

## O léxico na interface sintático-semântica: perspectivas e limitações computacionais

Ana Maria Ibaños<sup>1</sup>, Carlos A. Prolo<sup>2</sup>, Jorge Campos da Costa<sup>3</sup>

<sup>1</sup>Faculdade de Letras – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

<sup>2</sup>Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

<sup>3</sup>Faculdade de Letras – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

ana.ibanos@pq.cnpq.br, carlos.prolo@pucrs.br, jcampos@pucrs.br

**Resumo.** *O objetivo principal do texto é apresentar as bases teóricas que motivaram o projeto de nosso grupo de estudos Formalismos Lingüísticos e Computação (FLEC), que visa a construção de um sistema de relações interdisciplinares entre lingüística e computação, em que ambas as áreas sejam beneficiadas pelo conhecimento compartilhado de formalismos relevantes no tratamento da interface léxico-sintaxe-semântica.(Campos (2004)).*

**Abstract.** *The purpose of this paper is to present the theoretical foundations from a project developed by our group Linguistic Formalisms and Computer Sciences (FLEC) that aims at building a system of interdisciplinary relations between Linguistics and Computational Sciences in which both areas can profit by the shared knowledge in terms of relevant formalisms in the treatment of the lexico-semantic-syntactic interface.*

**Palavras-chave:** formalismos lingüísticos, modelos computacionais; léxico

### 1. Caracterização do problema

As pesquisas ligadas à área de Processamento da Linguagem Natural (PLN) não são recentes, mas, ainda hoje, observa-se que há questões que precisam ser mais bem trabalhadas para que se possa assegurar o desenvolvimento constante dos sistemas que manipulam o código lingüístico. O processamento automático da linguagem natural requer tanto conhecimentos da Computação quanto da Lingüística. Entretanto, na maioria das vezes, os trabalhos realizados nessa área são desenvolvidos nos Cursos de Ciências da Computação com a finalidade de buscar soluções para problemas de implementação, sem preocupação com questões teóricas lingüísticas. Esses trabalhos acabam muitas vezes ignorando as possíveis contribuições que as teorias lingüísticas podem trazer para a criação e aperfeiçoamento de sistemas computacionais que necessitam processar a linguagem natural, bem como o impacto de PLN sobre as investigações propriamente lingüísticas. Sag & Wason (1999) grifam que mesmo tecnologias já eficientes podem se beneficiar de um conhecimento mais sofisticado das propriedades da linguagem natural. Os sistemas de PLN, por mais simples que sejam,

tendem a exigir uma descrição lingüística rica e adequada à sua aplicação. Nos últimos anos, tornou-se evidente que os recursos lingüísticos e, em particular, os recursos lexicais, formam a base de qualquer sistema computacional que pretenda processar a linguagem natural (cf. Ranchhod (2001)). Implementações de ferramentas computacionais de impacto baseadas nos mais diversos paradigmas têm-se ancorado nos princípios de que as regras nos diversos níveis da descrição lingüística estão associadas às palavras no léxico (vide por exemplo, Schabes (1990), XTAG Research Group (2001) para LTAGs; Charniak (1997) para Gramáticas Livres de Contexto Probabilísticas Lexicalizadas; e as Gramáticas "Markovianas" de Collins (1997) e Charniak (2000) ). Assim, quanto mais informações lingüísticas estiverem armazenadas no léxico do sistema computacional, maior será sua eficácia e eficiência.

Pustejovsky e Boguraev (1996) afirmam que, independentemente da sofisticação do sistema, seu desempenho deve ser medido em grande parte pelos recursos do léxico computacional associado a ele. Então, para o tratamento automático da linguagem natural, é necessário que se tenham descrições sistemáticas e completas, pois a insuficiência de informações lingüísticas adequadas pode gerar falhas e limitações no processamento automático.

Um dos obstáculos enfrentados na avaliação e aperfeiçoamento dos sistemas computacionais é justamente a falta de um trabalho cooperativo entre as duas áreas, pois, de um lado, os cientistas da computação têm domínio limitado das teorias lingüísticas, não conseguindo lidar com alguns problemas inerentes à linguagem natural; e de outro lado, os lingüistas não têm uma noção mais precisa dos problemas que os cientistas da computação gostariam que eles descrevessem e explicassem.

Nosso projeto procura estabelecer formalizações de fenômenos lingüísticos que sejam passíveis de serem convertidas em um programa de computador de forma mais natural possível, evitando a utilização de símbolos artificiais — que trazem soluções *ad hoc* para os problemas no processamento automático da linguagem natural.

