GRAMÁTICAS LOCAIS PARA EXPRESSÕES CRISTALIZADAS E CONSTRUÇÕES COM VERBO-SUPORTE

Kailany Alves Silva, Thiago Tonelli da Silva, Juliana Pinheiro Campos Pirovani.

Universidade Federal do Espírito Santo/Departamento de Computação, Alto Universitário, S/N, Guararema - 295000-000 - Alegre-ES, Brasil, kailany.silva@edu.ufes.br, thiago.silva.66@edu.ufes.br, juliana.campos@ufes.br.

Resumo

O artigo visa aprimorar Gramáticas Locais (GLs) existentes para o reconhecimento de Construções com Verbo-Suporte (CVS) e Expressões Cristalizadas (EC), devido à sua relevância em várias aplicações no Processamento de Linguagem Natural (PLN), como sistemas de perguntas e respostas, tradução automática e reconhecimento de entidades nomeadas. A melhoria no reconhecimento dessas construções resulta em sistemas de PLN mais precisos e eficientes. A partir da análise dos trabalhos utilizados como base, foram implementadas alterações nas GLs existentes, que foram aplicadas no *corpus* aTribuna por meio de *shell scripts*. Os resultados para as CVS foram promissores, com um aumento de 88 construções reconhecidas. Entretanto, para as ECs, ainda não foram obtidas melhorias significativas.

Palavras-chave: Processamento de Linguagem Natural. Reconhecimento de Construções com Verbo-Suporte. Reconhecimento de Expressões Cristalizadas. Gramáticas Locais. Tábuas do Léxico-Gramática.

Área do Conhecimento: Ciências Exatas e da Terra - Ciência da Computação

Introdução

O Processamento de Linguagem Natural (PLN) é uma subárea da Ciência da Computação, juntamente com a Linguística, que visa analisar e interpretar a linguagem natural em textos de escrita livre.

Segundo a International Data Corporation (2019), estima-se que até 2025, os dados não estruturados irão compor cerca de 80% de todos os dados (apud Forbes, 2022). Com o crescimento de informações não estruturadas, os textos de escrita livre necessitam de um processamento adequado para serem utilizados em aplicações específicas, como sistemas de perguntas e respostas, tradução automática e reconhecimento de entidades nomeadas (Vereau; Pirovani, 2023).

Entretanto, o PLN encontra dificuldades no tratamento de textos não estruturados, devido à complexidade da linguagem humana. Um dos problemas cruciais é a interpretação de expressões com sentido não composicional, como as Expressões Cristalizadas (EC) e algumas Construções com Verbo-Suporte (CVS). Estas expressões apresentam significados não compostos pelo sentido literal de suas palavras, como no caso da CVS "João está de molho" que significa estar de repouso, doente.

As CVS são expressões formadas por um verbo-suporte (Vsup) e uma unidade predicativa não-verbal. Esta unidade pode ser um nome predicativo, como em *ter lábia* (Vsup+NPred), um adjetivo, como *estar liso* (Vsup+adj), ou uma expressão que se comporta com adjetivo, *estar azul de fome* (Vsup+Expadj). Por outro lado, as EC são compostas por múltiplas palavras que podem ser interpretadas como uma entrada lexical, como em “João tem cartaz”.

Uma abordagem para representar essas expressões são as tábuas do Léxico-Gramática (LG) (Gross, 1975) que apresentam em uma tabela um conjunto de expressões e suas possíveis variações. Ademais, essas expressões podem ser identificadas utilizando Gramáticas Locais (GLs) (Gross, 1997), que são regras descritas manualmente para reconhecer sentenças com características semelhantes em textos.

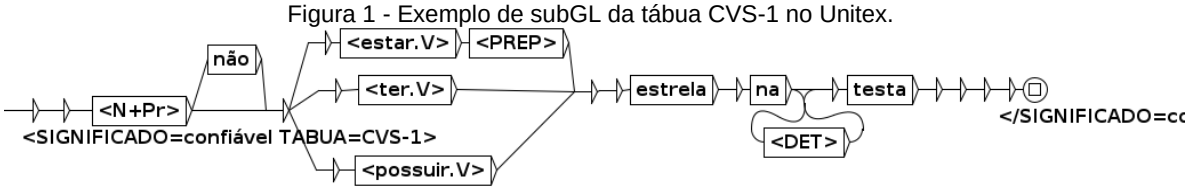
Este estudo tem como objetivo aprimorar as GLs existentes para o reconhecimento de CVS (Vereau; Pirovani, 2023) e EC (Santiago, 2022), visando melhorar o desempenho dessas GLs. Sendo assim, espera-se contribuir significativamente para o avanço do Processamento de Linguagem Natural (PLN).

