CNPQ – Conselho Nacional de Desenvolvimento Científico

Léxico na interface sintático-semântica:

Perspectivas e Limitações Computacionais

ANA MARIA TRAMUNT IBAÑOS
JORGE CAMPOS DA COSTA

1. Caracterização do Problema

As pesquisas ligadas à área de Processamento da Linguagem Natural (PLN) não são recentes (cf. Joshi, 2002), 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. Segundo Ranchhod (2001), nos últimos anos, tornou-se evidente que os recursos lingüísticos e, em particular, os recursos lexicais, são a base de qualquer sistema computacional que pretende processar a linguagem natural (COLOCAR AQUI AS REFERÊNCIAS). Assim, quanto mais informações lingüísticas estiverem armazenadas no léxico do sistema computacional, maior será sua eficiência. (RESSALTAR O PAPEL DO LÉXICO + REFERÊNCIAS QUE O PROLO VAI MANDAR)

Boguraev & Pustejovsky (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 apenas domínio superficial das teorias lingüísticas, não conseguindo lidar com alguns problemas inerentes à linguagem natural; e de outro lado, os lingüistas não têm noção dos problemas que os cientistas da computação gostariam que eles resolvessem, explicassem.

MOTIVAÇÃO

Nesse sentido, o presente projeto, apresenta a vantagem de contar entre seus proponentes especialistas da computação que efetivamente trabalham na implementação computacional da linguagem natural e de profissionais da lingüística que possuem um

maior domínio sobre os fenômenos lingüísticos. Com tal parceria, busca-se investigar teorias lingüísticas que não visem apenas a configurar os fenômenos da linguagem e sua resolução, mas também a eficiência necessária à sua inclusão em aplicações. Uma dentre as inúmeras aplicações de tais teorias é o apoio à implementação de ferramentas computacionais, como corretores ortográficos, tradutores automáticos, respondedores automáticos de questões, sumarizadores de texto, motores de busca da Internet, interpretadores de ordens, etc.

OBJETIVOS

Dados os aspectos acima considerados, o presente 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.

2. Objetivos

O objetivo do presente projeto é analisar algumas teorias lingüísticas importantes ao nível da interface sintaxe-semântica à luz de formalismos típicos com vistas à sua implementação computacional. Este projeto está vinculado ao Programa de Pós-Graduação em Letras da PUCRS, e conta também, como já foi mencionado, com profissionais da ciência da computação. Metodologicamente, unindo esforços dos profissionais das duas áreas, busca-se estar um passo à frente, ou ainda, procura-se ter formalizações dos 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.

O grupo envolvido com essa pesquisa para a qual se está solicitando auxílio por meio do presente projeto, além dos objetivos gerais de desenvolver um estudo acerca das teorias lingüísticas e dos formalismos usados para a sua descrição também possui outros objetivos secundários, como de promover eventos de intercâmbio científico no campo da Lingüística e da Computação através de palestras, seminários e cursos, de maneira a integrar Pós-Graduação e Graduação. Também se planeja participar de cursos, congressos, seminários, interagindo com outros grupos regionais, nacionais e internacionais. Cabe salientar que no momento, já se tem colaboração informal com outros grupos de pesquisadores: Vera Lúcia Strube de Lima — Faculdade de informática da PUCRS, Renata Vieira — Unisinos, Luís Sarmento — Engenharia Informática da FEUP/ Linguateca-Portugal, Eckhard Bick— VISL Project Leader.

Ainda como reflexo da pesquisa promovida nesse projeto, pretende-se produzir artigos para serem publicados em periódicos reconhecidos. Além disso, tem-se a intenção de construir um glossário com os termos técnicos utilizados pelos profissionais envolvidos

com pesquisas dessa natureza, pois, para processar automaticamente a linguagem natural, buscam-se conceitos de diferentes áreas do conhecimento, e, às vezes, esses conceitos são mal empregados, trazendo sentidos dúbios. É importante mencionar que, após a execução dessa primeira etapa do projeto para o qual se está pleiteando auxílio, pretende-se ampliar as tarefas do grupo com futuros projetos indo da implementação de teorias lingüísticas com vistas ao aperfeiçoamento de sistemas computacionais até a construções de novas ferramentas.

