PROLO, Carlos A. Tree-Adjoining Grammars: uma entrevista com Carlos Prolo. *Revista Virtual de Estudos da Linguagem - ReVEL*. Vol. 6, n. 10, março de 2008. ISSN 1678-8931 [www.revel.inf.br].

TREE-ADJOINING GRAMMARS – UMA ENTREVISTA COM CARLOS PROLO

Carlos A. Prolo

Pontifícia Universidade Católica do Rio Grande do Sul

ReVEL – Que tipo de formalismo sintático são as Tree-Adjoining Grammars (TAGs)?

Prolo – Antes de dizer o que são as TAGs, deixe-me dizer o que elas não são. Algumas pessoas lêem sobre as TAGs e imaginam que elas sejam uma formalização de Teoria Lingüística das Línguas Naturais. Esta confusão acontece por vários motivos. Por um lado, há outros formalismos conhecidos dos lingüistas, como LFG, GPSG, cujo desenvolvimento geralmente aparece associado ao desenvolvimento de concepções teóricas específicas da linguagem. Por outro lado, existe um projeto de longa duração, na Universidade da Pennsylvania (U. Penn, Penn), liderado pelo professor Aravind Joshi, de concepção de uma gramática da língua inglesa (XTAG) que tem por base o uso das TAGs. Então alguns lingüistas acabam fazendo uma associação das TAGs com uma tentativa de impor ou reafirmar elementos de um teoria da linguagem, e.g., como é a estrutura dos constituintes, quem é o núcleo do constituinte, até elementos mais sofisticados como caso gramatical, ECM (*Exceptional Case Marking*), etc. TAG não se prende a concepções específicas de uma teoria da sintaxe ou do léxico. Antes, ela é um formalismo, que permite ao lingüista descrever as suas teorias de acordo com suas concepções. Neste sentido, ela é como as Gramáticas Livres de Contexto (CFGs), que foram usadas por Chomsky para formalização da Teoria da Regência e Ligação, mas que também são usadas

para descrever linguagens de programação. Aliás, interessante notar, há vários trabalhos usando TAGs para tentar descrever a estrutura de proteínas como seqüências de aminoácidos.

Vamos agora ao que são as TAGs. Como eu disse, elas são um formalismo gramatical como o são as Gramáticas Livres de Contexto. E como as CFGs, suas limitações são bastante conhecidas dos lingüistas, vou partir daí para caracterizar as TAGs. Primeiro, as CFGs são gramáticas de reescrita de *strings*, enquanto as TAGs são gramáticas de reescrita de árvores. Em uma CFG, cada regra diz como um símbolo, representando uma categoria sintática, e.g., uma sentença ou sintagma nominal, pode ser reescrito, ou decomposto como uma seqüência de outros símbolos. Assim, uma sentença da linguagem é gerada a partir de um sintagma por um processo recursivo de aplicação destas regras, partindo do sintagma e terminando na seqüência de categorias lexicais correspondentes à sentença. Este processo é melhor visualizado como uma *árvore de derivação*, que é então entendida como a estrutura sintática da tal sentença. A principal crítica ao uso das CFGs para linguagem natural é que a expansão de cada nodo desta árvore é independente das expansões dos outros. Assim, é difícil caracterizar restrições da linguagem como, por exemplo, as ditadas pelos fenômenos de topicalização e *movimentos* em geral: o objeto aparece como expansão do VP, exceto se... ele já apareceu lá mais em cima na expansão da árvore... Isto não dá para fazer com CFG. Para estes problemas, que aliás são pervasivos na língua natural, Chomsky inicialmente propôs uma idéia mirabolante de transformações, as gramáticas transformacionais, que sofreu várias críticas ao longo de décadas, inclusive do próprio Chomsky. Uma das principais críticas foi a de que elas não eram realistas do ponto de vista computacional. Formalismos como GPSG e HPSG também usam CFGs como formalismo subjacente e apelam para um complexo sistema de unificação de *features* para modelar as restrições que as linguagem impõe, e a CFG por si não permite modelar.

