Revista brasileira de iniciação científica
ISSN: 2359-232X

Gramáticas Locais Para o Reconhecimento de Expressões Cristalizadas e Construções com Verbo-Suporte em Português

Local Grammars for the Recognition of Frozen Expressions and Support-Verb Constructions in Portuguese

Gramáticas Locales Para el Reconocimiento de Expresiones Cristalizadas y Construcciones Soportadas por Verbos en Portugués

Kailany Alves Silva[1]
Juliana Pinheiro Campos Pirovani[2]

Resumo: Este trabalho teve como objetivo melhorar Gramáticas Locais (GLs) existentes voltadas ao reconhecimento de Expressões Cristalizadas (ECs) e Construções com Verbo-Suporte (CVSs), dada a importância dessas estruturas para diversas aplicações no campo do Processamento de Linguagem Natural (PLN). A partir da análise de trabalhos anteriores, foram implementadas melhorias nas GLs existentes, construídas por meio de Tábuas do Léxico-Gramática (TLGs) e grafos parametrizados (GPs) usando a ferramenta Unitex. As GLs obtidas foram aplicadas no *corpora* aTribuna e PARSEME a partir de *shell scripts.* Para avaliação de desempenho, foi utilizado o cálculo da precisão. Os resultados foram significativos para ambas as construções. Para o *corpus* aTribuna, obteve-se um aumento de 21% na precisão para as ECs e 22% para as CVSs. No PARSEME, a precisão foi de 74% para as ECs e 78% para as CVSs.
Palavras-chave: Processamento de Linguagem Natural. Gramáticas Locais. Expressões Cristalizadas. Construções com Verbo-Suporte.

Abstract: This work aims to improve existing Local Grammars (LGs) for the recognition of Frozen Expressions (FEs) and Support-Verb Constructions (SVCs), due to their relevance in various applications in Natural Language Processing (NLP). Based on an analysis of previous work, improvements were made to existing LGs, built from Lexicon-Grammar Tables (LGTs) and parameterized graphs (PGs) using the Unitex tool. The GLs obtained were applied to the aTribuna and PARSEME corpus using shell scripts. The accuracy calculation was used to evaluate performance. The results were significant for both constructions. In the aTribuna corpus, a 21% increase in precision was obtained for the FEs and 22% for the SVC. For PARSEME, accuracy was 74% for FEs and 78% for SVC.
Keywords: Natural Language Processing. Local Grammars. Crystallized Expressions. Support-Verb Constructions.

[1] Graduanda em Ciência da Computação. Universidade Federal do Espírito Santo – UFES, Campus Alegre. ORCID: https://orcid.org/0009-0003-4778-464X. E-mail: kailany.silva@edu.ufes.br.
[2] Doutora em Ciência da Computação. Universidade Federal do Espírito Santo – UFES, Campus Alegre. ORCID: https://orcid.org/0000-0002-3727-4158. E-mail: juliana.campos@ufes.br.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Resumen: Este trabajo pretende mejorar las Gramáticas Locales (GLs) existentes para el reconocimiento de Expresiones Cristalizadas (ECs) y Construcciones con Soporte Verbal (CVSs), debido a su relevancia en diversas aplicaciones en Procesamiento del Lenguaje Natural (PLN). A partir del análisis de trabajos previos, se introdujeron mejoras en las GLs existentes, construidas a partir de Tablas Léxico-Gramáticas (TLGs) y grafos parametrizados (GPs) mediante la herramienta Unitex. Las GLs obtenidas se aplicaron a los corpus aTribuna y PARSEME mediante shell scripts. Para evaluar el rendimiento se utilizaron cálculos de precisión. Los resultados fueron significativos para ambas construcciones. En el corpus aTribuna, la precisión aumentó un 21% para los ECs y un 22% para los CVSs. En PARSEME, la precisión fue del 74% para las ECs y del 78% para las CVSs.
Palabras-clave: Procesamiento del Lenguaje Natural. Gramáticas Locales. Expresiones Cristalizadas. Construcciones Soportadas por Verbos.

Submetido 26/07/2025 **Aceito 30/09/2025** **Publicado 02/12/2025**

Revista brasileira de iniciação científica
ISSN: 2359-232X

Considerações Iniciais

Atualmente, grande parte das informações disponíveis estão contidas em dados não estruturados, como e-mails, postagens em redes sociais e outros tipos de textos em escrita livre. De acordo com Taylor (2025), estima-se que até 2028 a criação de dados ultrapassará 394 zettabytes. Diante desse cenário, surge a necessidade de adotar estratégias eficazes para extrair informações relevantes desses dados.

Nesse contexto, o Processamento de Linguagem Natural (PLN) desempenha um papel fundamental, permitindo a análise e interpretação de dados não estruturados, transformando-os em conhecimento útil, informação. O PLN consiste no estudo e desenvolvimento de técnicas computacionais voltadas para o entendimento e manipulação da linguagem humana por máquinas. Ele pode ser utilizado com abordagem linguística, aprendizado de máquina ou com métodos híbridos (Cavalcanti et al., 2021). De modo geral, a área do PLN busca soluções para desafios computacionais, ou seja, tarefas, sistemas, aplicações ou programas que envolvem o processamento de línguas naturais, tanto na forma escrita quanto falada (fala) (Caseli; Nunes, 2024). No entanto, a complexidade da linguagem natural ainda representa uma dificuldade para o PLN.

Um problema crucial é a interpretação de expressões multipalavras (*Multiword Expressions* – MWEs), que se caracterizam por transmitirem significados complexos que, frequentemente, não podem ser compreendidos a partir do sentido isolado de cada palavra (Caseli; Nunes, 2024). Essas expressões consistem em sequências de palavras com, no mínimo, uma relação sintática entre si, e apresentam algum grau de irregularidade em termos de forma ou significado (Savary et al., 2018). Um exemplo é a expressão *A ideia é sem pé nem cabeça*, utilizada para indicar que algo não faz sentido, ainda que, conceitualmente, ideias não possuam corpo, pés ou cabeça (Caseli; Nunes, 2024).

A dificuldade dos computadores em processar esse tipo de construção decorre, sobretudo, de seu significado não composicional. Por exemplo, a expressão *Pedro tem berço*, quando interpretada de forma literal, pelo seu significado composicional, indica a posse de um berço por um indivíduo. No entanto, no uso figurado típico da linguagem natural, a construção transmite a ideia de um sujeito bem-afortunado e/ou de família rica. Essa distinção entre sentido

Revista brasileira de iniciação científica
ISSN: 2359-232X

literal e figurado evidencia o desafio semântico que as MWEs impõem aos sistemas computacionais.

As Expressões Cristalizadas (ECs) e algumas Construções com Verbo-Suporte (CVSs) são subcategorias relevantes dentro das MWEs. Essas construções são amplamente utilizadas na comunicação cotidiana e sua correta identificação é essencial para diversas aplicações em PLN, como tradução automática, sumarização e análise de sentimentos.