## 2. Fundamentação Teórica

As Teorias e Formalismos a serem considerados são: X-barras (Chomsky 1981, 1986), Teorias das representações do léxico conceituais — Jackendoff (1990), Pustejovsky (1995), as LTAGs — Lexicalized Tree Adjoining Grammars — Schabes (1990), Joshi (1997, 1999) e a LFG — Lexical Functional Grammars — Kaplan e Bresnan (1982) e Head-Driven Phrase Structure Grammar — Pollard & Sag (1994), Gramáticas Categoriais como Steedman (1991), DRT – Discourse Representation Theory – Kamp & Reyle (1993).

O interesse por estudar a teoria X-barras, adotada pela Teoria de Princípios & Parâmetros e Modelo da Regência e Ligação (Chomsky, 1981 e Chomsky & Lasnik, 1995, respectivamente) se justifica por ser um modelo de análise sintagmática em que subteorias como Teoria  $\theta$  e Teoria do caso, assim como a relação computacional de mova  $\alpha$ , trabalham com o léxico em um sentido gerativo computacional, que dão grande suporte às questões relacionadas à construção do aparato sintático das sentenças.

A escolha pelas teorias das representações léxico-conceituais se dá pelo fato de que são teorias mais detalhadas em termos de conhecimento dos aspectos semânticos. Acredita-se que os componentes semânticos dos predicadores podem funcionar como restrições de seleção. Em Jackendoff (1983,1990), tem-se alguns recursos como as categorias conceituais, os primitivos conceituais e campos semânticos. Já em Pustejovsky (1991,1995), pode-se contar com uma proposta de um léxico *enriquecido*, pois as entradas lexicais contêm todas as informações consideradas necessárias para a caracterização das unidades lexicais. Essas informações encontram-se especificadas em vários níveis de representação (estrutura argumental, estrutura de eventos e estrutura qualia).

Desde que o formalismo da Tree Adjoining Grammar (TAG) foi originalmente introduzido em Joshi, Levi e Takahashi (1975) - embora aconselha-se que seja aprendido em sua formulação revisada atual como em Joshi e Schabes (1997) - , sua importância para a linguagem natural tem sido estabelecida de vários modos, entre outros: por sua relevância no domínio lingüístico, como uma elegante alternativa *não-transformacional* para descrição de gramáticas (e.g., Joshi (1985), Kroch e Joshi (1985), Frank (2002) );por permitir uma modelagem da linguagem natural em todos os seus aspectos (sintático, semântico e pragmático) como propriedades do léxico (Schabes (1990), Joshi e Schabes (1997)), através das TAGs Lexicalizadas (LTAGs); por sua utilidade na evocação de subsequente processamento semântico, devido ao conteúdo semanticamente rico de seus históricos de derivação (Joshi (1999), Kallmeyer (2005)); por suas implicações psicolinguísticas (e.g., Joshi, Becker e Rambow (2000), Joshi (1990), Kinyon (1999); e, naturalmente, por suas propriedades computacionais favoráveis (Vijay-Schanker (1988), Schabes (1990) -- dado seu altamente desejável poder de descrição, -- que a permitiu tornar-se uma das alternativas líderes na pesquisa para desenvolvimento de parsers para gramáticas de linguagem natural de grande cobertura (e.g. XTAG Research Group (2001), Joshi (2001), Prolo (2002, 2003, 2004).

A relevância lingüística das TAGs vem de sua adequação a um certo tipo de metodologia de projeto de gramáticas, que apresenta, entre outros aspectos, as propriedades de : Domínio de Localidade Estendido: Fatoração da Recursão e Descrições ricas.

Já a Lexical Functional Grammar (Kaplan e Bresnan, (1982)) é interessante por ser um formalismo elegante que apresenta regras gramaticais simples, e incorpora os aspectos complexos à representação do léxico. A LFG é um modelo sintático que tem como objetivo fornecer uma representação da linguagem computacionalmente precisa e psicologicamente realista (Sells (1985)). Utiliza-se de um modelo de múltiplos níveis de representação, cada um com sua própria arquitetura, vocabulário e restrições. Diferentemente de teorias como PP, são níveis derivados paralelamente e ligados através de um mapeamento restrito por princípios de correspondência. Para a LFG , sintaxe não está baseada somente em estrutura e tem como suposição inicial de que sujeito e objeto são constituintes primitivos. O mesmo pode, também, ser dito para a HPSG com seu princípio de lexicalismo estrito, em que a estrutura da palavra e a

estrutura do sintagma são governadas por princípios parcialmente independentes, isto é, palavras e sintagmas são dois tipos de signos.

Com relação às Gramáticas Categoriais (GTs) (Wood (1993); Ibaños (1996)), elas têm o forte apelo de construir modelos isomórficos entre sintaxe e semântica, e de serem, provavelmente, os mais antigos formalismos na área da lingüística computacional. As Gramáticas Categoriais podem ser caracterizadas por três propriedades básicas, independentemente das variedades de modelos que se abrigam dentro deles. Primeiro, são gramáticas em que os conceitos de função e argumento substituem a dupla sujeito/predicado da tradição, com precisão e objetividade que favorecem a implementação; segundo, porque sintaxe e semântica são trabalhadas de maneira homomórfica, dentro da assim chamada rule-to-rule hypothesis; finalmente, porque a concepção de léxico já traz incorporado o conjunto de regras da sintaxe, permitindo uma interface léxico/sintaxe/semântica mais adequada.

