Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

A Implementação de uma Minigramática do Português Brasileiro sob a Perspectiva da LFG

Daniel Soares[1], Francisco Nogueira[2], Leonel Figueiredo de Alencar[3]

[1]Departamento de Letras Vernáculas – Universidade Federal do Ceará (UFC)
Fortaleza, Brasil.

[2]Departamento de Letras Vernáculas – Universidade Federal do Ceará (UFC)
Fortaleza, Brasil.

[3]Departamento de Letras Estrangeiras – Universidade Federal do Ceará (UFC)
Fortaleza, Brasil.

{danielfbrasil,herrnogueira}@gmail.com, leonel.de.alencar@ufc.br

Abstract. *This paper describes the implementation of a grammar fragment of Brazilian Portuguese (BP) in the Lexical-Functional Grammar (LFG) formalism using the XLE system. This fragment analyses, among other phenomena, verbal and nominal agreement, adjective syntax, passive, verbal valence, complement clauses, prepositional phrases functioning as adjuncts, grade adverbs and control verbs. For the evaluation of this grammar, a parser was compiled in XLE and applied to a positive and a negative test set. The former contains 72 grammatical sentences, all of which were correctly analyzed. The latter contains 88 non-grammatical sentences, of which none were analyzed.*

Resumo. *Este artigo descreve a implementação de um fragmento de gramática do Português Brasileiro (PB) no formalismo da Gramática Léxico-Funcional (LFG), usando o sistema XLE. Essa minigramática do PB analisa, entre outros, os fenômenos de concordância verbal e nominal, a sintaxe de adjetivos, a passiva, a valência verbal, os complementos oracionais com 'que' e 'se', a função de adjunto exercida por sintagmas preposicionais, advérbios de gradação e verbos de controle. Para avaliação dessa gramática, foram testadas 72 sentenças gramaticais, das quais todas foram analisadas, e 88 sentenças agramaticais, das quais nenhuma foi analisada.*

1. Introdução

O presente trabalho descreve a construção de uma minigramática computacional para o português brasileiro (doravante PB). Esse fragmento de gramática do PB segue o formalismo da Gramática Léxico-Funcional (LFG, do inglês *Lexical Functional Grammar*) (KAPLAN e BRESNAN, 1982; BUTT, 1999; FALK, 2001; DALRYMPLE, 2005) e foi implementado no sistema XLE (*Xerox Linguistic Environment*), que constitui o estado da arte para a implementação e teste de gramáticas nesse formalismo (CROUCH *et al*., 2011).[1]

1 O XLE é a base do Projeto de Gramáticas Paralelas (do inglês *Parallel Grammar Project)*, que

Para nossa implementação, nos baseamos no fragmento de gramática do francês desenvolvido no capítulo 6 de Schwarze e Alencar (2016), que segue o formalismo LFG/XLE, variante notacional da LFG implementada no ambiente XLE (ALENCAR, 2017). Esse fragmento implementa os seguintes fenômenos do francês, entre outros: (i) concordância verbal e nominal, (ii) pronomes clíticos com função de sujeito, (iii) sintaxe dos adjetivos, (iv) voz passiva, (v) passado composto, (vi) valência verbal, (vii) complementos oracionais com *que* e *si,* (viii) sintagmas preposicionais na função de adjuntos, (ix) advérbios de gradação e (x) verbos de controle.

Nosso artigo se estrutura em mais três seções, além desta introdução. Na seção 2, apresentamos os princípios da LFG. Na seção 3, descrevemos a construção de um fragmento da nossa minigramática do PB de acordo com o formalismo LFG/XLE. Na seção 4, apresentamos os resultados dessa implementação. Na seção 5, finalmente, expomos as considerações finais.

2. A Gramática Léxico-Funcional

A LFG surgiu como alternativa gerativa à Teoria Padrão Estendida de Chomsky (DAVIES, W. D. e DUBINSKY, S., 2004), refutando a existência de transformações sintáticas. Continuou sendo desenvolvida como alternativa não transformacional aos modelos chomskyanos posteriores, como o Minimalismo. Dalrymple (2005) afirma que a LFG é uma teoria da estrutura da língua e de como diferentes aspectos da estrutura linguística são relacionados.

A LFG se caracteriza, principalmente, por dois aspectos, distinguindo-se das vertentes gerativas transformacionais: (i) por estruturar ricamente o léxico, codificando as relações lexicais em vez de transformações ou operações nas árvores de estrutura sintagmática; e (ii) por assumir as funções gramaticais (sujeito, objeto, etc) como primitivos da teoria (cf. FALK, 2001).