Ora, como uma gramática de reescrita de árvores, as "regras" de uma TAG são já pequenas árvores (às vezes nem tão pequenas), e isto permite que se especifiquem gramáticas em que estas pequenas árvores elementares já definam

exatamente as estruturas de realização sintáticas possíveis e impeçam as que não o são. Uma tal árvore elementar, por exemplo, pode ser uma que tem um VP com um verbo transitivo, sem seu complemento, mas contendo mais acima o mesmo complemento na posição da topicalização. As possibilidades são enormes. De modo a organizar as idéias de aplicação das TAGs à formalização da sintaxe das linguagens naturais, Joshi então sugeriu adicionalmente alguns princípios de modelagem, como o do "*Domínio Estendido da Localidade das Dependências*" (*Extended Domain of Locality*) que diz que estas árvores elementares devem conter todos os constituintes obrigatórios de uma dada realização de estrutura de projeção. Por exemplo, uma árvore elementar para um verbo transitivo como núcleo de sentença já conteria toda a estrutura projetiva do verbo até a categoria sentencial, incluindo a posição dos argumentos (sujeito, objeto, onde quer que estejam naquela realização), traços, etc. Claro que isto já começa a atuar no sentido de impor um estilo de modelagem e tem conseqüências. Então, estes princípios não devem ser vistos como uma restrição ou obrigação, mas apenas mostram a versatilidade das TAGs comparadas a outros formalismos, e como elas podem ser usadas para contornar de forma lingüisticamente plausível e computacionalmente factível problemas de modelagem de língua natural que outros formalismos tem dificuldade de resolver.

Até este ponto, poderíamos dizer que uma derivação de uma TAG começa com uma árvore elementar e progride por substituição de nodos nas extremidades por outras árvores elementares até que se chegue à arvore sintática desejada. O processo seria semelhante ao que acontece com uma CFG, só que a reescrita é de árvores ao invés de strings. Na verdade, pode-se provar que este processo de substituição pode até ser simulado através de CFGs equivalentes, embora de forma um tanto quanto deselegante. Aliás, esta a noção de gramáticas de substituição por si já é anterior às TAGs.

O que faz das TAGs um formalismo diferenciado e com poder de modelagem realmente superior às CFGs é uma operação adicional definida por Joshi, Levi e Takahashi em artigo seminal de 1975, hoje chamada de *adjunção* (*tree adjunction* ou *tree adjoining*). Esta operação permite a inserção de uma árvore

elementar, durante o processo de derivação, no meio da árvore de derivação parcialmente construída, como que partindo ao meio um nodo interno da árvore de derivação. Isto é impossível com o mecanismo de substituição, que opera apenas nas extremidades da árvore de derivação, da mesma forma que nas CFGs. O uso sugerido por Joshi para modelagem de linguagem é o princípio de "Fatoração da Recursão" (Factoring of Recursion), complementando o do Domínio Estendido da Localidade. A fatoração da recursão, por exemplo, permite que na estrutura básica de realização de um constituinte (para a qual a gramática teria uma árvore elementar como visto anteriormente), sejam inseridas árvores elementares para elementos opcionais, de forma ilimitada, implementando o conceito de adjunção de Chomsky. É bom notar que são coisas diferentes: a adjunção de Chomsky é um conceito teórico lingüístico, enquanto que a das TAGs é simplesmente uma operação de derivação do formalismo. No entanto, a semelhança do nome não é mera coincidência, visto que uma permite implementar formalmente dentro das TAGs a outra.

