Gramáticas Locais para Reconhecimento de Construções com Verbo Suporte em Português

Luís Enrique Santos Prado Vereau, Juliana Pinheiro Campos Pirovani

[1]Departamento de Computação - Universidade Federal do Espírito Santo; Alegre, ES - Brasil

luis.vereau@edu.ufes.br, juliana.campos@ufes.br

Abstract. *Natural Language Processing is an interdisciplinary subarea of Computer Science and Linguistics that aims to study the generation, representation and understanding of natural language by computers. This article describes the automatic generation of Local Grammars (LGs) for Supporting Verb Constructions (SVC) from six Lexicon-Grammar tables that describe a total of 468 CVS. For this, six parameterized graphs were created using the tool Unitex to extract information from these tables. The LGs generation was done through shell scripts and Unitex. The generated LGs were applied at the corpus aTribuna, where 211 structures in the form of the searched SVC were found.*

Resumo. *O Processamento de Linguagem Natural é uma subárea interdisciplinar da Ciência da Computação e Linguística que tem como objetivo o estudo da geração, representação e compreensão da linguagem natural por computadores. Este artigo descreve a geração semiautomática de Gramáticas Locais (GLs) para Construções com Verbo Suporte (CVS) a partir de seis tábuas do Léxico-Gramática que descrevem um total de 468 CVS. Para isto, foram criados seis grafos parametrizados utilizando a ferramenta Unitex para a extração de informações destas tábuas. A geração das GLs foi feita por meio de shell scripts e do Unitex. As GLs geradas foram aplicadas no corpus aTribuna, onde foram encontradas 211 estruturas no formato das CVS buscadas.*

1. Introdução

O Processamento de Linguagem Natural (PLN) é uma subárea interdisciplinar da Ciência da Computação e Linguística que se dedica a geração, representação e compreensão de linguagem natural de forma automática. Os próprios linguistas podem se beneficiar do PLN por meio das ferramentas construídas pelos profissionais da Computação, da mesma forma em que, simetricamente, a qualidade do PLN pode depender da descrição da língua pelos linguistas [Picoli et al. 2015]. [Chowdhury 2003] descreve PLN como sendo "uma área de pesquisa e aplicação que explora como computadores podem ser usados para entender e manipular texto e fala em linguagem natural para fazer coisas úteis".

Devido a necessidade atual de compartilhamento e compreensão de informação na era da internet, essa área tem ganhado destaque cada vez maior no meio acadêmico e na indústria. A grande quantidade de informação disponível atualmente em textos de escrita livre (não estruturados) precisa ser tratada para uso em aplicações que buscam informações específicas a partir deles. Exemplos de aplicações relevantes que necessitam das técnicas de PLN são sistemas de perguntas e respostas, tradução automática e reconhecimento de entidades nomeadas.

Entre os diversos problemas encontrados pelas máquinas ao processar a linguagem natural, há a interpretação de expressões com significado não composicional, que não podem ter seu significado extraído analisando o sentido literal de suas palavras individualmente. Expressões Cristalizadas (EC) como ”João comprou no mercado negro”e Construções com Verbo Suporte (CVS) como “João tem sangue frio para blefar” são exemplos dessas expressões. Este trabalho tem CVS como objeto de estudo, [Picoli 2020] expõe as diferenças entre EC e CVS.

CVS podem ser definidas como expressões compostas por um verbo que atua como Verbo Suporte (Vsup) e uma unidade predicativa não-verbal que pode ser um nome predicativo (Npred) como em “ter lábia”(Vsup+Npred), um adjetivo (adj) como na construção “estar liso”(Vsup+adj), ou uma expressão que se comporta como adjetivo (Expadj), “estar azul de fome”(Vsup+Expadj), por exemplo [Picoli 2020].

[Flores 2020] analisou ocorrências do verbo *dar* em duas coleções de textos em português brasileiro, o *corpus* do projeto Fala Goiânia[1] e o *corpus* do Grupo Discurso & Gramática[2]. Notou-se que das 190 ocorrências do verbo no *corpus* do projeto Fala Goiânia, em 103 (54,21%) ele atuava como Vsup, e no *corpus* do Grupo Discurso & Gramática, das 70 ocorrências encontradas, em 47 (67,14%) o verbo assume papel de Vsup, indicando grande presença das CVS na língua portuguesa do Brasil.