Metodologia

Inicialmente, foram analisadas as Gramáticas Locais construídas nos trabalhos de Vereau e Pirovani (2023) e Santiago (2022) para o reconhecimento de CVS e EC, respectivamente. Em seus trabalhos, os autores utilizaram a ferramenta Unitex[1], um conjunto de *software* livres para PLN, para construir essas GLs a partir de tábuas do Léxico-Gramática e grafos parametrizados. As tábuas utilizadas foram as de Picoli (2020).

GLs são autômatos de estados finitos que representam um conjunto de expressões de uma linguagem natural (Gross, 1997). Estas são representadas por grafos no Unitex, como o ilustrado na Figura 1, que reconhece expressões com a estrutura [Nome Próprio (código <N+Pr> no Unitex) + não (opcional) + verbo *estar*, *ter* ou *possuir* (código <Verbo.V> no Unitex, como <ter.V>) + estrela na testa].

Figura 1 - Exemplo de subGL da tábua CVS-1 no Unitex.



Fonte: o autor.

Tábuas do Léxico-Gramática são tabelas criadas para detalhar um conjunto de expressões e suas possíveis variações para determinados elementos, como o verbo utilizado, presença ou ausência de negação, comparação e intensificação, dentre outros (Vereau; Pirovani, 2023). A Figura 2 apresenta um exemplo de Tábua do Léxico-Gramática, no qual o símbolo "+" indica a aceitabilidade de uma determinada propriedade e o "-" a sua inaceitabilidade.

Figura 2 - Tábua do Léxico-Gramática (CVS-2).

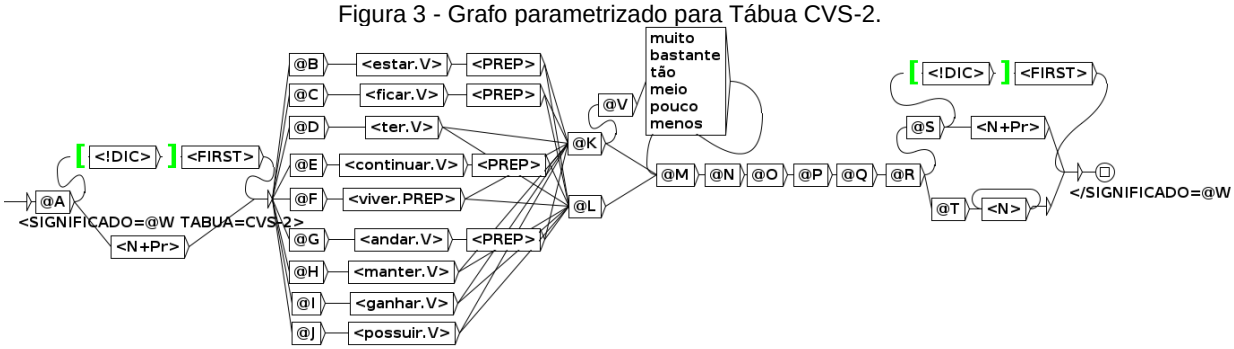
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	N0=:Nhum	estar (prep)	ficar (prep)	ter	continuar (prep)	viver (prep)	andar (prep)	manter	ganhar	possuir	DET=:E	DET=:Indef	N	ADJ	PREP/CONJ	N	V	Prep	N1=:Nhum	N1=:F	Comparação	Intensificação	Significado	Exemplo
2	+	+	-	+	+	-	-	-	-	-	+	-	estômago	<E>	para	<E>	conversar	com	+	-	+	+	disposto	João tem estômago para conversar com pessoas falsas
3	+	-	-	+	-	-	-	-	-	-	-	um	dedo	<E>	de	conversa	<E>	com	+	-	-	-	instante	João tem um dedo de conversa/prosa com Ana
4	+	+	-	+	+	-	-	-	-	+	-	uma	queda	<E>	<E>	<E>	<E>	por	+	-	-	-	sentimento	João tem uma queda por alguém
5	+	-	-	+	-	-	-	+	-	+	+	-	sangue	frio	<E>	<E>	<E>	para	-	+	+	+	calculista	João tem sangue frio para blefar
6	+	+	+	+	+	+	+	-	+	-	+	-	ares	<E>	<E>	<E>	<E>	de	+	-	+	-	aparenta	João tem ares de poeta

Fonte: Picoli, 2020.