As ECs são aquelas cujo significado não pode ser deduzido diretamente a partir do sentido isolado de cada uma de suas palavras, como em *Ana é surda que nem uma porta* (Vale, 1999). Já as CVSs consistem em estruturas compostas por um verbo com função de suporte (Vsup), que se associa a uma unidade predicativa de natureza não verbal. Essa unidade pode assumir diferentes formas, tais como: um nome com valor predicativo [*ter inveja* (Vsup + NPred)], um adjetivo [*estar liso* (Vsup+adj)] ou ainda uma construção adjetival [*estar vermelho de raiva* (Vsup+Expadj)] (Picoli, 2020).

Existem diversos critérios para classificar uma expressão como EC ou CVS, porém há muitas que estão no limite entre elas, visto que possuem características de ambas as classes. Por exemplo, a frase "Ana está de barriga" é formada por um verbo-suporte (*estar*) seguido de adjetivo cristalizado (*de barriga*). Isto é, por ter características de ECs e CVSs, pode-se afirmar que a construção *estar de barriga* está no limite entre as duas categorias (Picoli, 2020). Esta é uma das dificuldades de reconhecer e diferenciar tais expressões. A Tabela 1 traz alguns exemplos de ECs e CVSs, que estão representadas nas Tábuas do Léxico-Gramática (TLGs) de Picoli (2020).

Tabela 1 – Exemplos de ECs e CVSs

ECs	**CVSs**
João tem palavra	João tem o nome sujo
João ficou cara a cara com Ana	João tem ares de poeta
Fazer faculdade não está com nada	A casa está de pernas pro ar

Fonte: o autor (2025).

Revista brasileira de iniciação científica
ISSN: 2359-232X

Uma forma de representar as ECs e CVSs é por meio das Tábuas do Léxico-Gramática (TLGs) (Gross, 1975), que organizam, em formato tabular, diversas expressões que são acompanhadas de suas possíveis variações. Além disso, a extração dessas expressões pode ser realizada por meio de Gramáticas Locais (GLs) (Gross, 1997), que consistem em regras manuais criadas para identificar padrões em textos. Apesar do esforço manual na construção de GLs, elas são úteis para extração de informação quando ainda não existe um corpus para treino, quando existe um padrão ou contexto muito bem definido nas expressões a serem reconhecidas e também para uso com técnicas de aprendizado de máquina em abordagens híbridas (Pirovani, 2019).

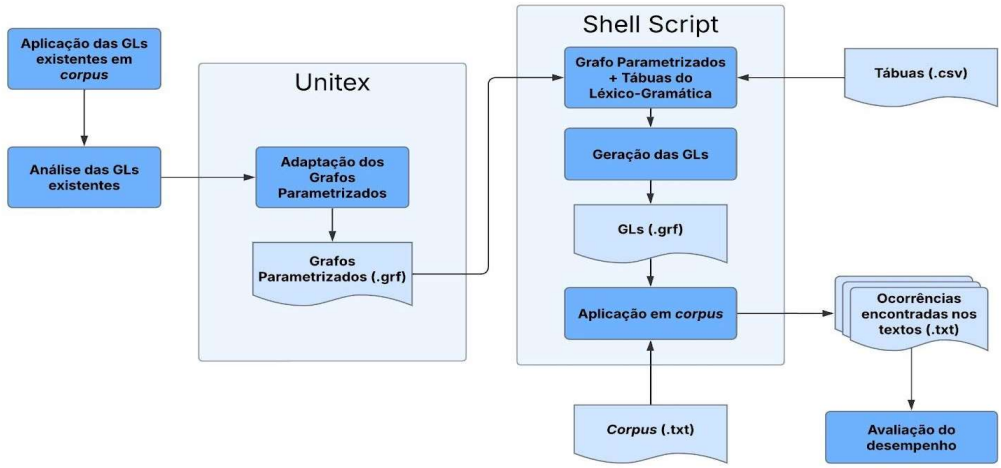
Diante do exposto, o objetivo deste estudo foi aprimorar as GLs existentes para o reconhecimento de ECs (Santiago, 2022) e CVSs (Vereau; Pirovani, 2023), visando melhorar seu desempenho e, consequentemente, contribuir para uma extração mais eficiente de informações em textos de linguagem humana.

Metodologia

O presente estudo adota um procedimento experimental, descritivo, de natureza aplicada e abordagem mista, combinando procedimentos qualitativos e quantitativos. A Figura 1 ilustra detalhadamente as principais etapas do trabalho, destacando desde a análise das GLs até os resultados gerados pelas GLs adaptadas. Além disso, são evidenciadas as ferramentas utilizadas para cada fase do processo.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 1 – Ilustração da metodologia utilizada neste trabalho



Fonte: o autor (2025).

Como ilustrado na Figura 1, a primeira etapa para alcançar o objetivo deste trabalho consistiu na aplicação das GLs de Santiago (2022) e Vereau e Pirovani (2023) ao *corpus* aTribuna, composto por textos jornalísticos de diferentes gêneros publicados pelo jornal do Espírito Santo A Tribuna[3]. Em seguida, foi conduzida uma análise dos resultados obtidos, com o objetivo de identificar melhorias que pudessem ser implementadas para aprimorar o desempenho das GLs. Essa análise foi realizada manualmente; para cada expressão, verificou-se o contexto em que estava inserida e, consequentemente, se correspondia a uma EC ou a uma CVS. Na ausência dessa correspondência, buscou-se identificar o padrão em que essas expressões se encaixavam, de modo a mitigar sua identificação pela GL.

A etapa seguinte consistiu na adaptação dos grafos parametrizados (GPs) construídos por Santiago (2022) e Vereau e Pirovani (2023) por meio da ferramenta Unitex[4], um conjunto de *softwares* livres para o PLN. Os GPs correspondem a GLs com parâmetros e serão detalhados na subseção *Grafos Parametrizados (GPs)*. As modificações foram realizadas de forma manual, com base nas observações obtidas na etapa anterior. Assim, para cada ponto identificado como passível de melhoria, foram feitas alterações nos GPs como forma de solução.

Posteriormente, as ferramentas do Unitex foram chamadas dentro de *shell scripts* com

[3] https://tribunaonline.com.br/
[4] https://unitexgramlab.org/pt/

o objetivo de automatizar a geração das novas GLs voltadas ao reconhecimento de ECs e CVSs. Para isso, utilizaram-se os GPs adaptados juntamente com as TLGs [que serão detalhadas na subseção *Tábuas do Léxico-Gramática (TLGs)*]. Além disso, nesses *scripts* a tarefa de aplicação das GLs adaptadas em *corpus* foi automatizada, gerando um arquivo .txt com o registro de todas as ocorrências identificadas nos textos

Na última etapa, realizou-se o cálculo da precisão, métrica que indica a proporção de acertos em relação ao total de construções reconhecidas, expressa em porcentagem (Pirovani, 2019). Dessa maneira, quanto maior for a precisão, menor será o número de falso-positivos. A fórmula a seguir representa o cálculo dessa medida.

$$Precisão = \frac{Quantidadedeexpressõesreconhecidascorretamente}{Quantidadedeexpressõesreconhecidas}$$

Tábuas do Léxico-Gramática (TLGs)

O modelo teórico-metodológico do Léxico-Gramática foi desenvolvido por Gross (1975) com o objetivo de explicar o funcionamento dos elementos vocabulares na estrutura das frases, fundamentando-se nos conceitos relacionados ao uso e às possíveis variações das palavras (Picoli, 2020). Esse modelo propõe um procedimento sistemático para a descrição e organização dos dados linguísticos, por meio de tábuas ou matrizes, nas quais as linhas correspondem às entradas lexicais e as colunas representam propriedades sintáticas e semânticas, como verbos, adjetivos e demais categorias (Calcia, 2022).