Uma gramática léxico-funcional de uma língua particular se constrói minimamente a partir de duas especificações: regras sintagmáticas e entradas lexicais, ambas providas das chamadas *anotações funcionais* (FALK, 2001, p. 68; BUTT *et al*., 1999, p. 23).

Na LFG, a representação das informações acerca da estrutura sintática de uma língua particular consiste, por um lado, da estrutura de constituintes (doravante estrutura-c, do inglês *c-structure*) e, por outro, da estrutura funcional (doravante estrutura-f, do inglês *f-structure*)[2].

A estrutura-c é gerada diretamente do módulo das regras sintagmáticas, tendo itens lexicais como nós terminais, e é nessa estrutura que são codificadas "as relações de precedência, dominância e constituição, as quais definem a boa-formação gramatical das sentenças" (ALENCAR, 2004, p. 3). A estrutura-f, por outro lado, é gerada a partir da estrutura-c por meio das anotações funcionais das regras sintagmáticas e dos itens lexicais, em que cada nó da árvore projeta uma estrutura funcional. Trata-se, segundo Alencar (2004, 2017), de um nível mais abstrato de análise, representando aspectos

desenvolve gramáticas para o inglês, francês, alemão, norueguês, japonês, urdo etc (SULGER *et al*., 2013).

2 Kaplan e Bresnan (1982) afirmam que é impossível construir uma teoria da sintaxe apenas com um formalismo de regras sintagmáticas, uma vez que elas não dão conta da complexidade sintático-semântica das línguas naturais.

universais da linguagem e constituindo *input* para o processamento semântico.

A estrutura-f é representada por meio de uma matriz de atributos e valores (AVM, do inglês *atribute-value matrix*) (cf. FALK, 2001). As AVMs formalizam a noção de traço (*feature*)[3] e desempenham papel fundamental na descrição de línguas naturais, uma vez que um atributo de uma AVM pode ter como valor uma outra AVM, modelando, dessa forma, a recursividade das estruturas sintáticas das línguas naturais (ALENCAR, 2017, p. 356). Na seção seguinte, descrevemos a implementação da nossa minigramática LFG/XLE para o PB.

3. A implementação de uma gramática léxico-funcional no XLE

A construção de gramáticas computacionais tem duas vantagens na descrição de línguas naturais: (i) o uso da gramática em aplicações tecnológicas, como em tradutores automáticos, programas de extração de informações, de perguntas e respostas etc; e (ii) a possibilidade de testar automaticamente a coerência interna e a adequação empírica das análises em conjuntos de dados em grande escala (ALENCAR, 2017, p. 356).

Além disso, a implementação de uma gramática léxico-funcional no XLE se aplica também à medição da complexidade de abordagens distintas de fenômenos gramaticais (ALENCAR, 2017, p. 355), ao permitir testar automaticamente hipóteses distintas sobre o mesmo fenômeno gramatical.

Com base na implementação de Schwarze e Alencar (2016), adaptamos nossa minigramática do PB para analisar, entre outras, sentenças como: *a fada é amável*, *o corajoso cavaleiro chega*, *a fada é esperada por um cavaleiro*, *o cavaleiro crê ver a fada* e *a fada pergunta ao cavaleiro se a rainha quer que a dama saiba que o anão é mau.*

As sentenças supramencionadas exemplificam os seguintes fenômenos implementados: (i) concordância de gênero e número entre determinantes, adjetivos e nomes; (ii) concordância entre sujeito e verbo; (iii) relação dos verbos com seus argumentos; (iv) sintaxe de adjetivos; (v) verbos de controle; (vi) sintaxe da passiva; e (vii) complementos oracionais com *que* e *se*.

Nossa minigramática consiste de 213 entradas lexicais, 8 regras, 29 estados, 44 arcos e 49 disjuntos. Essas informações indicam a complexidade espacial da gramática. O número de regras, estados, arcos e disjuntos são fornecidos após a leitura da gramática pelo sistema XLE.

Na figura 1, exemplificamos a estrutura-c:

3 Na LFG, um traço consiste de um atributo e seu valor (cf. ALENCAR, 2017, p. 358). Por exemplo, o item *fadas* possui as propriedades 'feminino' e 'plural'. Nesse caso, o atributo 'gênero' tem o valor 'feminino', enquanto o atributo 'número', o valor 'plural'.

