Estudos Lingüísticos XXVII (Anais dos Seminários do GEL)
Trabalho apresentado no XLV Seminário do GEL na UNICAMP - Campinas em 1997
UNESP-IBILCE
São José do Rio Preto (1998), 807-812

O SIGNIFICADO DA PALAVRA PARA O PROCESSAMENTO DE LINGUAGEM NATURAL

João Luís Garcia ROSA (Pontifícia Universidade Católica de Campinas)

ABSTRACT: *Semantics is often seen as the study of word meaning. Frege (1978) said "Only in a sentence context the words have meaning". This phrase shows how finite lexical meanings can be combined to give infinite sentencial meanings. This paper considers the word meaning for Natural Language Processing systems.*

KEY WORDS: *natural language processing, semantics, word meaning, thematic roles*

0. Introdução

A semântica freqüentemente é vista como o estudo do significado das palavras. Frege (1978) dizia "Apenas no contexto de uma sentença as palavras têm significado". Esta frase mostra como finitos significados lexicais podem ser combinados para fornecer infinitos significados sentenciais. Ou seja, os significados das palavras devem ser capazes de prover uma base finita apropriada para uma teoria recursiva adequada de infinitos significados sentenciais. Este trabalho, que trata do significado da palavra para o uso em sistemas de Processamento de Linguagem Natural, está baseado, principalmente, em Chierchia & McConnell-Ginet (1990).

1. Como as palavras estão relacionadas semanticamente

Decomposição lexical. Uma idéia bastante explorada na abordagem do significado da palavra é que os significados das palavras - os conceitos que as palavras rotulam - são construídos a partir de componentes semânticos que ocorrem nos significados de diferentes palavras. As palavras não são átomos fechados. Exemplo: *pais*, *mãe*, *pai*, *casado*, *esposa*, *marido*, *feminino*, *masculino*. Pode-se relacionar estas palavras através de expressões do cálculo lambda:

(1) a. mãe' $= \lambda x[\text{pais}'(x) \wedge \text{feminino}'(x)]$
b. pai' $= \lambda x[\text{pais}'(x) \wedge \text{masculino}'(x)]$
c. esposa' $= \lambda x[\text{casado}'(x) \wedge \text{feminino}'(x)]$
d. marido' $= \lambda x[\text{casado}'(x) \wedge \text{masculino}'(x)]$

Estas traduções, apesar de fornecerem uma representação explícita das relações dos conceitos associados com as palavras, mantêm abertos os valores semânticos de *pais*, *feminino*, *masculino* e *casado*.

A decomposição serve para identificar uma quantidade de componentes semânticos universais a partir dos quais todas as línguas constróem os conceitos que seus léxicos rotulam. É desejável um vocabulário básico, independente de língua, para a linguagem da representação semântica, algum tipo de lógica interpretada ou cálculo semântico (uma lógica cujas constantes não lógicas são interpretadas).

As considerações decomposicionais do significado lexical têm sido propostas para elucidar relações semânticas entre itens lexicais.

Análise lexical sem decomposição. Um postulado do significado para o Cálculo de Predicados Intensional (IPC) é uma fórmula deste cálculo que deve ser verdadeira em qualquer modelo admissível para interpretação do IPC. A abordagem do postulado do significado para a análise do significado da palavra é substancialmente diferente da abordagem decomposicional? Se pegarmos a forma lógica (*lf*) como uma caracterização de nossa competência semântica, as duas abordagens devem ter implicações psicológicas diferentes. De fato, a evidência experimental sugere que não existe correspondência detectável entre a complexidade no nível da decomposição em características e dificuldade no processamento.