A primeira linha das tabelas exibe os tipos sintáticos, ou seja, a descrição do conteúdo de cada coluna. Na Figura 2, que exemplifica uma TLG, observa-se, por exemplo, o campo "N0 = Nhum" em que *N0* representa o sujeito (nome ou grupo nominal) e *Nhum,* um sujeito não humano. Cada célula é marcada com o símbolo "+", quando a propriedade é aceitável e com "-", quando é inaceitável.

Figura 2 – Tábua do Léxico-Gramática

Revista brasileira de iniciação científica
ISSN: 2359-232X

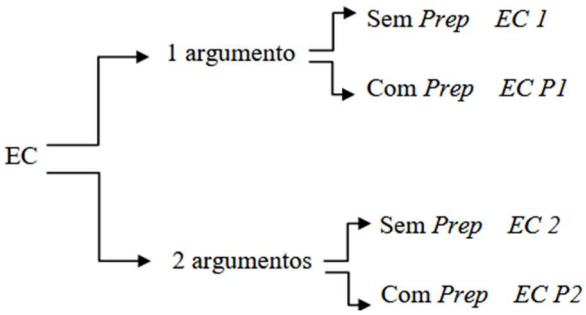
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	N0=:Nhum	estar (prep)	ficar (prep)	ter	continuar (prep)	viver (prep)	andar (prep)	manter	ganhar	possuir	DET=:E	DET=:Indef	N	ADJ	PREP/CONJ	N	V	Prep	N1=:Nhum	N1=:F	Comparação	Intensificação	Significado	Exemplo
2	+	+	-	+	+	-	-	-	-	-	+	-	estômago	<E>	para	<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).

Santiago (2022) e Vereau e Pirovani (2023) usaram as TLGs desenvolvidas por Picoli (2020) para construir os GPs e gerar as GLs. Neste trabalho utilizaram-se as mesmas tábuas, sendo quatro para as ECs (EC-1, EC-2, EC-P1 e EC-P2) e seis para as CVSs (CVSs-1, CVSs-1cop, CVSs-2, CVSs-2cop, CVSs-P1 e CVSs-P2). A Figura 2 ilustra parte da tábua CVSs-2 de Picoli (2020).

No total, as tábuas possuem 560 construções formadas com os verbos *ser*, *estar*, *ficar* e *ter* e são subdivididas de acordo com a quantidade de argumentos da frase e se possuem ou não preposição entre o verbo e o complemento fixo (Picoli, 2020). A Figura 3 ilustra essa subdivisão aplicada às tábuas de ECs.

Figura 3 – Subdivisão das ECs



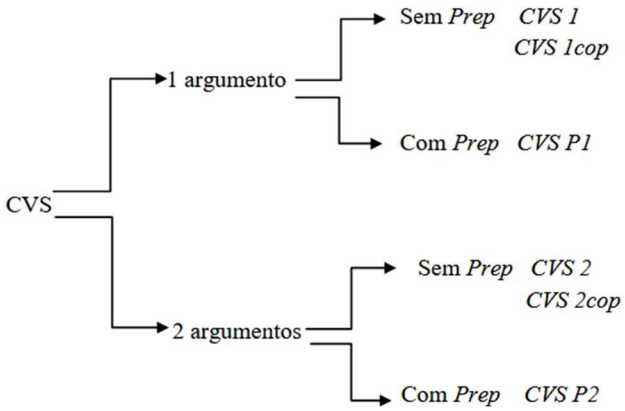
Fonte: Picoli (2020).

A Figura 4, por sua vez, detalha a subdivisão das CVSs. A classificação utiliza os mesmos critérios de divisão da Figura 3 (quantidade de argumentos e ausência ou presença de preposição), mas introduz uma distinção adicional para as construções não preposicionadas: a natureza copulativa, que se refere aos casos em que o verbo exerce função de ligação, estabelecendo uma relação entre o sujeito e um predicativo que expressa uma característica,

Revista brasileira de iniciação científica
ISSN: 2359-232X

estado ou identidade. Essa diferenciação resulta nas tábuas CVSs1 e CVSs2 e em suas respectivas variantes, CVSs1-cop e CVSs2-cop, que se caracterizam pelo uso de um verbo de cópula.

Figura 4 – Subdivisão das CVSs



Fonte: Picoli (2020).

Para aplicar as TLGs no processo de geração de GLs, é necessário adaptá-las a um formato interpretável pelo computador. Para isso, utilizam-se Grafos Parametrizados (GPs), que mapeiam estruturas matriciais para autômatos finitos, permitindo o reconhecimento das estruturas linguísticas.

Grafos Parametrizados (GPs)

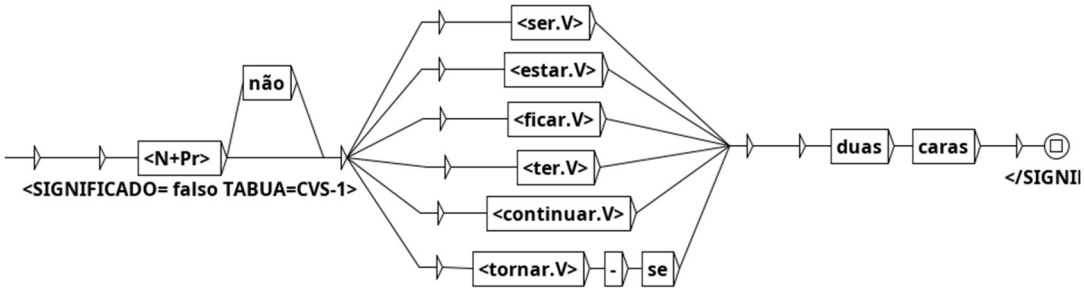
Por definição, as GLs são autômatos ou gramáticas de estados finitos utilizados para representar conjuntos de expressões pertencentes a uma linguagem natural (Gross, 1997). Essas gramáticas possibilitam o reconhecimento de expressões em textos e são amplamente empregadas em tarefas de Reconhecimento de Entidades Nomeadas (REN). Essa tarefa consiste em identificar e categorizar automaticamente entidades, como pessoas, empresas, organizações e locais (Silva, 2022). As GLs podem ser utilizadas em metodologias linguísticas, como a adotada neste trabalho, ou em abordagens híbridas que integram técnicas de aprendizado de máquina.