Os grafos parametrizados são GLs que empregam variáveis para referenciar as colunas de uma tabela do Léxico-Gramática, onde as variáveis como @[Letra do alfabeto em maiúsculo] correspondem às colunas da tabela em ordem crescente, por exemplo, @A refere-se à primeira coluna, @B à segunda e assim sucessivamente (Paumier, 2016). A Figura 3 apresenta o grafo parametrizado desenvolvido para a Tábua do Léxico-Gramática exposta na Figura 2.

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

Figura 3 - Grafo parametrizado para Tábua CVS-2.



Fonte: o autor.

As GLs de Vereau e Pirovani (2023) e Santiago (2022) foram aplicadas ao *corpus* aTribuna[2], também utilizado por esses autores. Esse *corpus* é composto por um conjunto de textos jornalísticos de diversos gêneros. Os resultados foram analisados para identificar possíveis melhorias que poderiam ser realizadas nos seus grafos parametrizados, com o intuito de gerar GLs com melhor desempenho.

Após observar as possibilidades de melhorias, os grafos parametrizados foram adaptados e tornou-se necessário gerar as GLs atualizadas. Utilizando *shell scripts* para automatizar o processo de criação e aplicação das GLs em *corpus*, as GLs foram aplicadas no *corpus* aTribuna. Como resultado, foi gerado um arquivo .txt contendo todas as ocorrências encontradas no texto.

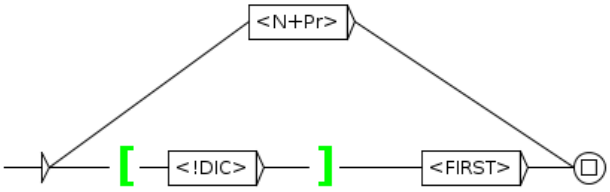
Para avaliar o desempenho das GLs adaptadas, foi utilizada a precisão, que indica a quantidade total de acertos, em porcentagem, no total de expressões reconhecidas (Pirovani, 2019). Quanto maior a precisão, menor o número de falso-positivos. A fórmula abaixo representa o cálculo desta medida.

$$Precisão = \frac{Total\ de\ reconhecimento\ correto\ de\ expressões}{Total\ de\ reconhecimento\ de\ expressões}$$

Resultados

Observou-se que alguns substantivos próprios – como nomes, sobrenomes, organizações, etc. – não estavam sendo reconhecidos no dicionário do UNITEX, isto é, não eram identificados pela entrada lexical <N+Pr>. Como solução, para os grafos parametrizados que contém a *tag* <N+Pr>, foi adicionado um subgrafo que reconhece palavras que não são reconhecidas pelo dicionário e começam com letra maiúscula, conforme apresentado na Figura 4.

Figura 4 - Subgrafo que reconhece substantivos próprios.

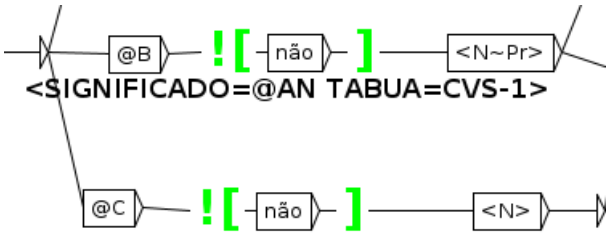


Fonte: o autor.

Outra adversidade observada no dicionário do UNITEX refere-se à classificação de palavras, o que ocasiona em falso-positivos. Por exemplo, “não era de coração” é um falso-positivo reconhecido pela GL, pois o dicionário classifica “não” como substantivo masculino, não apenas como advérbio. Portanto, quando a GL identifica um sujeito (<N>) ou sujeito não humano (<N~Pr>) a palavra “não” é equivocadamente reconhecida. Para corrigir esse problema, foi necessário adicionar um nó que não reconhece expressões iniciadas por “não”, como demonstrado na Figura 5.

Figura 5 – Nó adicionado para não reconhecer expressões iniciadas por *“não”*.

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



Fonte: o autor.

Além disso, observou-se que, nas tábuas do Léxico-Gramática (Picoli, 2020), havia propriedades transformacionais. As transformações de intensificação admitem advérbios intensificadores como *muito, meio e bastante.* As comparativas consistem em inserir estruturas como *mais/menos __ do que* e *tanto/tão ___ quanto* (Picoli, 2020). Tem-se como exemplos “João é *muito* espírito de porco”; e “João é *mais* espírito de porco *do que* Maria”. As transformações de intensificação foram incluídas nos grafos parametrizados, como exemplificado na Figura 6.