Em relação ao fenômeno da aquisição da linguagem, a abordagem decomposicional sugere que o simples conceito de *pai'* que a criança adquire primeiro é diferente do conceito complexo que irá construir mais tarde a partir de *masculino'* e *pais'*. E talvez, este último venha a substituir o primeiro. Em contraste, a abordagem do postulado do significado precisa apenas supor um enriquecimento do IPC para incluir expressões básicas como *masculino'* e *pais'*. O vocabulário cresce com o conhecimento das conexões conceituais, mas os últimos estágios do conhecimento lexical geralmente estendem, ao invés de descartarem, os estágios iniciais. Por exemplo, a criança que adiciona *feminino*, *masculino* e *pais* a um vocabulário que já contém *mãe* e *pai* pode ao mesmo tempo adicionar postulados de significado como

(2) a. $\forall x \forall y$ [pai'$(x,y) \leftrightarrow$ (masculino'$(x) \wedge$ pais'(x,y))]
b. $\forall x \forall y$ [mãe'$(x,y) \leftrightarrow$ (feminino'$(x) \wedge$ pais'(x,y))]
c. $\forall x$ [masculino'$(x) \rightarrow \neg$feminino'(x)]
d. $\forall x$ [feminino'$(x) \rightarrow \neg$masculino'(x)]

para o cálculo semântico juntamente com as novas expressões básicas *feminino'*, *masculino'* e *pais'*.

A teoria decomposicional sugere que o primeiro "pai'(x)" é substituído, mais tarde, por um "masculino'$(x) \wedge$ pais'(x)." Em contraste, a análise lexical através de postulados do significado é perfeitamente consistente com diversas representações mentais dos significados da palavra. E estas representações podem mudar. De fato esta análise não diz nada sobre como os conceitos associados com palavras individuais são estruturados ou como eles são processados na compreensão. Entretanto, ela impõe um tipo de integridade mental às palavras da linguagem natural, associando-as com unidades básicas do cálculo semântico armazenadas num léxico mental, ao invés de identificá-las como expressões complexas. Pode-se dizer que adotar a abordagem do postulado do significado não é ignorar os dados em conexões entre itens lexicais ou eliminar a possibilidade de elementos universais nas representações semânticas.

Algumas teorias decomposicionais são muito reducionistas. Por outro lado, os postulados do significado podem facilmente ligar palavras interdefiníveis sem tratar uma como mais básica do que a outra.

2. Papéis Temáticos

As abordagens que consideram papéis temáticos nas quais os papéis não são rótulos arbitrários, mas elementos com conteúdo semântico interessam ao significado da palavra.

Na literatura os relacionamentos semânticos mais específicos entre verbos e seus argumentos são referidos em termos de papéis temáticos ou papéis theta (papéis-θ). Sabe-se que o verbo *matar* tem dois argumentos aos quais ele atribui um papel-θ: o papel agente ao argumento sujeito da sentença e o papel paciente ao argumento objeto. O verbo marca-theta seus argumentos. Predicados em geral têm uma estrutura temática. O componente da gramática que controla a atribuição de papéis temáticos é chamada de teoria theta.

Ainda que muitos lingüistas concordem com a importância da estrutura temática para certos processos sintáticos, a teoria dos papéis temáticos está ainda no princípio. Por exemplo, no estágio presente da teoria não existe concordância sobre quantos papéis temáticos existem e o que significam. Alguns tipos são aceitos como gerais (Haegeman (1991)):

- *Agente/Ator*: aquele que intencionalmente inicia a ação expressa pelo predicado.
- *Paciente*: a pessoa ou coisa que sofre a ação expressa pelo predicado.
- *Tema*: a pessoa ou coisa movida pela ação expressa pelo predicado.
- *Experienciador*: a entidade que experimenta algum estado (psicológico) expresso pelo predicado.
- *Benefativo/Beneficiário*: a entidade que se beneficia com a ação expressa pelo predicado.
- *Meta*: a entidade para a qual a atividade expressa pelo predicado é dirigida.
- *Fonte*: a entidade a partir da qual alguma coisa é movida como um resultado da atividade expressa pelo predicado.
- *Localidade*: o lugar no qual a ação ou estado expresso pelo predicado está situado.