No Unitex, elas são modeladas por meio de grafos, conforme demonstrado na Figura 5, que ilustra a identificação de expressões com a seguinte configuração: [Nome Próprio (código

Revista brasileira de iniciação científica
ISSN: 2359-232X

<N+Pr> no Unitex) + não (opcional) + verbo *ser*, *estar* seguido de preposição (código <PREP> no Unitex), ficar, *ter, continuar* ou *tornar* (código <Verbo.V> no Unitex, como <ter.V>) + duas caras]. Um exemplo de sentença reconhecida por esse grafo é *João é duas caras.*

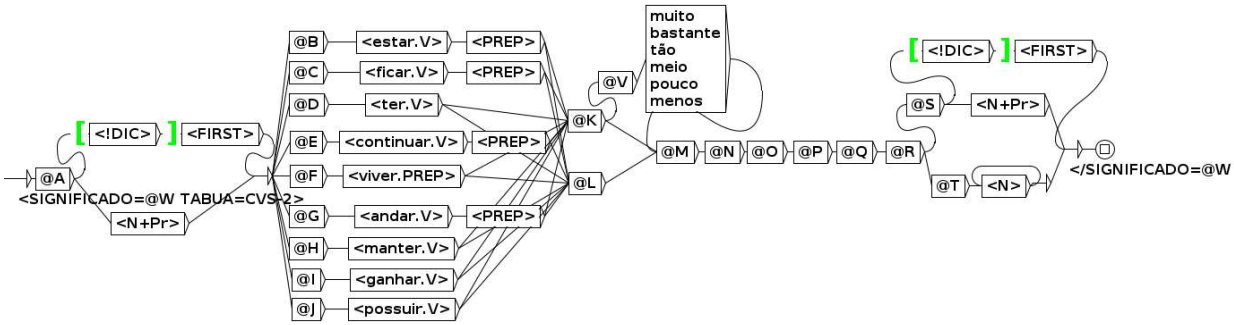
Figura 5 – Exemplo de GL no Unitex.



Fonte: o autor (2025).

Os GPs correspondem a GLs que utilizam variáveis para referenciar as colunas de uma TLG. Nessas representações, variáveis indicadas por @[Letra maiúscula] remetem às colunas da tábua, em ordem crescente (Paumier, 2021). Logo, @A representa a primeira coluna, @B a segunda e assim sucessivamente. Por meio dos GPs, torna-se possível gerar GLs automaticamente para todas as expressões representadas nas TLGs, ou seja, uma GL é gerada para cada linha da tábua correspondente ao grafo parametrizado (GP). A Figura 6 ilustra o GP elaborado com base nas TLGs da Figura 2.

Figura 6 – Grafo parametrizado para Tábua CVSs-2.



Fonte: Silva; Silva; Pirovani (2024).

A partir dos GPs, é possível gerar subGLs para cada linha das TLGs, assegurando que todos os registros possuam uma GL correspondente. Essa etapa é essencial para viabilizar a

aplicação da modelagem linguística em *corpora*.

Corpora

Os *corpora* utilizados para aplicar as GLs foram aTribuna e um *corpus* do PARSEME[5] (PARSing and Multi-word Expressions) (PARSEME, 2025). O *corpus* aTribuna contém 45.907 textos jornalísticos publicados pelo jornal local, A Tribuna do Espírito Santo, distribuídos em 21 classes, representando as várias seções do jornal, incluindo diferentes gêneros como política, economia, cultura, tecnologia etc.. A vantagem de utilizá-lo neste estudo foi a ampla variedade de expressões idiomáticas. Esse *corpus* não possui qualquer tipo de anotação linguística, ou seja, é necessário realizar a análise manual de todas as expressões reconhecidas para avaliação de desempenho. Apesar disso, é possível estabelecer uma comparação com trabalhos anteriores (Santiago, 2022 & Vereau; Pirovani, 2023) que utilizaram o mesmo *corpus*.

O PARSEME é uma rede científica interdisciplinar dedicada ao estudo das expressões multipalavras (Multiword Expressions – MWEs), reunindo especialistas de diversas áreas, como linguistas computacionais e cientistas da computação, dentre outros. O principal objetivo da organização é enfrentar os desafios que as expressões multipalavras representam para o PLN, promovendo a colaboração entre pesquisadores e o desenvolvimento de recursos para sua identificação e tratamento.

Uma das iniciativas da organização é a Tarefa Compartilhada[6], um esforço coletivo da comunidade de PLN para identificar automaticamente as MWEs. Seu principal objetivo é criar um *corpus* anotado com MWEs em diversos idiomas e avaliar os sistemas participantes por meio de métricas estatísticas. Para isso, foram realizadas edições ao longo de alguns anos, nas quais equipes colaboradoras disponibilizaram *corpora* de treinamento e teste em várias línguas, além de desenvolverem sistemas de PLN voltados à identificação automática dessas expressões.

Neste estudo, foi utilizado o *corpus* da edição 1.2 do PARSEME (Guillaume, 2020), que é composto por 19 textos, fornecidos no formato CONLL-U. Assim como o *corpus* aTribuna, os textos não possuem anotação linguística, impossibilitando a comparação automática das concordâncias obtidas na realização da avaliação de desempenho. Nesse caso,

[5] https://gitlab.com/parseme/corpora/-/wikis/home
[6] https://typo.uni-konstanz.de/parseme/index.php/results/shared-task

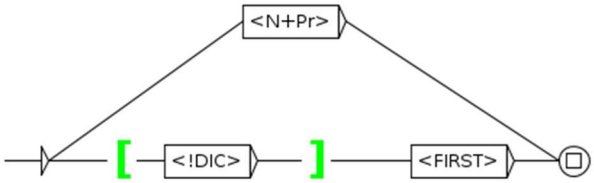
Revista brasileira de iniciação científica
ISSN: 2359-232X

também não havia resultados anteriores para comparação.

Análise dos Dados e Resultados

A aplicação das GLs existentes ao *corpus* aTribuna mostrou que alguns substantivos próprios – como nome, sobrenome e organização – não eram corretamente reconhecidos pelos dicionários do Unitex, ou seja, não eram detectados por meio da entrada lexical <N+Pr>. Como medida corretiva, foi desenvolvido um subgrafo capaz de reconhecer palavras ausentes no dicionário (código <!DIC> no Unitex) e que começam com letra maiúscula (código <FIRST> no Unitex), como ilustrado na Figura 7.

Figura 7 – Subgrafo que reconhece substantivos próprios.



Fonte: Silva; Silva; Pirovani (2024)

Com essa modificação, o número de ocorrências de ECs aumentou em 17, das quais 11 (65%) foram classificadas como verdadeiro-positivos e 6 (35%) como falso-positivos. Para as CVSs, o acréscimo foi de 18 ocorrências, todas corretamente identificadas como verdadeiro-positivos, indicando uma melhoria nos resultados.

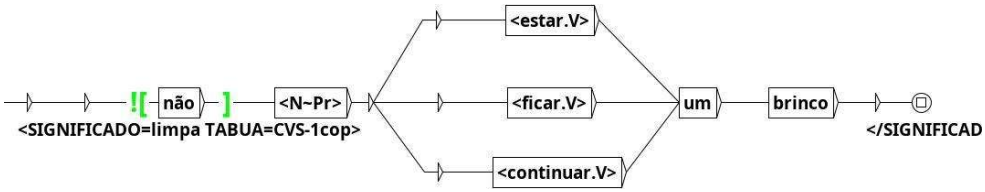
Contudo, outra adversidade foi observada, também relacionada ao dicionário do Unitex. O problema decorre da classificação incorreta de algumas palavras, que resultaram em falso-positivos. Por exemplo, a expressão “Não é o máximo”, com significado de ótimo, foi erroneamente reconhecida pela GL como CVSs, pois o dicionário classificou a palavra “não” também como substantivo masculino, além de advérbio, o que levou ao reconhecimento da palavra como sujeito humano (<N>) ou sujeito não humano (<N~Pr>). Isso ocorreu porque na aplicação de seus dicionários, o Unitex atribui todas as possíveis classificações gramaticais às palavras.