A DRT, Discourse Representation Theory, (Kamp e Reyle (1993); Andrade (2002)) é uma das teorias mais atraentes do ponto-de-vista semântico. Trata-se de uma abordagem formal, dentro da tradição montaguena, mas diferindo desta ao trabalhar o significado enquanto interdependente em relação ao contexto. Faz uma abordagem discursiva, propondo um algoritmo que captura a dinamicidade do processo semântico, sendo compatível com algumas das teorias sintáticas mais expressivas como a GB, a LFG, a HSPG e Gramáticas Categoriais, tais como anteriormente apresentadas. Além disso, propõe-se, de forma explícita, a ser implementável computacionalmente.

Cabe observar que as teorias lingüísticas devem não apenas configurar os fenômenos da linguagem e sua resolução, mas também devem oferecer uma descrição passível de implementação computacional. Destaca-se também que a facilidade e mesmo a possibilidade do uso computacional de uma teoria lingüística depende das características da teoria em si e do modo como ela é descrita. Características desejáveis intrínsecas à teoria são correteza — capacidade de explicar corretamente os fatos — e a completude — condição de explicar o maior número de fenômenos possíveis no escopo para o qual é proposta.

No entanto, para a implementação computacional não basta a conveniência das propriedades intrínsecas, é preciso que sua formulação seja passível de ser convertida em um programa de computador. Muitas vezes, uma teoria lingüística oferece uma descrição lingüística sofisticada, mas é de difícil implementação computacional. Quanto mais formalizada é a teoria, mais fácil a sua implementação computacional. Acredita-se que o cerne de uma boa descrição da teoria é o uso de formalismos adequados.

Dados os aspectos acima considerados, o projeto tem como propósito:

- (a) avaliar teorias lingüísticas formais com relação ao papel do léxico ;
- (b) comparar formalismos lingüísticos em relação às suas potencialidades e limitações para a modelagem do léxico;
- (c) investigar as possibilidades de implementação computacional de tais teorias;

- (d) propor, considerando-se (a), (b) e (c), modelos de implementação computacional de uma teoria lexical, tendo em vista futuro desenvolvimento de aplicações computacionais.

O que o nosso GT apresenta, é justamente uma amostra das discussões que vem sendo analisadas no FLEC.

### 3. Referências

BEARDON, C.; LUMSDEN, D. & HOLMES, G. *Natural Language and Computational Linguistics*, England: Ellis-Horwood, 1991.

CAMPOS, J. **Os enigmas do nome - na interface lógica/semântica/pragmática**. Porto Alegre: AGE/EDIPUCRS, 2004.

CHARNIAK, Eugene. **Statistical Parsing with a Context-free Grammar and Word Statistics**. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park, CA, USA. 1997.

CHARNIAK, Eugene. **A Maximum-Entropy-Inspired Parser**. In: *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2000)*. Seattle, WA, USA, 2000. p. 16-23.

CHOMSKY, Noam. **Lectures on Government and Binding**. Dordrecht: Foris, 1981.

CHOMSKY, Noam. **Knowledge of Language Its Nature, Origin and Use**. New York: Praeger, 1986.

CHOMSKY, N. **The Minimalist Program**. Cambridge: The MIT Press, 1995.

CHOMSKY, N. e H. LASNIK (1995) **The theory of principles and parameters**. In: CHOMSKY, *The minimalist program*. Cambridge, MA.: MIT Press, 1995.

COLLINS, Michael. **Lexicalized Models for Statistical Parsing**. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain, 1997. p. 16-23

CRUSE, D.A. **Lexical Semantics**. Cambridge University Press, 1986.

FRANK, Robert. **Phrase Structure Composition and Syntactic Dependencies**. MIT Press, 2002.

GRISHMAN, R. *Computational Linguistics: An Introduction*. **Studies in Natural Language Processing**. Cambridge: Cambridge University Press, 1992.

IBAÑOS, Ana Maria. Gramática e Linguagens categoriais. **Ciências e Letras**, 1996.