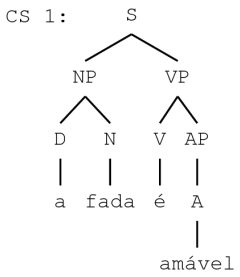

Figura 1. Exemplo de estrutura-c

Essa estrutura-c é gerada a partir de uma gramática de estrutura sintagmática (PSG, do inglês *phrase structure grammar*) com anotações funcionais. Em (1)[4], segue um fragmento das regras sintagmáticas que geram a estrutura-c da figura 1.

(1) S --> NP: (^ SUBJ)=! ; VP.
NP --> D N.
VP --> V AP: (^ XCOMP)=!.
AP --> A.

Essas regras codificam os constituintes das sentenças e as funções sintáticas exercidas pelos argumentos do verbo. O fenômeno de concordância, por outro lado, não é codificado nas regras sintagmáticas, mas no léxico. Os exemplos (2)-(5) exibem as entradas lexicais da sentença *a fada é amável*:

(2) a D * (^ GEN)=FEM
(^ NUM)=SG
(^ SPEC)=DEF.

(3) fada N * (^ PRED)='FADA'
(^ GEN)=FEM
(^ NUM)=SG.

(4) é V * (^ PRED)='SER<(^ SUBJ)(^ XCOMP)>'
(^ SUBJ)=(^ XCOMP SUBJ)
(^ SUBJ PERS)=3
(^ SUBJ NUM)=SG.

(5) amável A * (^ PRED)='AMÁVEL<(^ SUBJ)>'
(^ SUBJ NUM)=SG.

Codificadas as regras sintagmáticas e o léxico de nossa minigramática,

4 As abreviações usadas nesse exemplo são as seguintes: AP = sintagma adjetival, NP = sintagma nominal, VP = sintagma verbal, D = determinante, S = sentença, N = nome, V = verbo, GEN = gênero, NUM = número, PERS = pessoa, PRED = predicado, 3 = terceira pessoa, DEF = definido, FEM = feminino, MAS = masculino, SG = singular, SUBJ = sujeito e XCOMP = predicativo.

apresentamos, na figura 2, um exemplo de estrutura-f, gerada automaticamente pelo XLE:

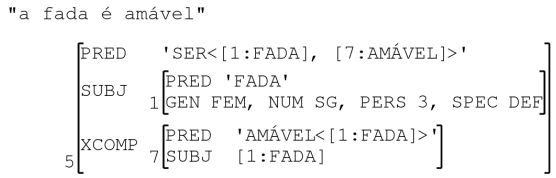

Figura 2. Exemplo de estrutura-f

Nessa estrutura-f, a sentença consiste de um SUBJ, um PRED e um XCOMP. O SUBJ tem o significado lexical 'FADA', é definido e suas propriedades de concordância são FEM, SG e PERS 3. O verbo tem o significado lexical 'SER', sua valência exige um SUBJ e um XCOMP, realizados, respectivamente, pelas AVMs 1 e 7, cujos predicados são 'FADA' e 'AMÁVEL'. O XCOMP tem o significado lexical 'AMÁVEL' e sua valência exige um SUBJ, realizado por 'FADA'.

Para avaliar nossa minigramática, criamos um conjunto-teste positivo e um conjunto-teste negativo (cf. BUTT, 1999, p. 212). O conjunto-teste positivo consiste de um corpus com sentenças gramaticais, ou seja, sentenças que a gramática deve analisar e gerar. Conforme Alencar (2017, p. 361), "o conjunto-teste positivo define o recorte gramatical implementado". Esses conjuntos-teste foram construídos a partir de uma adaptação dos conjuntos-teste de Schwarze e Alencar (2016).

Em relação à construção do conjunto-teste positivo, as sentenças foram adaptadas de acordo com os fenômenos gramaticais em análise. Por exemplo, a sentença em (6), tradução nossa do francês *cette dame est dans la chambre,* não representa a estrutura do sintagma preposicional em análise na língua francesa. A sentença adaptada em (7), por sua vez, representa com exatidão a estrutura do sintagma preposicional *dans la chambre*.

(6) esta senhora está no quarto
(7) esta senhora está sob a ponte

O conjunto-teste negativo consiste de um corpus com sentenças agramaticais, que não devem ser analisadas e geradas pela gramática, evitando dessa forma hipergeração. Esse conjunto-teste também é uma adaptação da minigramática do francês. Vale ressaltar que em cada sentença do conjunto-teste negativo é infringida apenas uma regra por vez. Dessa forma, verifica-se com exatidão o tipo de má formação da sentença[5].