Para mitigar essa questão, foi preciso inserir um nó que impedisse o reconhecimento de construções iniciadas pela partícula negativa "não". A Figura 8 exemplifica essa solução,

Revista brasileira de iniciação científica
ISSN: 2359-232X

mostrando a inclusão de um nó com contexto negativo à direita em uma GL. Um contexto à direita é definido ao limitar o nó com [e], que representa, respectivamente, o início e o fim de um contexto (Paumier, 2021). O contexto negativo à direita ocorre de maneira em que o começo do contexto seja encontrado e percorrido, e caso atinja o estado final do contexto, a expressão é considerada como uma falha. Por contrário, se não for possível alcançar o estado final, continua-se explorando o resto da expressão após o fim do contexto. O contexto negativo à direita pode ser adicionado em qualquer lugar da GL, inclusive no começo (Paumier, 2021), como mostrado na Figura 8.

Figura 8 – Exemplo do nó adicionado, com contexto negativo à esquerda, para não reconhecer expressões iniciadas por “não”.



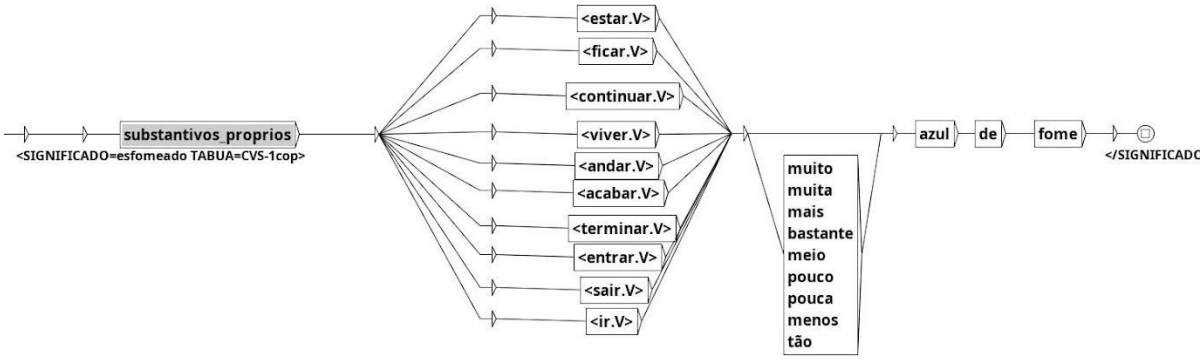
Fonte: o autor (2025).

Essa modificação resultou em uma redução de 32 ocorrências para as CVSs. Dentre elas, 8 (25%) eram verdadeiro-positivos e 24 (75%) falso-positivos. Embora 8 expressões reconhecidas corretamente tenham sido removidas, a diminuição significativa no número de falso-positivos compensou essa perda, tornando o reconhecimento mais preciso. Para as ECs, a modificação teve impacto mínimo, resultando na remoção de apenas uma ocorrência, que era classificada como falso-positivo, por ser a única expressão reconhecida que iniciava com “não”.

Ao analisar as TLGs desenvolvidas por Picoli (2020), identificou-se a presença de propriedades transformacionais de intensificação e comparação, que não haviam sido implementadas nos GPs utilizados em trabalhos anteriores. As transformações de intensificação incluem advérbios como *pouco*, *meio* e *muito*. As comparativas consistem na inserção de estruturas como *menos __ do que, mais __ do que, tanto ___ quanto e tão ___ quanto* (Picoli, 2020). Por exemplo: “João tem muita sede de vingança” e “João tem mais sede de vingança do que Maria”. Ambas as modificações foram implementadas nos GP. A Figura 9 ilustra a aplicação de propriedades de intensificação.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 9 – GL com transformações de intensificação.

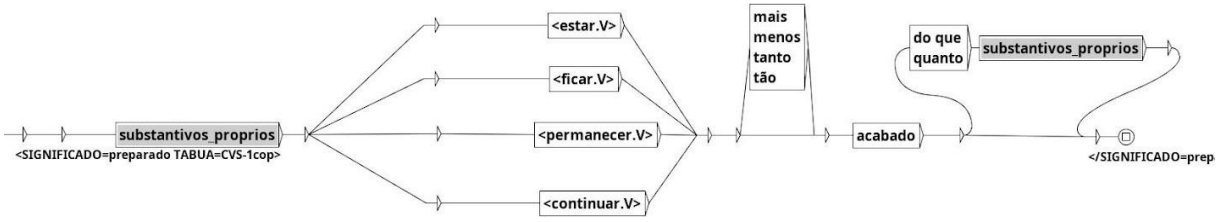


Fonte: o autor (2025).

Nas transformações voltadas à intensificação, foram utilizados os advérbios de intensidade mais comuns em textos, considerando que o Unitex não dispõe de uma marcação específica para essa categoria, contendo apenas a tag <ADV>, que abrange todos os advérbios, independentemente de sua função semântica. Ainda que esses advérbios apareçam com frequência entre o verbo e o complemento, observou-se também sua ocorrência antes de adjetivos, como no exemplo: “Maria ficou com o rosto bastante vermelho”. Por esse motivo, os grafos foram ajustados para permitir, opcionalmente, a inserção desses advérbios imediatamente antes de adjetivos.

De modo semelhante, para as transformações de comparação, foram adicionadas apenas as estruturas descritas por Picoli (2020), por serem as mais recorrentes em textos, conforme ilustrado na Figura 10. No entanto, essa modificação não adicionou novas ocorrências para as ECs e reconheceu apenas três novas expressões para as CVSs. Apesar da baixa quantidade, todas foram corretamente reconhecidas como verdadeiro-positivos.

Figura 10 – GL com transformações de comparação.

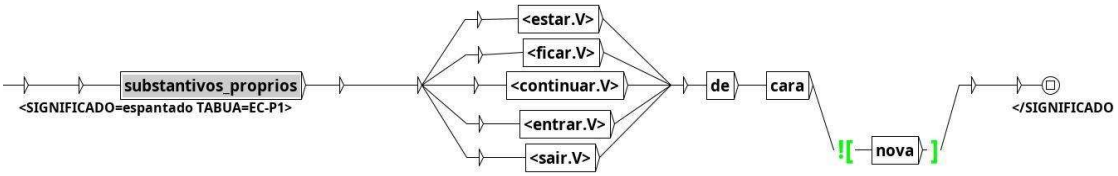


Revista brasileira de iniciação científica
ISSN: 2359-232X

Fonte: o autor (2025).

