BrGram: uma gramática computacional de um fragmento do português brasileiro no formalismo da LFG*

Leonel F. de Alencar[1]

[1]Programa de Pós-Graduação em Linguística – Universidade Federal do Ceará (UFC)
Av. da Universidade, 2683 – 60.020-181 – Fortaleza – CE – Brazil

leonel.de.alencar@ufc.br

Abstract. *This paper deals with BrGram, a grammar for deep syntactic parsing of a fragment of Brazilian Portuguese. Presently, this grammar, implemented in XLE, the state of the art in computational grammar development within the LFG generative framework, covers an extensive subset of Brazilian Portuguese. BrGram aims at bridging the gap in Brazil between research in formal syntax and natural language processing.*

Resumo. *Este artigo trata da BrGram, uma gramática para a análise sintática automática profunda de um fragmento do português brasileiro. No momento, essa gramática, implementada no sistema XLE, que representa o estado da arte no desenvolvimento de gramáticas computacionais no modelo gerativo da LFG, abrange um amplo subconjunto do português brasileiro. A BrGram almeja superar o fosso, vigente no Brasil, entre a pesquisa em sintaxe formal e o processamento automático da linguagem natural.*

1. Introdução

Todo falante do português padrão reconhecerá como gramaticalmente bem formados exemplos do tipo de (1), (2) e (3) e mal formados exemplos como (4). As intuições de um falante sobre as sentenças de sua língua, porém, não se restringem a julgamentos de gramaticalidade. Referem-se, também, à estrutura sintática e às propriedades semânticas dessas construções, como o fato de que o sujeito de (1) é a expressão em itálico ou o fato de que o sujeito de (2) é parafraseável por *nós*, ao passo que o de (3), superficialmente idêntico, se parafraseia por *eles*.

(1) *Todos nós três agrônomos* temos rejeitado a proposta.

(2) Os três temos rejeitado a proposta.

(3) Os três têm rejeitado a proposta.

(4) *Os três temos rejeitando a proposta.

No presente artigo, partimos da carência no Brasil, no cenário atual da linguística computacional e processamento automático de linguagem natural (doravante PLN), de esforços voltados para a construção de uma gramática computacional do português do Brasil que tenha ampla cobertura e se baseie nas pesquisas mais recentes em gramática gerativa, especialmente no âmbito dos modelos não derivacionais, representados, sobretudo, pela Gramática Léxico-Funcional (doravante LFG, do inglês *Lexical-Functional Grammar*) (FALK, 2001) e a HPSG (SAG; WASOW; BENDER, 2003). Esses modelos oferecem os recursos necessários para a descrição formal de fenômenos gramaticais não triviais, como os exemplificados acima, e têm tido sua eficácia comprovada na construção de gramáticas computacionais de ampla cobertura para diversas línguas (MÜLLER, 2010). Visamos, portanto, ao contribuir para preencher essa lacuna, eliminar o descompasso que se observa, entre nós, no âmbito da

* Este trabalho, desenvolvido com o apoio da CAPES, Processo nº BEX 10175/12-1, constitui um desdobramento de Alencar (2013).

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 183–188,
Fortaleza, CE, Brazil, October 21–23, 2013. ©2013 Sociedade Brasileira de Computação

análise sintática automática, em relação aos EUA e muitos países da Europa, Ásia etc., onde a implementação de gramáticas computacionais para novas línguas e o aperfeiçoamento das gramáticas existentes continua na ordem do dia das pesquisas tanto na linguística quanto na informática.

Para tanto, descrevemos aspectos da BrGram, uma gramática de um fragmento do português brasileiro implementada no sistema XLE, o qual representa o estado da arte no desenvolvimento de gramáticas computacionais no modelo da LFG. Uma comparação da nossa proposta com as iniciativas mais importantes de *parsing* profundo do português, como o LX-Parser (SILVA *et al.*, 2010), LXGram (BRANCO; COSTA, 2010) e o *parser* do projeto VISL (BICK, 2000) fica de fora do presente artigo por motivo de espaço. Independentemente dessa comparação, a nossa proposta é original, na medida em que adota o modelo da LFG para formalização da gramática e o sistema XLE para implementação, o que ainda não foi feito, com a mesma extensão, para o português (MÜLLER, 2010; MISTICA *et al.*, 2012).