3. Metodologia e estratégias de ação

Como já foi dito, propõe-se, nessa pesquisa, analisar as teorias lingüísticas cruzadas com formalismos típicos usados para sua descrição, à luz do princípio lexicalista em voga. Para realizar as etapas previstas no projeto, o grupo de trabalho é composto por Ana Maria Tramunt Ibanos e Carlos Augusto Prolo (responsáveis), Gilberto Keller de Andrade, Jorge Campos da Costa e Simone Sarmento (pesquisadores colaboradores); Gabriel Othero de Ávila, Gabriela B. Hinrichs Conteratto, Gustavo Brauner, Karina Molsing (pesquisadores auxiliares). O projeto terá como eixo a metodologia de trabalho em PLN elaborada por Dias-da-Silva (1996), a qual prevê o desenvolvimento do trabalho em três fases, cada uma abordando questões específicas:

> **1- Fase Lingüística -** analisar algumas teorias lingüísticas.
> **2- Fase Representacional** – estudar determinados formalismos de representação para os conhecimentos reunidos no domínio lingüístico que sejam computacionalmente tratáveis. Ou ainda, cruzam-se as teorias lingüísticas com alguns formalismos típicos.
> **3- Fase implementacional -** refletir a viabilidade de futuras aplicações. Ou até mesmo, a criação de novas ferramentas computacionais no âmbito da PLN.

Teorias e Formalismos a serem considerados: X-barra (Chomsky 1981, 1986), Teorias das representações do léxico conceituais — Jackendoff (1990), Pustejovsky (1995), as LTAGs — Lexicalized Tree Adjoinng Grammars — (Schabes1990), (Joshi 1997, 1999) e a LFG — Lexical Functional Grammars — (Kaplan e Bresman (1982)) O interesse por estudar a teoria X-barra, adotada pela Teoria de Princípios & Paramêmtros e Modelo da Regência e Ligação (Chomsky&Lasnik 91e Chomsky 81, respectivamente) se justifica por ser um modelo de análise sintagmática com um grande grau de difusão e aceitação entre sintaticistas gerativistas.

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 (1990), tem-se alguns recursos como as categorias conceptuais, os primitivos conceptuais e campos semânticos. Já em Pustejovsky (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. Estas informações encontram-se especificadas em vários níveis de representação (estrutura argumental, estrutura de eventos e estrutura qualia).

As LTAGS (Schabes (1990) e Joshi (1997,1999)) são relevantes para esse estudo por serem consideradas como gramáticas lexicalizadas. Já a LFG (Kaplan e Bresman (1982)) é interessante por ser um formalismo elegante que apresenta regras gramaticais simples, e incorpora os aspectos complexos à representação do léxico. Vale mencionar 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. Cabe destacar também que a facilidade e mesmo a possibilidade do uso de uma teoria lingüística computacionalmente 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 é a sua implementação computacional.Acredita-se que o cerne de uma boa descrição da teoria é o uso de formalismos adequados. Vale lembrar que, com o advento de grandes corpora de materiais lingüísticos, a avaliação das teorias ficou mais representativa e produtiva. Nosso projeto também deverá contemplar a importância do uso de corpus nas análises a serem feitas.

O potencial de aplicação deste projeto é enorme, pois um dos maiores empecilhos da pesquisa no tratamento computacional da linguagem são formalizações inadequadas, incorretas ou incompletas. Geralmente, a dimensão dos problemas só é considerada após implementações, através de números indicadores de percentual de cobertura e correção. Os pesquisadores não se dão ao trabalho de averiguar quais os motivos de números baixos na avaliação. Por vezes, dada a necessidade de justificar os resultados para a escrita de artigos, explicações superficiais e descompromissadas são elaboradas — que não podem ser verificadas pelos leitores. Ainda é muito comum o pesquisador não revelar os problemas de cobertura e correção, por medo de que tais falhas, quando devidamente qualificadas, venham a depor contra sua ferramenta, ou a abordagem de implementação da teoria semi-formal, ou mesmo contra a teoria de que ele se tornou adepto.