Na análise dos resultados das ECs, detectou-se um problema envolvendo a expressão *estar de cara*, usada no sentido de *espanto*. Essa construção estava sendo equivocadamente reconhecida em contextos diferentes, como em *"*Nestlé está de cara*"*, em que o sentido se aproxima de *mudança*. Com base na observação do *corpus*, verificou-se que essas ocorrências apareciam com frequência seguidas do termo *"nova"*. Para evitar esse tipo de falso-positivo, foi introduzido um nó com restrição de contexto à direita, impedindo esse padrão indesejado de reconhecimento, conforme apresentado na Figura 11.

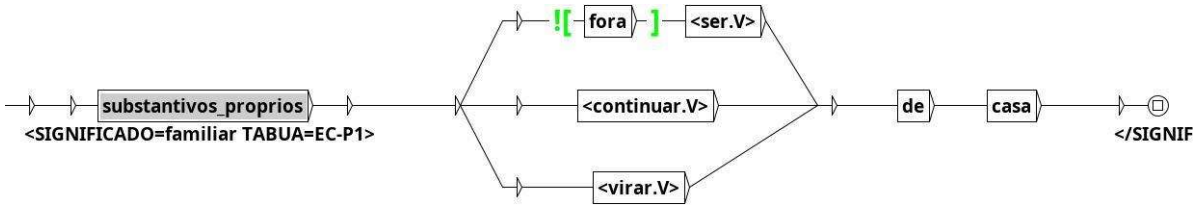
Figura 11 – GL que reconhece expressões *estar de cara*.



Fonte: o autor (2025).

Da mesma forma, a expressão *ser de casa*, com significado de *familiar*, também apresentou um valor significativo de falso-positivos. Dentre as 126 ocorrências totais de ECs, 29 (23%) foram classificadas erroneamente devido a essa expressão. O problema ocorreu devido ao dicionário do Unitex, que atribui a uma palavra todas as suas classificações gramaticais possíveis, como dito anteriormente. Como consequência, expressões como "Comemorar a vitória fora de casa" foram reconhecidas indevidamente, uma vez que o dicionário interpretou "fora" como verbo. Como solução, foi inserido um nó com contexto negativo à direita, impedindo o reconhecimento de "fora" quando antecede o verbo "ser", conforme ilustrado na Figura 12.

Figura 12 – GL que reconhece expressões *ser de casa*.

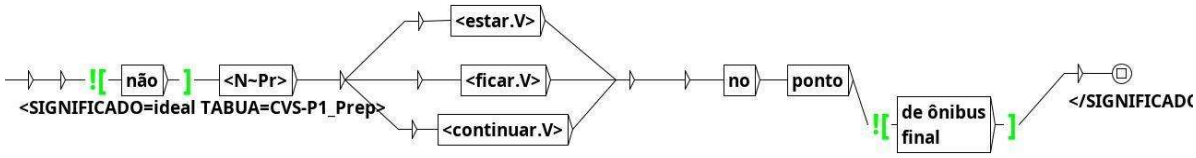


Revista brasileira de iniciação científica
ISSN: 2359-232X

Fonte: o autor (2025).

Para as CVSs, a construção *estar no ponto,* com sentido de *ideal/pronto*, destacou-se com o número de falso-positivos. Das 20 ocorrências analisadas, 15 (75%) foram classificadas incorretamente. A análise do contexto revelou que a palavra "ponto" era frequentemente seguida pelas expressões "de ônibus" e/ou "final", indicando o significado de estar no ponto final de ônibus. Para resolver o problema, foi criado um nó com contexto negativo à direita, indicando que a palavra "ponto" não pode ser sucedida pelo conjunto de palavras "de ônibus" e/ou "final", como ilustrado na Figura 13.

Figura 13 – GL que reconhece expressões no *ponto*.



Fonte: o autor (2025).

Também foram identificadas algumas adversidades nas TLGs que influenciaram nos resultados obtidos. Por exemplo, algumas GLs geradas estavam desconexas, ou seja, sem a possibilidade de alcançar o estado final, devido à ausência da marcação de aceitabilidade (+) na tábua. Ademais, observaram-se inconsistências na representação de contrações envolvendo preposições e determinantes. Por exemplo, a contração "do", formada pela preposição "de" e pelo determinante "o". No entanto, algumas tábuas tratavam essas palavras separadamente, resultando em construções incorretas, como "de o" em vez de "do". Com essa correção, foi possível reconhecer um número maior de ocorrências.

A partir da implementação das melhorias nos GPs, as GLs foram geradas novamente e aplicadas ao *corpus* aTribuna. Foram identificadas 400 expressões no total, das quais 117 correspondem a ECs e 283 a CVSs. Para as ECs, obteve-se uma precisão de 55%, enquanto para as CVSs a precisão alcançou 79%.

A Tabela 2 compara os resultados obtidos em Santiago (2022) com aqueles alcançados pelas GLs adaptadas neste estudo. Observa-se que, embora o número total de ECs reconhecidas

a mais tenha sido pequena (110 contra 117), as GLs adaptadas obtiveram um aumento expressivo no número de verdadeiros-positivos (37 para 64) e uma redução proporcional nos falsos-positivos (73 para 53). Consequentemente, a precisão passou de aproximadamente 34% para 55%, evidenciando o impacto positivo das adaptações realizadas.

Tabela 2 – Comparação com os resultados de Santiago (2022).

GLS	ECs RECONHECIDAS	VERDADEIRO-POSITIVOS	FALSO-POSITIVOS	PRECISÃO (%)
Santiago (2022)	110	37	73	≈ 34%
GLs adaptadas	117	64	53	≈ 55%

Fonte: o autor (2025).

Para as CVSs, a Tabela 3 apresenta a comparação entre os resultados obtidos por Vereau e Pirovani (2023) e aqueles alcançados pelas GLs adaptadas. Nesse caso, o ganho de desempenho foi ainda maior: o número de verdadeiros-positivos aumentou de 111 para 224, enquanto os falsos-positivos diminuíram de 84 para 59, resultando em um aumento da precisão de 57% para 79%.

Tabela 3 – Comparação com os resultados de Vereau e Pirovani (2023).

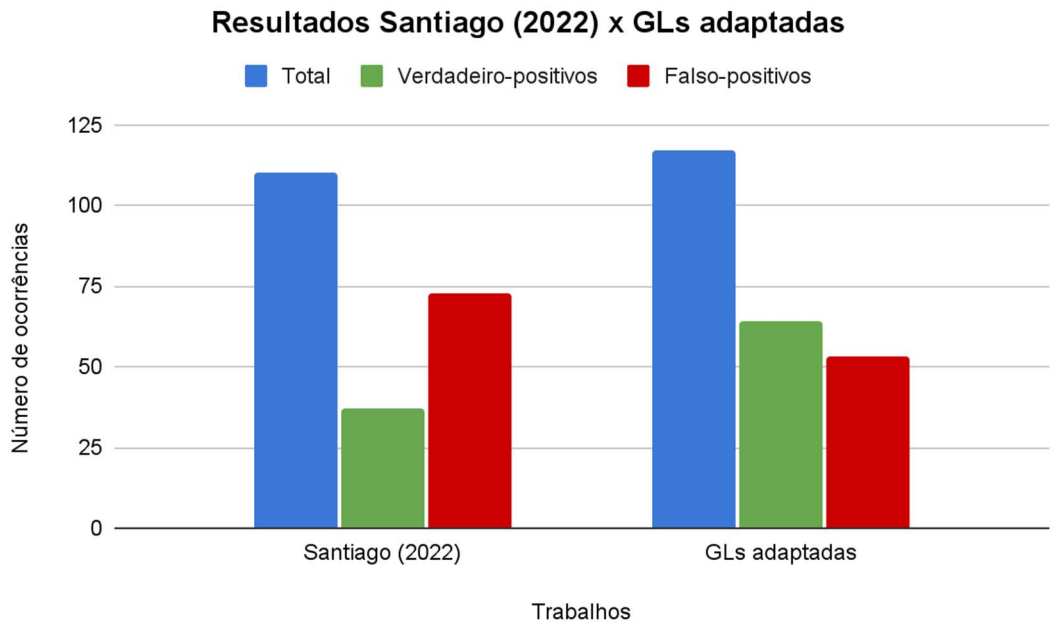
GLs	CVSs RECONHECIDAS	VERDADEIRO-POSITIVOS	FALSO-POSITIVOS	PRECISÃO (%)
Vereau e Pirovani (2023)	195	111	84	≈ 57%
GLs adaptadas	283	224	59	≈ 79%