GT Inferências em Linguagem Natural, uma abordagem via interface externa lingüístico-lógico-computacional e interface interna sintático-semântico-pragmática dos operadores sentenciais

- JACKENDOFF, R. **Semantics and Cognition**. Londres: MIT Press, 1983.
- JACKENDOFF, R. S. **Semantic Structures**. Cambridge/Mass.: The MIT Press, 1990.
- JOSHI, A.; LEVY, L. & TAKAHASHI, M. Tree Adjunct Grammars. *Journal of the Computer and System Sciences*, v.10, n.1, New York: Academic Press, 1975.
- JOSHI, A. K.; SCHABES, Y. Tree-Adjoining Grammars. In: **Handbook of Formal Languages**, Salomaa, A.; Rozemberg, G. (eds.), v. 3. Springer-Verlag, 1997. p.69-123.
- JOSHI, A.; VIJAY-SHANKER, K. Compositional Semantics with LTAG: How Much Underspecification is Necessary? In: *Proceedings of the 3<sup>rd</sup> International Workshop on Computational Semantics*. Tilburg, The Netherlands. 1999.
- JOSHI, A. K.; BECKER, Tilman; and RAMBOW, Owen. Complexity of Scrambling: A new Twist to the Competence/Performance Distinction. In: **Tree Adjoining Grammar: formalisms, linguistic analysis and processing**. ABEILLE, Anne; RAMBOW, Owen (eds.) CSLI, Stanford, USA, 2000.
- JOSHI, A. K. The XTAG Project at Penn. In: *Proceedings of the 7<sup>th</sup> International Workshop on Parsing Technologies (IWPT 2001)*. Beijing, China. 2001.
- KALLMEYER, Laura. Tree-local Multicomponent Tree Adjoining Grammars with Shared Nodes. **Computational Linguistics**, v. 31, n. 2, p. 187-225, 2005.
- KAMP, H. & REYLE, U. **From Discourse to Logic**. Kluwer Academic Publishers: London, 1993.
- KAPLAN, R. & BRESNAN, J. Lexical-functional grammar: a formal system for grammatical representation. In J. Bresnan, editor, **The Mental Representation of Grammatical Relations**, pages 173--281, 1982.
- KROCK, Anthony S.; JOSHI, A. K. The Linguistic Relevance of Tree Adjoining Grammar. *Technical Report MS-CIS-85-15*, University of Pennsylvania, 1985.
- LEVIN, B.; RAPPAPORT HOVAV, M. **Unaccusativity : at the syntax-lexical semantics interface**. Cambridge (MA) : MIT Press, 1996.
- POLLARD, C., SAG, I. A. **Head-driven phrase structure grammar**. Chicago: University of Chicago Press, 1994.
- PROLO, Carlos A. Fast LR Parsing Using Rich (Tree Adjoining) Grammars. In: *Proceedings of the Seventh Conference on Empirical Methods in Natural Language Processing*. Philadelphia, USA, 2002, p. 103-110.
- PROLO, Carlos A. Generating the XTAG English Grammar Using Metarules. In: *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, Taipei, Taiwan, 2002. p. 814-820.

- PROLO, Carlos A. *LR Parsing for Tree Adjoining Grammars and its Application to Corpus-based Natural Language Parsing*. Ph.D. Thesis. Department of Computer and Information Science, University of Pennsylvania. Junho, 2003.
- PROLO, Carlos A. An efficient LR parser generator for Tree Adjoining Grammars. In: **Parsing Technologies**. Harry Bunt and John Carroll and Giorgio Satta (eds.). Kluwer Academic Publishers, Boston, MA, USA. 2004.
- PUSTEJOVSKY, James. The syntax of event structure. **Cognition**, v. 41, p. 47-81, 1991.
- PUSTEJOVSKY, James. **The generative lexicon**. Cambridge: The MIT Press, 1995.
- PUSTEJOVSKY, J., B. Boguraev. **Lexical Semantics: The Problem of Polysemy**, Oxford University Press, pp. 1-14, 1996, 1996.
- RANCHHOD, Elisabete Marques (2001), **O Uso de Dicionários e de Autómatos Finitos na Representação Lexical das Línguas Naturais**. In Ranchhod, Elisabete M. (org.) *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações*, Lisboa: Caminho (pp. 13-47).
- SAG, I.A. & WASOW, T. **Syntactic Theory: A formal introduction**. Stanford: CSLI Publications, 1999.
- SAINT-DIZIER, Patrick; VIEGAS, Evelyne Viegas. **Computacional Lexical Semantics**. Cambridge University Press, 1995.
- SCHABES, Yves. *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph.D. Thesis. Department of Computer and Information Science, University of Pennsylvania. 1990.
- SELLS, Peter. **Lectures on contemporary syntactic theories: an introduction to government-binding theory, generalized phrase structure grammar, and lexical-functional grammar**. Stanford: University of Chicago Press, 1985.
- STEEDMAN, Mark. Type-raising and directionality in combinatory grammar. **Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics**, Berkeley CA, p. 71-79, 1991.
- WALLACE, L. Chafe. **Significado e Estrutura Lingüística**. Rio de Janeiro: Livros Técnicos e Científicos, 1979.
- WOOD, Mary McGee. **Categorial Grammars**. New York: Routledge, 1993.