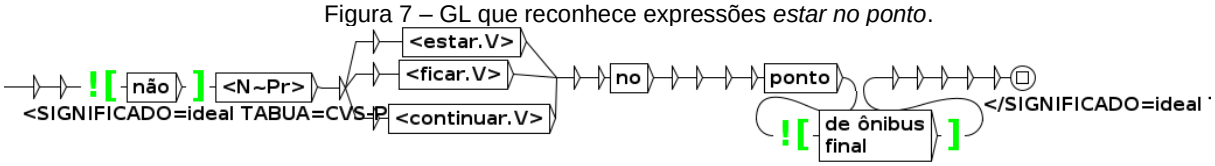
Figura 6 - Advérbios de intensidade adicionados aos grafos parametrizados.

Fonte: o autor.

É importante destacar que o UNITEX não possui uma *tag* específica para advérbios de intensidade, apenas a tag <ADV>, que reconhece todos os tipos de advérbios. Por essa razão, optou-se por utilizar os advérbios de intensidade mais frequentes em textos. Normalmente, estes advérbios aparecem entre o verbo e o complemento. No entanto, observou-se casos em que ocorrem antes do adjetivo, como em “João está de cabeça muito quente”. Assim, foram adicionados advérbios de intensidade, de forma opcional, antes da ocorrência de adjetivo nos grafos.

Nos estudos de Vereau e Pirovani (2023), a expressão *estar no ponto* foi destacada, visto que, das 20 ocorrências observadas, 15 (75%) foram classificadas como falso-positivos. Analisando o contexto no *corpus*, foi possível perceber que a palavra “ponto” frequentemente aparecia acompanhada de “de ônibus” ou “final”, indicando o sentido de estar no ponto [final] de ônibus. Para solucionar o problema, foi inserido um nó especificando que a palavra “ponto” não pode ser seguida por “de ônibus” ou “final”, conforme ilustrado na Figura 7.

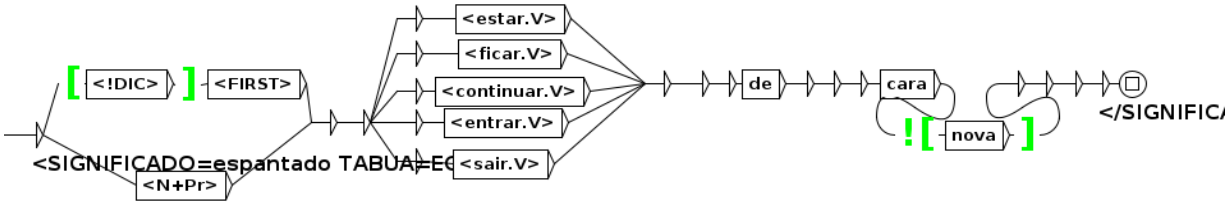
Figura 7 – GL que reconhece expressões *estar no ponto*.



Fonte: o autor.

Também foi observado que a EC *estar de cara* estava gerando falso-positivos, pois reconhecia expressões associadas ao sentido de mudança, como “Philco está de cara”. Após analisar o *corpus*, confirmou-se que a expressão era seguida pela palavra “nova”. Portanto, foi inserido um nó que impede o reconhecimento de expressões seguidas pela palavra “nova”, como exposto na Figura 8.

Figura 8 – GL que reconhece expressões *estar de cara.*



Fonte: o autor.

Após realizar as melhorias nas GLs de CVS (Vereau; Pirovani, 2023) e EC (Santiago, 2022), foram encontradas 409 expressões, onde 283 eram CVS e 126 EC. Para as CVS foi obtido 79% de precisão e para as EC 38%. As Tabelas 1 e 2 apresentam uma comparação dos resultados obtidos com as GLs adaptadas neste estudo e as GLs dos trabalhos anteriores.

Tabela 1 - Comparação com os resultados de Vereau e Pirovani (2023).

GLs	CVS reconhecidas	Verdadeiro-positivos	Falso-positivos	Precisão (%)
Vereau e Pirovani (2023)	195	111	84	57%
GLs adaptadas	283	224	59	79%

Fonte: o autor.

Tabela 2 - Comparação com os resultados de Santiago (2022)

GLs	CVS reconhecidas	Verdadeiro-positivos	Falso-positivos	Precisão (%)
Santiago (2022)	110	37	73	34%
GLs adaptadas	126	48	78	38%

Fonte: o autor.