A identificação dos papéis temáticos não é sempre fácil. Entretanto, intuitivamente a idéia parece clara. A informação sobre o relacionamento semântico entre o predicado e seus argumentos é parte do conhecimento lexical do falante nativo e deve ser gravada no léxico. Na Teoria da Regência e Ligação isto é representado através de uma rede temática, ou rede theta, que é parte da entrada lexical do predicado. *Matar* teria a seguinte representação:

matar: verbo

Agente (SN)	Paciente (SN)
i	*j*

Esta representação especifica que *matar* atribui dois papéis temáticos (agente e paciente), que são dois sintagmas nominais (SN). Os índices *i* e *j* que aparecem abaixo dos papéis temáticos representam os constituintes envolvidos como em:

(3) José$_i$ matou Pedro$_j$

A *José* o verbo *matar* atribui papel temático de agente (índice *i*) e a *Pedro* o verbo atribui papel temático de paciente (índice *j*). A cada argumento é atribuído um e apenas um papel temático e cada papel temático é atribuído a um e apenas um argumento (*Critério Theta*).

Os papéis temáticos podem ser pensados como tipos de papéis específicos, tal que podemos defini-los formalmente como propriedades comuns a conjuntos de papéis específicos. A teoria baseado em eventos de papéis temáticos associa sentenças da linguagem natural com fórmulas em um cálculo de papel temático (Chierchia & McConnell-Ginet (1990)):

(4) a. José chutou Pedro.
b. $\exists$e [chutar"(e) $\wedge$ agente(e) = jose $\wedge$ tema(e) = pedro]
c. José gosta de Pedro.
d. $\exists$e [gostar_de"(e) $\wedge$ experienciador(e) = jose $\wedge$ tema(e) = pedro]

Este tipo de cálculo foi implementado na linguagem de programação lógica Prolog, considerando léxico e regras limitados (Rosa & Gonçalves (1995)). Segundo Dowty (1989), para o qual uma teoria precisa dos papéis temáticos virá da semântica, deve-se reconhecer o conteúdo semântico dos papéis temáticos, distinguindo-se um argumento de outro semanticamente, independentemente da linguagem.

Para Dowty (1989), existe o papel temático individual, que é específico para um determinado verbo e para uma determinada posição argumental deste verbo, e existe o tipo de papel temático, conjunto L, que é a interseção de todos os papéis temáticos individuais, ou seja, o conjunto que inclui Agente, Paciente, Fonte, etc. O conjunto dos papéis temáticos deve obedecer a três critérios:

a. *Completeza*: todo papel temático individual contém algum tipo de papel temático do conjunto L;

b. *Distinção*: toda posição argumental de todo verbo é distinta de toda outra posição argumental do mesmo verbo pelos tipos de papel temático do conjunto L atribuídos a essas posições;

c. *Independência*: as propriedades em um tipo de papel temático do conjunto L devem ser caracterizáveis independentemente das relações (denotadas por verbos da linguagem natural) que as acarreta.

Os critérios (a) e (b) lembram o critério-θ da Teoria da Regência e Ligação.

3. Microcaracterísticas semânticas

As palavras são representadas como listas de microcaracterísticas semânticas (Waltz & Pollack (1985); McClelland & Kawamoto (1986)). Para substantivos e verbos, as

características são agrupadas em muitas dimensões. Cada dimensão consiste de um conjunto de valores mutuamente exclusivos e, em geral, cada palavra é representada por um vetor no qual um, e apenas um, valor em cada dimensão está ligado para a palavra e todos os outros valores estão desligados.

Segundo o trabalho de McClelland & Kawamoto (1986), foram escolhidas as dimensões e os valores em cada dimensão para capturar o que se considerou dimensões importantes de variações semânticas nos significados de palavras que tinham implicações para a atribuição de casos das palavras.