Desta forma, o reconhecimento automático ou semiautomático de CVS é importante para o PLN pela riqueza de significado de seus elementos, a dificuldade de entendimento da não-composicionalidade pelo computador, bem como pelo frequente aparecimento de expressões desse tipo no português. A geração automática de resumos, tradução de máquina e *chatbots* [Tan et al. 2021] são possíveis tarefas nas quais o reconhecimento dessas expressões poderá ser aplicado.

Um dos métodos para descrever CVS é o Léxico-Gramática [Gross 1975] que consiste em criar tabelas, também chamadas de tábuas, que detalham um conjunto de expressões e possíveis variações para certos elementos, como verbo utilizado, ausência ou existência de negação, comparação, intensificação, dentre outras.

Uma forma de reconhecer expressões com características sintáticas e semânticas em comum, como as CVS, são as Gramáticas Locais (GLs). As Gramáticas Locais [Gross 1997] são “gramáticas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões de uma língua natural”.

Este trabalho tem como objetivo a extração de informações presentes nas seis tábuas do Léxico-Gramática que descrevem CVS em [Picoli 2020] para geração semiautomática de GLs que sejam capazes de reconhecer essas expressões e anotar seus significados. O trabalho busca, por fim, fornecer recursos que possam ser utilizados em aplicações de PLN e auxiliem tanto profissionais da computação quanto da linguística.

2. Metodologia

Unitex[3] é um conjunto de software livres para PLN que permite, além do pré-processamento de textos, a construção de GLs e a construção automática de GLs a partir

[1]https://gef.letras.ufg.br/p/11947-projetos-tematicos

[2]https://discursoegramatica.wordpress.com/

[3]https://unitexgramlab.org/pt/

de uma tabela (neste caso, tábua do Léxico-Gramática) e de um grafo parametrizado.

As GLs no Unitex são representadas por grafos como o apresentado na Figura 1 que reconhece a estrutura [Nome Próprio (reconhecido pelo código lexical <N+Pr >) + não (opcional) + verbo *ser*, *ter*, *permanecer* ou *continuar* (reconhecido pelo código lexical <Verbo.V >, como <ser.V >) + coração grande].

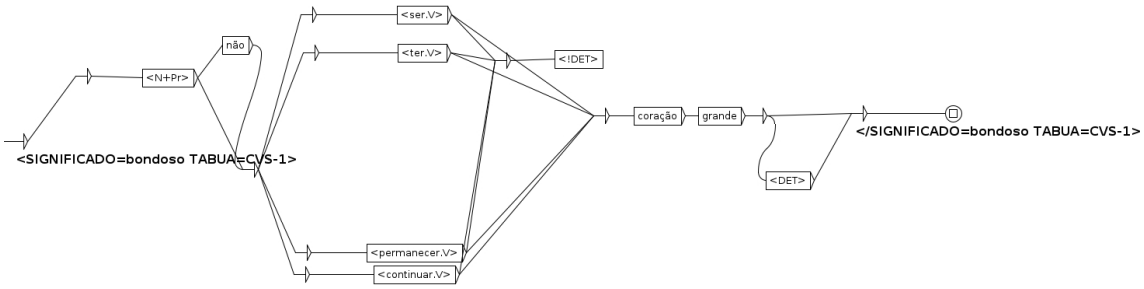

Figura 1. SubGL CVS-1

Um exemplo de tábua do Léxico-Gramática é apresentado na Figura 2. Na linha 3 desta tábua está descrita a CVS representada pela subGL da Figura 1.