Fonte: o autor (2025).

O Gráfico 1 apresenta uma comparação visual dos resultados obtidos por Santiago (2022) e os alcançados com as GLs adaptadas. Observa-se que, com as melhorias implementadas, não houve um grande aumento no número total de expressões reconhecidas. Porém, em ambos os casos (ECs e CVSs), os verdadeiro-positivos aumentaram e os falso-positivos diminuíram, melhorando a precisão e indicando que as alterações melhoraram o desempenho das GLs.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Gráfico 1 – Comparação visual com os resultados de Santiago (2022).

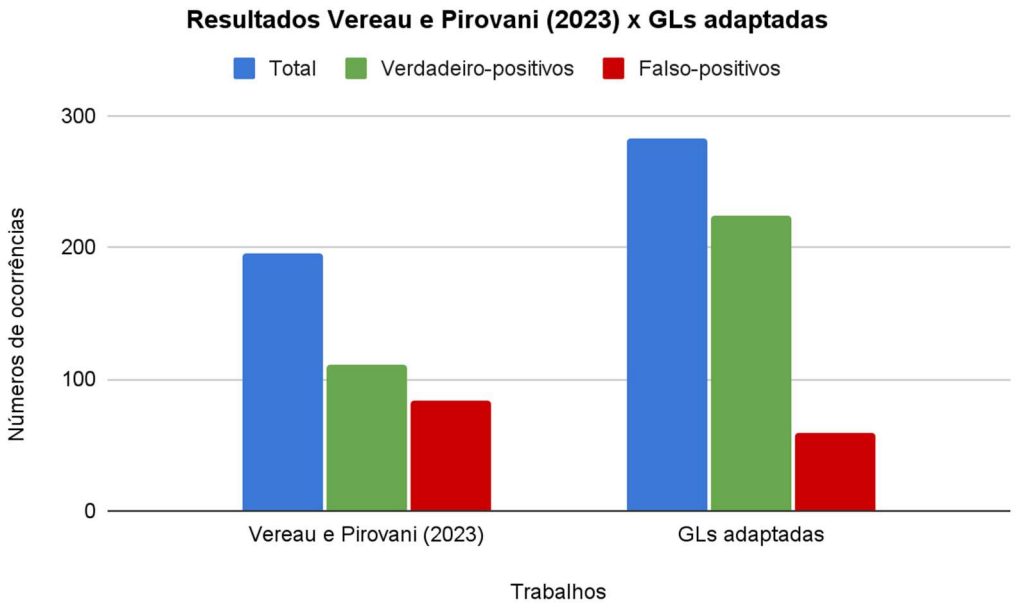


Fonte: o autor (2025).

O Gráfico 2 ilustra a comparação entre os resultados de Vereau e Pirovani (2023) e aqueles obtidos com as GLs adaptadas neste trabalho, seguindo o mesmo padrão observado anteriormente. Houve um aumento no número total de ocorrências e de verdadeiro-positivos, acompanhado por uma redução significativa na quantidade de falso-positivos. Contudo, no caso das CVSs, o crescimento no número de expressões reconhecidas foi mais expressivo, demonstrando que as adaptações realizadas aprimoraram a identificação dessas construções no *corpus* utilizado.

Gráfico 2 – Comparação visual com os resultados de Vereau e Pirovani (2023).

Revista brasileira de iniciação científica
ISSN: 2359-232X



Fonte: o autor (2025).

Para o *corpus* do PARSEME, os resultados foram obtidos apenas para uma TLG de cada tipo de construção linguística. Como o *corpus* utilizado não estava anotado, a análise foi feita manualmente, o que dificultou analisar todas as 1355 expressões reconhecidas. Assim, foram selecionadas as tábuas EC-P2 para as ECs e CVSs-1 para as CVSs. A Tabela 4 apresenta os resultados obtidos.

Tabela 4 – Resultados para o *corpus* PARSEME para as tábuas EC-P2 e CVSs-1

MÉTRICAS	**ECs (Tábua EC-P2)**	**CVSs (Tábua CVSs-1)**
Total	62	93
Verdadeiro-positivos	46	73
Falso-positivos	16	20
Precisão (%)	≈ 74%	≈ 78%

Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Os resultados exibidos na Tabela 4 podem ser considerados satisfatórios, dado o baixo número de falso-positivos identificados, totalizando apenas 16 para ECs e 20 para CVSs. Grande parte dessas ocorrências estão diretamente relacionadas à dificuldade de contemplar todos os contextos possíveis, ou seja, a ausência de um contexto padronizado. Por exemplo, a frase "As vias de comunicação são em geral conveniência de localidades, sem se *ter em vista* os lugares melhores, mais fáceis e mais econômicos", em que a EC *tendo em vista* deveria ser interpretada como *almejar*, porém, nesse caso, assume o significado de *considerar* e/ou *levar em conta.* Para as CVSs ocorre o mesmo problema, como em "Sofia passou a *ter meios*-irmãos que incluíam Nicolau e a duquesa Altburg", em que é reconhecida a sequência "meios-irmãos" para a expressão *ter meios*, com significado de *ter dinheiro,* o que ocorre devido à desconsideração do contexto das frases.

Dentre os falso-positivos das ECs, todas as concordâncias – ou seja, 100% – são geradas pela expressão *ter em vista*, com significado de *almejar,* como em "Qualquer desenvolvedor interessado pode também ajudar a melhorar o código do SecureDrop *tendo em vista* a sua natureza open-source". Já as CVSs não seguem um mesmo padrão: 13 expressões (65%) são relacionadas com *ter meios*, 2 (10%) com *ser duas caras*, 2 (10%) com *ter sangue quente*, 2 (10%) com *ter boa vontade* e, por fim, 1 (5%) com *ter a cabeça no lugar*.

Por fim, verificou-se que algumas expressões presentes nas tábuas apresentavam significados ausentes ou extremamente específicos, não correspondendo aos significados apresentados por Picoli (2020). A Tabela 5 apresenta exemplos dessas ocorrências, com o significado atual e uma sugestão de reformulação, visando a uma interpretação mais abrangente para diferentes contextos.