4. Principais Referências Bibliográficas

ALLEN, James. Natural ***Language Understanding***. Benjamin/Cummings, 2nd edition, 1995.

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

BOURBEAU, L.; Carcagno, D.; Goldberg, E.; Kittredge, R.; Polguère, A. ***"Bilingual generation of Weather Forecasts in an Operations Environment***", Hans Karlgren (ed.), *Proceedings of COLING'90* (Helsinki, 1990), Vol. 1 (pp.90-92),1990.

CARLSON, G. **Reference to kinds in English.** Tese de Doutorado, University of Massachusetts: Amhrest, 1977.

COHEN, D.I. **Introduction to Computer Theory**. New York: Wiley & Sons, 1990.

CHOMSKY, Noam. ***Aspectos da Teoria da Sintaxe.*** Traduzido por J. A. Meireles e E. P. Raposo. Coimbra: A. Amado, 1975. Tradução de: Aspects of the Theory of Syntax.

CHOMSKY, Noam. ***O Conhecimento da Língua. Sua Natureza, Origem e Uso.*** Traduzido por Anabela Gonçalves e Ana Teresa Alves. Lisboa: Caminho, 1994. Tradução de: Knowledge of Language. Its Nature, Origin and Use.

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

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

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

CRUSE, D.A. ***Lexical Semantic*s**, Cambridge University Press, 1986.

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

DIMARCO, Chrysanne; Foster, Mary Ellen (1997), ***"The automated generation of Web documents that are tailored to the individual reader"***, *Natural Language Processing for the World Wide Web, Papers from the 1997 AAAI Symposium* (Stanford, March 24-26, 1997), Menlo Park, California: AAAI Press (pp.44-53).

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

JACKENDOFF, R . ***Semantics and Cognition*.** Londres: MIT Press, 1983.

JACKENDOFF, R. S. **Semantic Structures**. Cambridge/Mass.: The MIT Press, 1990.

JESPERSEN, O. ***The philosophy of grammar***. Great Britain : Unwin Brothers LTD, 1948.

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. Tree-Adjoining Grammars: How much context-sensitivity is required to provide reasonable descriptions?**.** In: **Natural Language Parsing**, Dowty, Kartunnen, Zwick (eds.). Cambridge University Press, 1995. p.206-250.

JOSHI, A.K. Parsing Techniques. In: COLE, R.A; et al. (eds.) **Survey of the State of the Art in Human Language Technology**. Philadelphia, PA: University of
Pennsylvania, 1994.

JOSHI, A.K. Tree Adjoining Grammars and Lexicalized Grammars. In: M. Nivat and A. Podelski (eds.) **The Automata and Languages**, Philadelphia, PA, 1992.

KROCH, A.S. & JOSHI, A.K. Analyzing Extraposition in a Tree Adjoining Grammar, **Syntax and Semantics**, v.20, 1987. p.107-149.

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.

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 Lingüística Computacional e suas Aplicações*, Lisboa: Caminho (pp. 13-47).

SAINT-DIZIER, Patrick; VIEGAS, Evelyne Viegas. ***Computacional Lexical Semantics.*** Cambridge University Press, 1995.

SHIEBER, S. M. ***The design of a computer language for linguistic information***. In Proceedings of the 10th International Conference on Computational Linguistic (pp 362-366), Stanford University, CA, 1984. COLING.

WALLACE, L. Chafe. ***Significado e Estrutura Lingüística.*** Traduzido por Maria Helena de Moura Neves. Rio de Janeiro: Livros Técnicos e Científicos, 1979. Tradução de: Meaning and the Structure of Language.