Uma descrição mais completa e didática certamente fugiria às características de uma entrevista e de nosso objetivo aqui. Há uma vasta literatura sobre o formalismo das TAGs, seu uso em formalização lingüística, suas vantagens e desafios. Também há outros elementos importantes das TAGs como recursos de modelagem lingüística que foram adicionados ao longo do tempo, como o da lexicalização, que permite a modelagem da linguagem como uma extensão do léxico: nas *TAGS lexicalizadas* (*LTAGS, Lexicalized TAGs*), cada árvore elementar é vista como a projeção de uma palavra da linguagem. As *LTAGS baseadas em features* (*FB-LTAGS, Feature-Based LTAGS*) permitem a especificação de restrições que não são apropriadas para o nível de reescrita tradicional, como por exemplo, regras de concordância de gênero e número. E assim vai ...

ReVEL – Qual foi a finalidade das TAGs? Por que e como se percebeu a necessidade da criação desse formalismo sintático?

Prolo – Joshi pesquisava o processamento da linguagem natural desde os seus primórdios. Trabalhou na construção de um dos primeiros programas para análise sintática de sentenças de linguagem natural, há muito, muito tempo atrás, no início da década de 60, em um projeto coordenado por Zellig Harris. Aliás, por curiosidade, Noam Chomsky e Lila Gleitman trabalhavam no mesmo projeto. Desde aquela época conhecia as limitações dos formalismos tradicionais, como CFG para descrição de sintaxe de linguagem natural. Não é surpresa então que Joshi procurasse por um formalismo que tivesse expressividade adequada para a descrição das linguagens naturais, em particular no ponto de vista sintático. A primeira publicação foi o artigo de 1975 que mencionei acima. Seguiram-se uma série de desenvolvimentos que apoiaram a inserção do formalismo como alternativa à implementação computacional da sintaxe da linguagem, como os princípios de uso que mencionei, a lexicalização, inserção de *features*, o projeto *XTAG* de construção da gramática de larga cobertura do inglês. Um aspecto importante foi a colaboração de lingüistas, em particular a partir dos trabalhos com Tony Kroch.

ReVEL – Conte-nos como foi seu trabalho na universidade da Pennsylvania, junto ao criador do formalismo, Aravind Joshi.

Prolo – Quando fui fazer meu doutorado na University of Pennsylvania com Aravind Joshi, em 1995, ele já tinha mais de 65 anos. Hoje, com quase 80, se você for visitar as páginas do departamento de computação da Universidade, vai ver que ele ainda está ativo. Note que não é como membro *emeritus faculty*. É professor regular mesmo. É uma figura fascinante, ávido pelo trabalho de pesquisa e com uma capacidade de organização e mobilização incrível. Naquela época, após ter sido peça fundamental na construção do departamento, além de dedicar-se ao ensino e pesquisa, ele dirigia o IRCS (Institute for Research in Cognitive Science, http://www.ircs.upenn.edu/), que ajudou a fundar como um *NSF-STC Center for Research in Cognitive Science*, com verba do NSF

americano. Joshi sempre enfatizou a importância do trabalho cooperativo, a existência de uma forte comunidade integrada de pesquisa. O IRCS consolidava a tradição de integração de pesquisadores nas áreas de Lingüística, Lógica Matemática, Filosofia, Computação e Neurociência da Universidade em torno da pesquisa em ciência cognitiva. Esta foi certamente o melhor ganho que tive. Você tinha a oportunidade de ver, ouvir, interagir com pesquisadores da estatura de Henry e Lila Gleitman, Tony Kroch, Mark Liberman, Robin Clark, Mitch Marcus, Fernando Pereira, Martha Palmer, Mark Steedman, Bonnie Webber, Jean Gallier, Michael Kearns, John Trueswell, Scott Weinstein. Isto para citar alguns, além de Joshi é claro. Isto criava um ambiente extremamente dinâmico, com pesquisadores visitantes de todas as partes do mundo, bastante procurado por alunos para doutorado e pós-doutorado. Uma oportunidade de aprendizado realmente muito boa.