2. O formalismo da LFG e o sistema XLE

Atualmente, no *parsing* sintático profundo baseado em regras, vários formalismos gramaticais se destacam (LJUNGLÖF; WIRÉN, 2010). Todos dispõem de sistemas bastante sofisticados para o desenvolvimento de gramáticas computacionais e têm sido usados na implementação de *parsers* de ampla cobertura para diversas línguas (MÜLLER, 2010). Nesse quadro, quando consideramos tanto as implicações desses formalismos para a linguística teórica quanto as suas aplicações nas tecnologias da linguagem natural, a HPSG e a LFG sobressaem. Uma extensa gramática do português europeu, a LXGram, desenvolvida no formalismo da HPSG, está disponível. A LFG, por sua vez, tem sido usada como base para a elaboração de gramáticas computacionais de mais de duas dezenas de línguas, especialmente no âmbito do Projeto ParGram, voltado para a descrição de línguas tipologicamente diversas, a partir de um aparato conceitual comum (PARGRAM, 2012). O português, contudo, como ressaltamos, ainda não dispõe de uma gramática de maior extensão baseada na LFG.

A característica mais marcante de *parsers* baseados na LFG é a geração de vários níveis de representação sintática para as sentenças gramaticais, minimamente a estrutura de constituintes (*c-structure*) e a estrutura funcional (*f-structure*) (FALK, 2001), exemplificadas, respectivamente, na Figura 1 e na Figura 2. Esse segundo tipo de estrutura representa uma dimensão mais abstrata da informação gramatical, codificando a estrutura de predicado e argumentos, as relações gramaticais sujeito (SUBJ), objeto direto (OBJ) etc. bem como as propriedades morfossintáticas dos diferentes constituintes, como pessoa, número, gênero, tempo, modo etc. Por abstrair da estrutura superficial das sentenças, a estrutura funcional é utilizável como interlíngua em sistemas de tradução automática (CROUCH *et. al.*, 2011), entre outras aplicações.

No momento, os três principais ambientes de desenvolvimento de gramáticas computacionais em LFG, os quais oferecem as correspondentes facilidades de *parsing* em interfaces amigáveis, são, por um lado, o Xerox Linguistic Environment (XLE) (CROUCH *et al.*, 2011), e, por outro, o XLFG5 (CLÉMENT, 2010) e o LFG Parser (MINOS, 2013). Enquanto o primeiro visa à implementação tanto de *parsers* quanto de geradores para aplicações de escala industrial, os dois últimos, dotados de bem menos recursos, focam o ensino da LFG e a pesquisa linguística nesse arcabouço. Desse modo, optamos pelo XLE para implementação de nossa gramática.

3. Aspectos da BrGram

Na construção da BrGram, utilizamos a metodologia consagrada para elaboração de gramáticas formais de uma língua natural, que consiste em definir, inicialmente, um fragmento cobrindo um leque inicial de fenômenos, fragmento esse ampliado sucessivamente, resultando em versões cada vez mais abrangentes da gramática

(FRANCEZ; WINTNER, 2012).

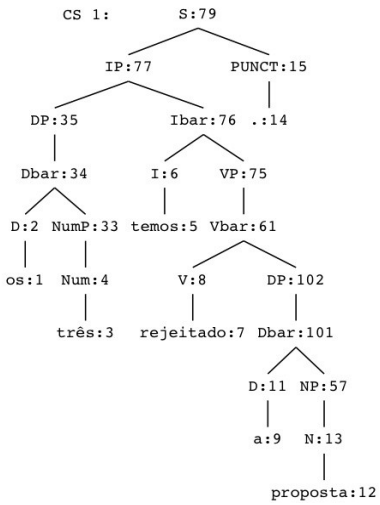

Figura 1. Exemplo de estrutura de constituintes gerada no XLE conforme a BrGram

Dada a complexidade de uma língua natural, o desenvolvimento de uma gramática voltada para o *parsing* sintático profundo é algo correspondentemente complexo, sobretudo quando se utilizam formalismos gramaticais da envergadura da LFG e da HPSG. A experiência dos projetos internacionais mais importantes nessa área, que alcançaram resultados significativos no *parsing* de textos autênticos em línguas como inglês, alemão, francês etc., demonstra que se trata de tarefa que necessariamente consome vários anos e implica um trabalho de equipe. No entanto, uma vantagem crucial da utilização de um desses arcabouços na implementação de uma nova gramática é que, no tratamento de grande parte dos fenômenos, não se precisa "reinventar a roda", bastando transferir ou adaptar, para a nova língua em questão, soluções anteriormente propostas para outras línguas.

