico; tais traços ficam "à espera" dos traços do substantivo. Nesse traço, inclusive, foi inserido um índice numérico que indica o argumento do qual o adjetivo herdará os valores dos traços GEN e NUM. Na Figura 1, o índice 1 indica que o adjetivo herdará os traços do argumento marcada com esse índice na estrutura SUBCAT.

A estrutura de traços SINTSEM, por sua vez, abrange as informações previstas pelo componente EA. A denominação SINTSEM foi dada a essa estrutura por entender que a *estrutura de argumentos* ou *valência semântica* é uma estrutura de contato entre TG e FS. Cabe salientar que a estrutura de argumentos dos adjetivos seria representada nesse modelo de entrada por meio de *papeis temáticos* e *restrições selecionais*. Além disso, faz-se uso de *índices* para "ligar" um papel temático a um determinado argumento da estrutura SUBCAT.

No modelo proposto, ressalta-se que é possível representar as seguintes informações sobre os adjetivos valenciais: realização gráfica; categoria gramatical; tipo sintático; subcategorização; valência semântica (estrutura de argumentos); papéis temáticos; restrições selecionais; categorias de número e gênero. Representadas formalmente, tais informações podem ser utilizadas na compilação de léxicos (monolíngües) lingüístico-computacionais.

---

[1] Entende-se por "item lexical canônico" a forma lexical abstrata que pertence ao léxico e a partir da qual o módulo morfológico do sistema de PLN geraria as formas flexionadas.

[2] Na Teoria do Léxico Gerativo de Pustejovsky (1996), por exemplo, a *forma semântica* de um predicador é especificada em termos de quatro estruturas: *estrutura de argumentos; estrutura qualia; estrutura de evento* e *estrutura de herança.* Dessa forma, a estrutura de argumentos é "parte" da representação semântica de um item lexical.

Referências

BIERWISCH, M., SCHREUDER, R. From concepts to lexical items. *Cognition*, 42, p.23-60, 1992.

BOCK, J. K. Towards a Cognitive Psychology of Syntax. *Psychological Review*, 89, p. 1-47, 1982.

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

DIAS-DA-SILVA, B. C. A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais. 1996. 272p. Tese (Doutorado em Letras. Área de Concentração: Lingüística e Língua Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara.

DIK, S. C. *The theory of functional grammar*. Berlin, New York: Mouton de Gruyter, 1997.

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

KOENIG, J-P. *Lexical relations*. Stanford, CA: 1999.

LEVELT, W .J. M. Accessing words in speech production: stages, processes and representations. *Cognition*, 42, p.1-22, 1992.

____. *Speaking:from intention to articulation*. Cambridge, Mass.: The MIT Press, 1993.

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

NIRENBURG, S. *Machine translation*. San Mateo, Morgan Kaufmann, 1992.

PALMER, M. Miltilingual resources – Chapter 1. In: Hovy, E., et al. (Eds.). *Linguistica Computazionale*, v.14-15, 2001.

POLLARD, C., SAG, I. A. *Information-based Syntax and Semantics*. Volume I: Fundamentals. Stanford: CSLI Publications, 1987.

PUSTEJOVSKY, J. *The generative lexicon*. Cambridge: Mass.: The MIT Press, 1996.

RASKIN, V., NIRENBURG, S. *Lexical Semantics of Adjectives: a microtheory of adjectival meaning.* Technical Reporty Computing Research Laboratory (CRL)/New Mexico State University (NMSU), MCCS-288, 1995.

ROSNER, M, JOHNSON, R (Eds.). *Computational linguistics and formal semantics*. Cambridge: CUP, 1992.

SAINT-DIZIER, P., VIEGAS, E. *Computational lexical semantics*. Cambridge: Cambridge University Press, 1995.

SHIEBER, S. M. *An introduction to unification-based approaches to grammar*. CSLI Lecture Notes 4. Center for the Study of Language and Information. Stanford: CA, 1986.

VIEGAS, E., RASKIN, V. Computational semantic lexicon acquisition: methodology and guidelines. Relat. Téc. CRL/NMSU, MCCS-315, 1998.