Entre outros círculos de interação existentes, Joshi mantinha um grupo particular chamado de *XTAG Group*, que embora aberto a quem estivesse interessado, tinha como membros cativos, com alguma rotatividade ao longo do tempo, talvez uns quinze a vinte alunos, professores e colaboradores, adeptos do formalismo das TAGs, trabalhando nos aspectos teóricos computacionais, lingüísticos e psicolingüísticos, na implementação de recursos computacionais, implementação de gramáticas, como a para a língua inglesa que mencionei anteriormente. Havia também gente trabalhando no uso de TAGs para modelagem da estrutura genética. Toda semana, em geral na quinta-feira de manhã, havia um encontro, o XTAG meeting, que era sagrado para ele, onde se discutiam os aspectos acima citados, progressos feitos, *practice talks* de artigos aceitos em conferências. Isto além da interação com outros grupos de interesses correlatos.

Mais do que propriamente a orientação do doutorado, Joshi oportunizou e sempre incentivou a participação ativa de grupos, o convívio e discussão com pessoas que tinham algo relevante a dizer.

ReVEL – As TAGs sempre estiveram muito relacionadas à figura de Aravind Joshi e à universidade da Pennsylvania. Existem trabalhos de descrição sintática e *parsing* com as TAGs sendo realizados fora dos Estados Unidos?

Prolo – Sim. Primeiro, em termos de diversidade lingüística, existem trabalhos de formalização de linguagem natural em TAGs para várias línguas, além do Inglês, como Francês, Alemão, Coreano, Japonês, Português. Quanto à sua pergunta propriamente dita das comunidades, há ou houve grupos e indivíduos com pesquisa relevante, mais ou menos ativos, em universidades e centros de pesquisa em vários países. A citar alguns que me vêm à cabeça: França (e.g., Paris 7, INRIAs), Alemanha (Saarbrucken, DFKI), Inglaterra (Sussex, Brighton, Edinburgh), Itália (Giorgio Satta), Holanda (Mark-Jan Nederhof), Japão (Keio), Brasil, Canadá (Chung-Hye Han, Anoop Sarkar)... Em geral, a origem ou desenvolvimento destes pólos têm a ver com a Penn e o IRCS, a partir de pesquisadores que fizeram doutorado ou pós-doutorado com o Joshi ou na Penn, ou que se interessaram pelas TAGs e passaram a colaborar com o grupo do Joshi. Alguns destes pólos são voltados à computação, outros à lingüística. Ah, não se pode esquecer do Workshop on Tree-Adjoining Grammars (ou Tag+), um evento bi-anual que este ano vai para sua nona edição este ano. As anteriores foram em Dagstuhl (1990), Philadelphia (1992), Paris (1994), Philadelphia (1998), Paris (2000), Veneza (2002), Vancouver (2004), Sydney (2006).

ReVEL – O senhor poderia sugerir algumas leituras para quem deseja começar ou prosseguir seus estudos no formalismo das TAGs?

Prolo – Vou sugerir três referências e um *site* da Internet. A primeira referência é um texto que eu sempre recomendo, um *overview* bastante abrangente. A segunda é um artigo do Joshi como palestrante convidado, dando um panorama sobre os desenvolvimentos em TAGs. A terceira é relatório técnico descrevendo com detalhes a implementação da gramática de larga

cobertura do projeto XTAG para a língua inglesa usando TAGs. O *site* é o do projeto *XTAG*.

Joshi, Aravind K. and Yves Schabes. 1997. Tree-Adjoining Grammars. In A. Salomaa and G. Rozenberg, editors, *Handbook of Formal Languages*, volume~3. Springer-Verlag, Berlin, pages 69—123.

Joshi, Aravind K. 2001. The XTAG project at Penn. *Proceedings of the 7th International Workshop on Parsing Technologies (IWPT-2001)*, Beijing, China. Invited speaker.

The XTAG Research Group. 2001. *A Lexicalized Tree Adjoining Grammar for English*. Technical Report IRCS 01-03, University of Pennsylvania. Disponível em ftp://ftp.cis.upenn.edu/pub/xtag/release-2.24.2001/tech-report.pdf.

http://www.cis.upenn.edu/~xtag