1	N0=Hum	N0=Nhum	N0=Nhor	ser	estar	estar (part)	ficar	ficar (part)	ter	permanecer	continuar	continuar (part)	tornar (se)	tornar	virar	virar (part)	andar	andar (part)	ir	aguentar	receber	perder	possuir	manter	haver	colocar	pôr	ERE=E	ERE=De4	ERE=Ne4	ERE=BRON	ADJ+ADV	N	ADJ	PREP+DET	PRON	N	N	A	ADJ	Composição	Intensificação	Significado	Exemplo
2	+	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	<E>	estrela	<E>	na	<E>	testa	<E>	<E>	<E>	-	-	confiável	João (não) tem estrela na testa
3	+	-	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	<E>	coração	grande	<E>	<E>	<E>	<E>	<E>	<E>	-	-	bondoso	João é coração grande
4	+	-	-	+	+	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	duas	<E>	de	<E>	caras	<E>	<E>	<E>	-	+	falso	João é duas caras
5	+	-	-	+	-	+	-	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	espírito	<E>	de	<E>	porco	<E>	<E>	<E>	+	+	inconveniente	João é espírito de porco
6	+	-	-	+	-	+	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	mão	furada	<E>	<E>	<E>	<E>	<E>	<E>	+	+	gastador	João é mão furada
7	+	-	-	+	+	-	+	-	+	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	mão	<E>	de	<E>	ferro	<E>	<E>	<E>	-	-	opressor	João é mão de ferro
8	+	-	-	+	-	+	-	+	-	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	pavio	curto	<E>	<E>	<E>	<E>	<E>	<E>	+	+	estressado	João é pavio curto
9	+	-	-	+	+	-	+	-	+	-	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<E>	sangue	quente	<E>	<E>	<E>	<E>	<E>	<E>	+	-	irritadiço	João é sangue quente
10	+	-	-	-	-	+	-	+	+	-	-	+	-	-	+	+	+	-	-	-	-	-	+	-	-	+	-	-	+	-	<E>	peso	<E>	na	<E>	consciência	<E>	<E>	<E>	+	+	despreocupado	João tem peso na consciência
11	+	-	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	<E>	boca	suja	<E>	<E>	<E>	<E>	<E>	<E>	-	-	fala palavrão	João tem a boca suja
12	+	-	-	-	-	+	-	+	+	-	-	+	-	-	+	+	+	-	-	-	-	-	+	-	-	-	-	-	-	-	<E>	cabeça	cheia	de	<E>	minhocas	<E>	<E>	<E>	-	-	preconceituoso	João tem a cabeça cheia de minhocas
13	+	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	a	-	-	<E>	cabeça	<E>	no	<E>	lugar	<E>	<E>	<E>	-	-	íntegro	João tem a cabeça no lugar
14	+	-	-	-	-	+	-	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	a	-	-	<E>	faca e o queijo	<E>	na	<E>	mão	<E>	<E>	<E>	-	-	preparado	João tem a faca e o queijo na mão
15	+	-	-	-	-	+	-	-	+	-	-	+	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-	a	-	-	<E>	ficha	limpa	<E>	<E>	<E>	<E>	<E>	<E>	-	-	honesto	João tem a ficha limpa

Figura 2. Tábua CVS-1

Por meio da ferramenta Unitex, foi construído manualmente um grafo parametrizado para cada uma das seis Tábulas do Léxico-Gramática que descrevem CVS em [Picoli 2020] (denominadas CVS-1, CVS-1cop, CVS-2, CVS-2cop, CVS-P1_Prep e CVS-P2_Prep).

Grafos parametrizados, cujo exemplo é apresentado na Figura 3, são GLs que usam variáveis para se referir às colunas de uma tábua do Léxico-Gramática, onde @[Letra do alfabeto em maiúsculo] referem-se às colunas da tábua em ordem crescente. Isto é, @A refere-se à primeira coluna, @B à segunda e assim por diante. Os grafos parametrizados geram, de forma automática, uma subGL para cada linha da tábua e uma GL principal que engloba todas as subGLs.