Discussão

Os resultados para o reconhecimento de CVS indicaram uma melhora significativa, evidenciada pelo aumento da precisão (ganho de 22%, como pode ser visto na Tabela 1). A expressão que obteve destaque foi *ter os dias contados,* apresentando 45 ocorrências (15,9%), sendo todas classificadas como verdadeiro-positivos. Em contrapartida, a expressão *ser duro* apresentou a maior quantidade de falso-positivos, com 13 ocorrências – 5 (≈38,5%) verdadeiro-positivos e 8 (≈61,5%) falso-positivos. Não foi possível realizar novas melhorias para esta expressão, devido a ausência de um padrão consistente entre os falso-positivos reconhecidos.

No reconhecimento de EC, até o momento, não foram obtidas grandes melhorias. As alterações realizadas aumentaram o número de verdadeiro-positivos, mas também aumentaram o número de falso-positivos. Acredita-se que a dificuldade no reconhecimento dessas expressões, também relatada por Santiago (2022), esteja relacionada à estrutura delas.

A expressão mais problemática foi *ser de casa*, com significado de familiar, que aparece 29 vezes, todas como falso-positivos, como em “Holanda fora de casa”. Não foi possível resolver este problema pois ele está associado ao dicionário do Unitex, que atribui todas as classificações gramaticais possíveis para as palavras, interpretando “fora” como verbo também e permitindo o reconhecimento dessa expressão como EC.

Algumas tábuas do Léxico Gramática, como a EC-P1, apresentaram problemas, devido a um erro na configuração relacionado às preposições e determinantes. Em vez de reconhecer “no”, estava reconhecendo apenas “em”. Isso gerou a ocorrência de falso-positivos, como “Vitória está em céu”, e interferiu nos resultados obtidos para as EC. Além disso, houve falta de informações necessárias para implementar as transformações comparativas nos grafos parametrizados. Assim, torna-se necessário fazer alterações para que os resultados sejam cada vez melhores.

Conclusão

Neste trabalho, foram realizadas melhorias em GLs existentes para o reconhecimento de Construções com Verbo-Suporte e Expressões Cristalizadas, elementos cruciais em diversas aplicações do Processamento de Linguagem Natural. Com base na análise do *corpus* aTribuna e nas GLs dos trabalhos anteriores, diversas melhorias foram implementadas. Foram utilizados *shell scripts* para geração e aplicação automatizada das novas GLs no *corpus*.

Os resultados obtidos para as CVS foram significativos (ganho de 22%), evidenciando um aumento no reconhecimento das expressões e diminuição de falso-positivos em relação ao número total de reconhecimento. Entretanto, para as EC, foi obtido um ganho de 4% com relação ao trabalho de Santiago (2022).

Portanto, para futuros trabalhos, é importante dar ênfase às Expressões Cristalizadas, buscando aprimorar seus resultados. Além disso, é essencial revisar e implementar melhorias nas tábuas do Léxico-Gramática, a fim de mitigar a ocorrência de falso-positivos e acrescentar informações para reconhecimento das transformações comparativas. Pretende-se também aplicar as novas GLs em outros *corpus* de textos.

Referências

GROSS, M. **Méthodes en Syntaxe: Régime des Constructions Complétives**. 1975. p. 414.

GROSS, M. **The construction of local grammars**. Finite-state language processing. 1997. p. 329-354.

O'REILLY, M. **The Unseen Data Conundrum**. 2022. Disponível em: https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/. Acesso em: 24 jun. 2024.

PAUMIER, S. **Unitex 3.3 User Manual**. 2021. Disponível em: https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf. Acesso em: 24 jun. 2024.

PICOLI, L. **Contínuo e limite entre expressão cristalizada e construção com verbo-suporte à luz do Léxico-Gramática**. 2020. 179 f. Tese (Doutorado em Linguística) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, 2020.

PIROVANI, J. P. C.. **CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português**. 2019. Tese (Doutorado em Ciência da Computação) – Centro Tecnológico, Universidade Federal do Espírito Santo, 2019.

SANTIAGO, D. H. **Gramáticas Locais para Reconhecimento de Expressões Cristalizadas em Português.** 2022. In: JORNADA DE INICIAÇÃO CIENTÍFICA DA UFES, 13., 2022, Vitória. Anais… Vitória: UFES, 2022. Disponível em: https://anaisjornadaic.sappg.ufes.br/desc.php?&id=19144. Acesso em: 15 ago. 2024.

VEREAU, L. E. S. P.; PIROVANI, J. P. C. **Gramáticas Locais para Reconhecimento de Construções com Verbo Suporte em Português**. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14., 2023, Belo Horizonte. **Anais**... Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 347-351.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) pelo financiamento do projeto que deu origem a este trabalho.

Expressamos nossa sincera gratidão à nossa orientadora, Juliana, pela atenção, paciência e orientação, que têm sido essenciais para a realização deste trabalho e para minha formação acadêmica.