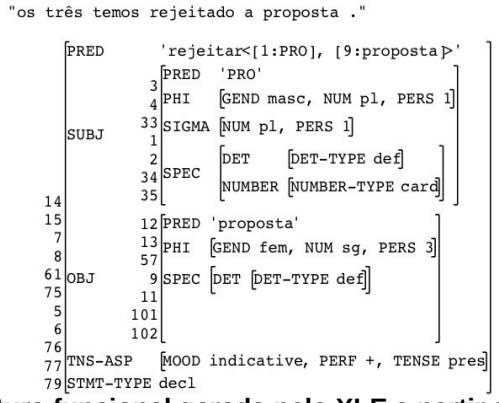

Figura 2. Estrutura funcional gerada pelo XLE a partir da BrGram

Em seu atual estágio, fruto de esforços aplicados no projeto, de forma não contínua, ao longo de um ano, a BrGram cobre um leque não trivial de fenômenos gramaticais do português, como já sugerem a Figura 1 e a Figura 2. Essa última utiliza, essencialmente, a notação do projeto PARGRAM (KING, 2004), que se estabeleceu como um padrão internacional para gramáticas desenvolvidas no âmbito do XLE.

O Curupira (MARTINS; OTHERO, 2012), a exemplo de outras iniciativas brasileiras anteriores de implementação de analisadores sintáticos automáticos do português levadas a cabo por equipes mais vinculadas à computação, não leva em conta os esforços mais recentes no âmbito da linguística gerativa para a descrição do

português, recorrendo, antes, a abordagens tradicionais. A BrGram, pelo contrário, dada a sua fundamentação numa teoria linguística como a LFG, toma exatamente esses trabalhos como ponto de partida, lançando mão, também, de descrições formais de outras línguas, notadamente o francês, cuja estrutura morfossintática, na qualidade de língua neolatina, apresenta significativas analogias com a portuguesa. Nesse sentido, há que destacar dois trabalhos. O primeiro é a implementação de um fragmento do francês no formalismo da LFG, realizada por Schwarze (1998), que nos orientou na escolha tanto do leque de fenômenos a serem cobertos inicialmente por nossa gramática do português quanto da solução a adotar no tratamento de várias questões gramaticais. O segundo trabalho é a modelação em termos da teoria X-barra, proposta por Othero (2009), de um considerável fragmento da sintaxe do português do Brasil no formalismo da CFG.

Com base nas sentenças geradas pela gramática francesa de Schwarze (1998), definimos um conjunto análogo de sentenças do português a serem analisadas pela BrGram em sua primeira etapa de desenvolvimento. Alguns exemplos:

A rainha é esperada. A fada vê o cavaleiro passar. A rainha quer dançar. A fada quer esperar o cavaleiro. A rainha começa a dançar. A rainha para de dançar. A rainha sabe que o cavaleiro quer dançar. A rainha quer que o cavaleiro saiba que a menina espera.

Esse conjunto foi em seguida bastante ampliado, de modo a incluir construções como (1), (2) e (3).

Dentre os fenômenos gramaticais modelados pela BrGram, destacamos: (i) concordância nominal, (ii) concordância verbal tanto sintática quanto semântica por meio da distinção entre traços phi e sigma (ALENCAR, 2013), (iii) valência verbal e diáteses verbais, (iv) auxiliares *ter* + particípio e *estar* + gerúndio, (v) verbos modais e aspectuais de controle do sujeito ou do objeto, (vi) recuperação do sujeito "lógico" implícito (como *nós* e *eles* em (2) e (3)), (vii) projeções funcionais no interior do sintagma determinante (DP), como sintagma quantificador, sintagma numeral e sintagma possessivo (OTHERO, 2009).

Como a LFG é um formalismo baseado em restrições, o XLE, ao carregar a BrGram, não apenas analisa as sentenças gramaticais do fragmento, como também acusa, no caso de construções agramaticais como (4), violações dessas restrições, afetando a concordância, a valência, o modo, a forma verbal etc.

4. Conclusão

Este trabalho tratou da BrGram, uma gramática computacional para um fragmento do português brasileiro. Com essa gramática, implementada no sistema XLE, o estado da arte no *parsing* com base no modelo gerativo da LFG, pretendemos preencher lacuna na linguística computacional e PLN no Brasil, onde, diferentemente dos grandes centros internacionais de pesquisa nessas áreas, não se tem ultimamente focado a análise sintática automática profunda.