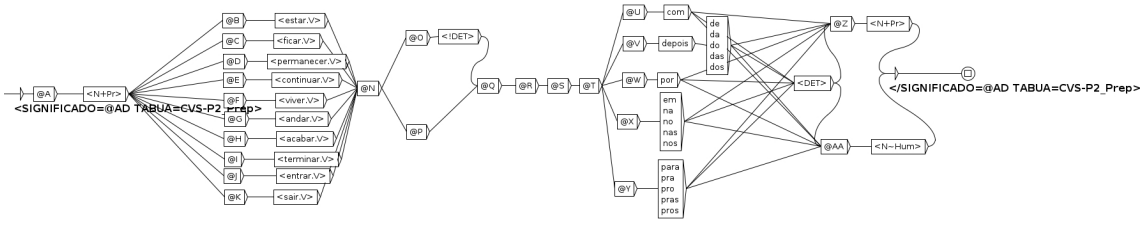

Figura 3. Grafo parametrizado para CVS-P2_Prep

As GLs geradas pelos grafos parametrizados foram aplicadas no *corpus* composto por um conjunto de textos publicados pelo jornal do Espírito Santo aTribuna[4]. Esse *corpus* contém textos jornalísticos de gêneros variados, tendo sido utilizado em [Santiago 2022]. Também foram criados *shell scripts* que automatizaram o processo de criação e aplicação das GLS ao chamar programas do Unitex responsáveis pela *tokenização* dos textos (Tokenize), aplicação de um Grafo Parametrizado em uma Tábua do Léxico-Gramática (Table2Grf), compilação das GLs (Grf2Fst2), dentre outros.

3. Resultados

Foram construídos 6 grafos parametrizados e geradas semiautomaticamente 473 GLs para reconhecimento de CVS. Para o *corpus* escolhido, foram encontradas 211 correspondências com estruturas no formato das CVS buscadas.

Algumas das expressão corretamente identificadas foram: ”Vaz tem carta branca”, ”três é show de bola”, ”eu sou o máximo”, ”vida é um mar de rosas”, ”senão fica o dito pelo não dito”, ”público não esteve lá essas coisas”, ”Eduardo está entre a cruz e a espada”, ”Motta é do ramo”e ”Sandra tem os dias contados”, sendo ”ter os dias contados”a CVS mais frequente, aparecendo 21 vezes (9,95%).

Observou-se também a presença de falso-positivos dentre as construções identificadas, i.e, expressões que foram identificadas como CVS, mas não são CVS. Por exemplo, anotou-se a expressão “o corpo foi achado”, falso positivo para a CVS *ser um achado*, como em “o livro é um achado”. Notou-se que a ocorrência deste falso-positivo em particular foi devido à descrição na tábua CVS-1cop que permite essas construções sem o artigo indefinido. Sabendo que este artigo é necessário para esta CVS, a tábua poderia ser alterada para corrigir essa situação melhorando a precisão da GL gerada.

Outra CVS que apresentou falso-positivos foi *ser o de menos*, que indica pouca relevância como em “essa informação será o de menos”. Um dos falso-positivos identificados foi “A academia também observou que a taxa de acréscimo de novas reservas em todo o mundo era de menos”, que indica que taxa de acréscimo de novas reservas em todo o mundo foi menor que certo valor, apresentando, desta forma, significado bem diferente ao da CVS.

Faz-se necessário, então, analisar os grafos parametrizados novamente a fim de mitigar a aparição de falso-positivos.

4. Conclusão

Neste trabalho foram geradas 473 GLs a partir de tábuas do Léxico-Gramatica e grafos parametrizados construídos no Unitex. As GLs identificaram no *corpus* aTribuna 211 estruturas correspondentes a CVS.

Como metas futuras, os grafos parametrizados construídos até então sofrerão novas análises e serão feitas melhorias. Igualmente serão feitas novas análises nas tábuas do Léxico-Gramática a fim de remover possíveis falhas. Também será realizada a análise de resultados, avaliando-se a precisão das correspondências encontradas.

Pode-se avaliar também em quais gêneros textuais houve mais ocorrências de CVS e analisar se, nesses textos, há alguma CVS não descrita nas tábuas.

[4]https://tribunaonline.com.br/

Referências

Chowdhury, G. (2003). Natural language processing. *Annual review of information science and technology*, pages 51–89.

Flores, E. (2020). A construção-suporte no português brasileiro.

Gross, M. (1975). Méthodes en syntaxe: régime des constructions complétives.

Gross, M. (1997). The construction of local grammars. *Finite-state language processing*, pages 329–354.

Picoli, L. (2020). Contínuo e limite entre expressão cristalizada e construção com verbo-suporte à luz do léxico-gramática.

Picoli, L. et al. (2015). Uso de uma ferramenta de processamento de linguagem natural como auxílio à coleta de exemplos para o estudo de propriedades sintático-semânticas de verbos. *Linguamática*, pages 35–44.

Santiago, D. (2022). Gramáticas locais para reconhecimento de expressões cristalizadas em português. *Relatório Parcial de Pesquisa, Programa Institucional de Iniciação Científica 2021/2022*.

Tan, K. et al. (2021). Review on light verb constructions in computational linguistics. pages 25–26.