Como a rede de papéis temáticos de um verbo pode ser representada como um conjunto de relacionamentos "semânticos" desta palavra, existe uma correspondência entre este tipo de representação e a representação das microcaracterísticas semânticas de McClelland & Kawamoto (1986) e Waltz & Pollack (1985), abordagem esta utilizada em Rosa (1997).

4. Imprecisão semântica

O que significa dizer que uma expressão é semanticamente imprecisa? A idéia básica é que sua contribuição às condições de verdade é indeterminada porque o critério adotado não é exato o suficiente. Considere *vermelho* por exemplo (Chierchia & McConnell-Ginet (1990)). As coisas podem ser menos vermelhas, tendendo para o laranja ou mais vermelhas tendendo para o violeta. As palavras que designam cores e muitos outros itens lexicais têm o seu conteúdo semântico fazendo referência a um protótipo que estabelece o padrão para a plena satisfação do conceito. Mas apelar para os protótipos nem sempre é relevante quando trata-se de imprecisão semântica.

Para tratar a imprecisão, ou seja, valores que podem variar entre dois níveis, por exemplo, *alto* e *baixo*, *quente* e *frio*, etc., que não podem ser representados pela lógica clássica, que considera apenas dois valores, podem ser usadas as lógicas nebulosas (*fuzzy logics*), que provêem uma multiplicidade de valores verdade entre a verdade absoluta e a falsidade absoluta (Klir & Folger (1988)).

5. Conclusão

Este texto deve ser entendido como um esboço de um estudo inicial sobre o significado da palavra, considerando alguns aspectos relevantes para o Processamento de Linguagem Natural. Foi baseado, principalmente, em Chierchia & McConnell-Ginet (1990) e foram introduzidas considerações a respeito de papéis temáticos e microcaracterísticas semânticas.

RESUMO: *A semântica freqüentemente é vista como o estudo do significado das palavras. Frege (1978) mostrou como finitos significados lexicais podem ser combinados para fornecer infinitos significados sentenciais. Este trabalho trata do significado da palavra para sistemas de Processamento de Linguagem Natural.*

PALAVRAS-CHAVE: *processamento de linguagem natural, semântica, significado da palavra, papéis temáticos.*

REFERÊNCIAS BIBLIOGRÁFICAS

CHIERCHIA, G. & MCCONNEL-GINET, S. (1990). *Meaning and Grammar - An Introduction to Semantics*. The MIT Press.

DOWTY, D. R. (1989). On the Semantic Content of the Notion of 'Thematic Role'. In Chierchia, G. & Partee, B. H. & Turner, R. (Eds.) *Properties, Types, and Meaning, vol. 2, Semantic Issues*. Dordrecht: Kluwer.

FREGE, G. (1978). *Lógica e Filosofia da Linguagem.* Editora Cultrix. São Paulo.

HAEGEMAN, L. (1991). *Introduction to Government and Binding Theory*. Cambridge: Blackwell.

KLIR, G. A. & FOLGER, T. A. (1988). *Fuzzy Sets, Uncertainty, and Information*. Prentice-Hall.

MCCLELLAND, J. L. & KAWAMOTO, A. H. (1986). Mechanisms of Sentence Processing: Assigning Roles to Constituintes of Sentences. In J. L. McClelland and D. E. Rumelhart (eds.) *Parallel Distributed Processing - Explorations in the Microstructure of Cognition*, Volume 2. The MIT Press.

ROSA, J. L. G. (1997). A thematic connectionist approach to Portuguese language processing. IASTED International Conference on Artificial Intelligence and Soft Computing. Banff - AB - Canada. July 27 - August 1 (a ser publicado).

ROSA, J. L. G. & GONÇALVES, R. F. (1995). Fragmento de Gramática com Advérbios. Trabalho Final da Disciplina Tópicos em Semântica II. IEL-UNICAMP, Campinas - SP. Fevereiro.

WALTZ, D. L. & POLLACK, J. B. (1985). Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretations. *Cognitive Science* 9, 51-74.