Por outro lado, a BrGram, que, no momento, constitui apenas um protótipo, não obstante abranger um leque significativo de fenômenos gramaticais do português, se pretende como uma alternativa futura às iniciativas europeias de *parsing* sintático do português, LXGram/LX-Parser e VISL. Enquanto essa última iniciativa não leva em conta a gramática gerativa, a BrGram oferece, na análise sintática automática do português baseada em regras, a perspectiva da LFG, teoria gerativa irmã, porém distinta da HPSG, em que se fundamenta a LXGram.

Referências

ALENCAR, L. F. de. Modelação computacional de padrões variáveis de concordância

em português. *Revista de Estudos da Linguagem*, Belo Horizonte, n. 21, vol. 1, p. 43-110, 2013. Disponível em: <http://relin.letras.ufmg.br/revista/upload/2112-ALENCAR.pdf> Acesso em: 17. mai. 2013.

BICK, E. *The parsing system "Palavras"*: automatic grammatical analysis of Portuguese in a Constraint Grammar framework. 2000. Tese (Dr. phil.) – Department of Linguistics, University of Århus, Århus, Dinamarca, 2000. 505 p. Disponível em: <beta.visl.sdu.dk/pdf/PLP20-amilo. ps.pdf> Acesso em: 27 out. 2009.

BRANCO, A.; COSTA, F. LXGram: A Deep Linguistic Processing Grammar for Portuguese. In: PARDO, T. A. S. *et al.* (Eds.). INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, n. 9, 2010, Porto Alegre. *Proceedings*... Berlin; Heidelberg: Springer, 2010. p. 86-89.

CLÉMENT, L. *XLFG5*. [S.l.]: [s.n.], 2010. Disponível em: <http://nlp.ioperm.org/lfg-parser.html> Acesso em: 19. abr. 2013.

CROUCH, D. *et al*. *XLE Documentation*. Palo Alto: Palo Alto Research Center, 2011. Disponível em: <http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html> Acesso em: 5. nov. 2012.

FALK, Y. N. *Lexical-functional grammar*: an introduction to parallel constraint-based syntax. Stanford, CSLI Publications, 2001.

FRANCEZ, N.; WINTNER, S. *Unification grammars*. Cambridge: CUP, 2012.

KING, T. H. *Starting a ParGram Grammar*. 2004. Disponível em: <http://www2. parc.com/isl/groups/nltt/xle/doc/PargramStarterGrammar/starternotes.html> Acesso em: 10. nov. 2012.

LJUNGLÖF, P. ; WIRÉN, M. Syntactic parsing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). *Handbook of Natural Language Processing*. 2. ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. p. 59-91.

MARTINS, R.; OTHERO, G. A. Parsing do português. In: ALENCAR, L. F. de; OTHERO, G. A. (Org.). *Abordagens computacionais da teoria da gramática*. 1. ed. Campinas: Mercado de Letras, 2012, p. 99-126.

MINOS, P. *LFG Parser*. [S.l.]: [s.n.], 2013. Disponível em: <http://nlp.ioperm.org/lfg-parser.html> Acesso em: 19. abr. 2013.

MISTICA, M. *et al*. LFG Bibliography. [S.l.]: [s.n.], 2012. Disponível em: <http://ww2.cs.mu.oz.au /~mmistica/ bibliography.html> Acesso em: 21. nov. 2012.

MÜLLER, S. *Grammatiktheorie*. Tübingen: Stauffenburg, 2010.

OTHERO, G. A. *A gramática da frase em português*: algumas reflexões para a formalização da estrutura frasal em português. Porto Alegre: Edipucrs, 2009. Disponível em:< http://www. pucrs. br/edipucrs/gramaticadafrase.pdf> Acesso em: 02.08.2010.

PARGRAM/ParSem: An international collaboration on LFG-based grammar and semantics development. [S.l.]: [s.n.], 2012. Disponível em: <http://pargram.b.uib.no/> Acesso em: 2. nov. 2012.

SAG; WASOW; BENDER. *Syntactic theory*: a formal introduction. 2. ed. Stanford: CSLI, 2003.

SCHWARZE, C. *Lexikalisch-funktionale Grammatik*: eine Einführung in 10 Lektionen mit französischen Beispielen. 2. ed. Konstanz: Fachgruppe Sprachwissenschaft der Universität Konstanz, 1998.

SILVA, J. *et al*. Out-of-the-Box Robust Parsing of Portuguese. In: PARDO, T. A. S. *et*

al. (Ed.). INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, n. 9, 2010, Porto Alegre. *Proceedings*... Berlin; Heidelberg: Springer, 2010. p. 75-85.