Para essa implementação, o conjunto-teste positivo consiste de 72 sentenças gramaticais, enquanto o conjunto-teste negativo de 88 sentenças agramaticais. Em

5 No formalismo LFG, há três condições de boa formação da sentença: (i) completude, (ii) coerência e (iii) unicidade (cf. KROEGER, 2004, p. 20). A primeira condição pressupõe que a estrutura-f de uma sentença contenha todas as relações gramaticais exigidas pelo PRED(icado). A segunda pressupõe que a estrutura-f de uma sentença não contenha nenhuma relação argumental não exigida pelo PRED. A terceira, finalmente, pressupõe que "nenhuma relação argumental deve ser atribuída mais de que uma vez em uma única estrutura-f" (KROEGER, 2004, p. 20).

seguida, apresentamos nossos resultados em relação aos testes realizados.

4. Resultados

A avaliação da gramática, realizada por meio do conjunto-teste positivo e do conjunto-teste negativo, foi satisfatória.

Nossa minigramática analisa corretamente todas as 72 sentenças gramaticais do conjunto-teste e não analisa nenhuma das 88 sentenças agramaticais do conjunto-teste negativo. Em (8)-(15), apresentamos um recorte do teste positivo:

(8) A fada é amável. (1 0.007 15)
(9) O corajoso cavaleiro chega. (1 0.004 15)
(10) A fada é esperada por um cavaleiro. (1 0.007 42)
(11) O cavaleiro crê ver a fada. (1 0.004 22)
(12) A fada pergunta ao cavaleiro se a rainha quer que a dama saiba que o anão é mau. (1 0.009 61)
(13) A fada vê o cavaleiro passar. (1 0.005 21)
(14) A fada pede ao cavaleiro para combater o gigante. (1 0.007 34)
(15) A rainha exige que a fada espere o cavaleiro. (1 0.006 31)

As informações que se seguem a cada sentença entre parênteses são geradas pelo XLE após análise do arquivo de teste. A primeira informação indica o número de análises da sentença. Espera-se que, quando há ambiguidade, a sentença tenha mais de uma análise. Nos casos supra, não identificamos ambiguidades e, por isso, o resultado é satisfatório. A segunda informação indica o tempo de processamento para análise da sentença. E, finalmente, a terceira informação especifica o número de *subtrees*[6].

Em (16)-(23), apresentamos um recorte do conjunto-teste negativo:

(16) A fada é sob a ponte. (0 0.006 24)
(17) O cavaleiro brancos chega. (0 0.004 7)
(18) A fada é esperado por um cavaleiro. (0 0.007 42)
(19) O cavaleiro crê vê a fada. (0 0.003 0)
(20) O gigante pergunta para a rainha se a fada espere. (0 0.006 31)
(21) A fada quer espera o cavaleiro. (0 0.004 0)
(22) A fada vê passa o cavaleiro. (0 0.003 0)
(23) A fada pede ao cavaleiro a combater o gigante. (0 0.005 34)

Como observado acima, nenhuma sentença é analisada pela gramática. Uma vez que essas sentenças são agramaticais em PB e nossa minigramática não as analisa, o resultado do teste negativo também é satisfatório.

Esses testes foram realizados em uma máquina com sistema operacional Linux Ubuntu 16.04, 64 bit, processador Intel® Celeron (R) CPU N2830 @ 2.16GHz × 2, com memória de quatro *gigabytes* (4GB).

6 O número de *subtrees* dá ao implementador uma indicação da complexidade do sistema de regras. Por exemplo, quando sentenças muito simples aparecem com um número alto de *subtrees,* isso é uma indicação de que há algo de errado com a escrita das regras (BUTT, 1999, p. 167).

5. Considerações finais

Apresentamos uma minigramática do PB no formalismo LFG/XLE, implementada a partir da adaptação para o PB da gramática do francês do capítulo 6 de Schwarze e Alencar (2016). Apesar de constituir um fragmento, nossa minigramática analisa sentenças relativamente complexas, como, por exemplo, sentenças passivas, com verbos de controle e/ou estruturas de complementação oracional. Os resultados da avaliação da gramática em um conjunto-teste positivo de 72 sentenças gramaticais e um conjunto-teste de 88 sentenças agramaticais foram satisfatórios.

Disponibilizamos nossa gramática *on-line* de forma livre, como uma alternativa à BrGram, de Alencar (2013), e à gramática de Santos (2014), as duas propostas alternativas recentes de gramáticas do PB no formalismo LFG/XLE[7]. Uma gramática mais antiga do PB nesse formalismo, e não disponível livremente, é o fragmento de Alencar (2004), que foca as estruturas oracionais de complementação verbal. Como a nossa proposta se relaciona com as duas alternativas mais recentes? Nossa gramática constitui, ao lado da de Santos (2014), a primeira gramática do PB no formalismo LFG/XLE com código aberto e livremente disponível para *download.* A vantagem de nossa gramática em relação à de Santos (2014) é que, apesar de menos abrangente, é mais adequada para utilização em cursos introdutórios sobre a LFG e o sistema XLE, por ser mais simples e seguir de perto o manual introdutório de Schwarze e Alencar (2016). A BrGram (ALENCAR, 2013) é também mais abrangente que o nosso fragmento, mas não está disponível *on-line* e apresenta um nível de complexidade sintática que a torna inadequada para iniciantes.

O nosso próximo passo é a adaptação da gramática final do francês de Schwarze e Alencar (2016), que é a gramática do capítulo 8, a qual integra um componente morfológico de estados finitos para análise de um conjunto de verbos da gramática, componente esse desenvolvido no capítulo 7 desse livro. Em seguida, adaptaremos a gramática do francês de Alencar (2017), a qual representa um avanço em relação à gramática final de Schwarze e Alencar (2016). Finalmente, expandiremos a cobertura de nossa gramática para cobrir fenômenos do PB sem correspondência nos fragmentos do francês de Schwarze e Alencar (2016) e Alencar (2017). Com isso, acreditamos contribuir, por um lado, para a difusão dos estudos e pesquisas sobre o desenvolvimento de gramáticas computacionais da língua portuguesa e, por outro, para tornar mais acessível o formalismo LFG/XLE para estudantes e pesquisadores do português.

Referências

Alencar, L. F. de. (2004), "Complementos verbais oracionais - uma análise léxico-funcional. In: Revista Lingua(gem), Santa Maria, v.1, n.1, p. 173-218.

Alencar, L. F. de. (2013), "BrGram: uma gramática computacional de um fragmento do português brasileiro no formalismo da LFG. In: Brazilian Symposium In Information And Human Language Technology – Stil, 9., 2013. Fortaleza. Proceedings. Fortaleza: Sociedade Brasileira de Computação. p. 183-188.

7 Essa minigramática léxico-funcional do PB implementada no sistema XLE está disponível em https://github.com/DanielFBrasil/lfg-portuguese-grammar.

Alencar, L. F. de. (2017), "Uma implementação computacional de construções verbais perifrásticas em francês", In: Alfa, São Paulo, v.61, n.2, p.351-380.

Butt, M., King, T. H., Niño, M., Segond, F. (1999), A grammar writer's cookbook, Stanford: CSLI publications.

Crouch, D. et al. (2011), XLE documentation, Palo Alto: Palo Alto Research Center. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html, Setembro.

Dalrymple, M. (2005), "Lexical-functional grammar", In: Encyclopedia of Language & Linguistics, Elsevier, 2nd edition.

Davies, W. D., Dubinsky, S. (2004), "Extended standard theory: Chomsky's 'conditions on transformations", In: The grammar of raising and control – a course in syntactic argumentation, Blackwell publishing.

Falk, Y. (2001), Lexical-functional grammar: an introduction to parallel constraint-based syntax, Stanford: CSLI Publications.

Kaplan, R. M., Bresnan, J. (1982), The mental representation of grammatical relations, Edited by Joan Bresnan, Cambridge.

Kroeger, P. R. (2004), Analyzing syntax: a lexical-functional approach, Cambridge: University Press.

Santos, A. F. dos. (2014), Uma gramática LFG-XLE para a análise sintática profunda do português. 178 f. Tese (Doutorado) – Centro de Humanidades, Departamento de Letras Vernáculas, Programa de Pós-Graduação em Linguística, Universidade Federal do Ceará, Fortaleza.

Schwarze, C. und Alencar, L. F. de. (2016), Lexikalisch-funktionale Grammatik – eine Einführung am Beispiel des Französischen mit computer-linguistischer Implementierung. Stauffenburg Verlag.

Sulger, S. et al. (2013), ParGramBank: the ParGram parallel treebank. In: Association for Computational Linguistics, 51., 2013. Proceedings… Sofia: Association for Computational Linguistics. p. 550-560.