Tabela 5 – Sugestões de reformulação dos significados atribuídos às expressões das TLGs.

Expressão	**Significado atual**	**Sugestão de significado**
O coronelismo *estava com os dias contados*	doente	próximo do fim
A audiência *não é lá essas coisas*	-	mediano
Brasil *é o máximo*	presunçoso	ótimo

Eu *tenho vergonha na cara*	vergonha	decência

Fonte: o autor (2025).

As observações destacadas na Tabela 5 evidenciam limitações nos significados atribuídos às expressões nas TLGs, principalmente em casos de sentidos ausentes ou demasiadamente específicos que não correspondem às interpretações amplamente reconhecidas na literatura. Dessa forma, torna-se imprescindível revisar e adaptar esses significados para garantir maior precisão semântica e adequação ao uso em diferentes contextos textuais.

Discussão

Para o *corpus* aTribuna, os resultados se mostraram significativos, conforme demonstrado pelo aumento da precisão em ambas as construções. Para o reconhecimento das ECs, as GLs adaptadas neste trabalho apresentaram um aumento de 21% na precisão em relação à GL do trabalho de Santiago (2022), enquanto para CVSs esse aumento foi de 22%. Esses resultados sugerem uma melhoria nas GLs utilizadas para o reconhecimento dessas construções, evidenciado que as modificações implementadas foram eficazes para aprimorar a identificação dessas expressões.

No caso das ECs, a expressão que mais se destacou foi *ter cartaz*, com sentido de *ser uma pessoa famosa*, registrando 10 ocorrências, todas classificadas como verdadeiro-positivos. Já a expressão *estar limpo*, com significado de *sem dinheiro*, apresentou um alto índice de falso-positivos. Das 13 ocorrências identificadas, nenhuma foi reconhecida corretamente. Um exemplo é a expressão “O carro estava limpo”, em que *limpo* refere-se a ausência de sujeira, e não a falta de dinheiro. Esse problema decorre da falta de um contexto bem definido para essas expressões, dificultando a sua identificação correta.

As CVSs apresentaram resultados um pouco melhores em relação às ECs, destacando-se a expressão *ter os dias contados*, que apareceu 45 vezes, todas classificadas como verdadeiro-positivos. Entretanto, a expressão *ser duro* apresentou o maior número de falso-positivos, com 13 ocorrências, das quais 5 eram verdadeiro-positivos e 8 falso-positivos. Como mencionado anteriormente, não é possível implementar melhorias nesse caso em razão da inexistência de um padrão consistente que permita o reconhecimento dessas expressões.

É importante ressaltar que, ao comparar os resultados, observa-se que o número de reconhecimentos das ECs não apresentou um aumento significativo dos reconhecimentos quanto o das CVSs, com um acréscimo de apenas sete expressões. Acredita-se que essa dificuldade está diretamente relacionada à complexidade estrutural das ECs, também relatado por Santiago (2022), em seu trabalho.

Em relação ao *corpus* do PARSEME, vale destacar que ele abrange diversos tipos de construções, além das ECs e CVSs. Além disso, contempla uma ampla variedade de expressões, não se restringindo às expressões listadas nas TLGs utilizadas neste trabalho. Apesar dessas diferenças, os resultados obtidos foram positivos, alcançando uma precisão de 74% para as ECs e 78% para as CVSs, considerando exclusivamente as tábuas EC-P2 e CVSs-1.

A análise de ocorrências mostrou que, no caso das ECs, apenas dois tipos de expressões foram reconhecidas: *ter em mente* e *ter em vista*, totalizando 46 verdadeiro-positivos e 16 falso-positivos. Observou-se que todas as 16 ocorrências reconhecidas erroneamente estavam associadas à expressão *ter em vista*, pois, embora tenham sido reconhecidas, estavam sendo empregadas no sentido de *considerar*, e não no sentido de *almejar*, que seria o significado correto da expressão. No entanto, não foi possível realizar ajustes para mitigar esse problema, uma vez que essas expressões não compartilham um contexto comum que permitisse uma distinção mais precisa.

No caso das CVSs, a expressão *ter os dias contados* se destacou, aparecendo em 48 ocorrências (aproximadamente 52% do total), todas corretamente classificadas como verdadeiro-positivos. Isso indica que mais da metade das expressões reconhecidas pertenciam a esse tipo de expressão e foram identificadas sem erros. Contudo, verificou-se um equívoco na atribuição de significado para essa expressão. De acordo com a TLG, *ter os dias contados* é associada ao significado de *doente*, enquanto o sentido mais adequado seria *próximo do fim.* Esse problema reflete a tendência das tábuas de atribuírem significados excessivamente específicos a certas expressões, sem considerar a amplitude de interpretações possíveis em diferentes contextos.

Diante desses resultados, observa-se que, embora as modificações nas GLs tenham aprimorado o reconhecimento das ECs e CVSs, desafios ainda persistem, especialmente no que se refere à ambiguidade semântica e à ausência de padrões consistentes para algumas

expressões. O desempenho positivo alcançado, tanto no corpus *aTribuna* quanto no PARSEME, reforça a eficácia da abordagem adotada, mas também evidencia a complexidade inerente ao reconhecimento automático de ECs e CVSs. Isso comprova a dificuldade que os modelos computacionais enfrentam na interpretação de linguagem natural, especialmente devido às múltiplas possibilidades de significado que certas expressões podem assumir em diferentes contextos.

Considerações Finais

Este estudo teve como objetivo melhorar a identificação de Expressões Cristalizadas (ECs) e Construções com Verbo-Suporte (CVSs) por meio de alterações realizadas em Gramáticas Locais (GLs), utilizando as Tábuas do Léxico-Gramática (TLGs) construídas por Picoli (2020) em seu trabalho sobre ECs e CVSs. Além disso, foram empregados *shell scripts* para automatizar a geração e aplicação das GLs adaptadas nos *corpora*. Os resultados obtidos nos *corpora* aTribuna e PARSEME demonstraram que as modificações implementadas aumentaram a precisão no reconhecimento dessas expressões, especialmente no *corpus* aTribuna, onde foi obtido um acréscimo de 21% para ECs e de 22% para as CVSs.

Apesar da melhoria na identificação, algumas dificuldades persistiram. A ambiguidade semântica de certas expressões e a ausência de um contexto bem definido dificultaram a eliminação de falso-positivos. No caso das ECs, expressões como *estar limpo* apresentaram dificuldades na classificação, pois seu significado pode variar de acordo com o contexto, o qual não é bem definido nas ocorrências dos *corpora* utilizados. Da mesma forma, nas CVSs, observou-se que a expressão *ter os dias contados* foi corretamente identificada, mas o significado atribuído pela tábua não correspondia ao sentido das expressões reconhecidas.

Os resultados obtidos reforçam o potencial das GLs utilizadas neste estudo. Foram adaptados dez GPs, que geraram 560 GLs, isto é, foi gerado uma GL para cada expressão da TLG, que é capaz de reconhecer diversas variações linguísticas dessas expressões. As GLs adaptadas podem ser utilizadas individualmente ou em conjunto com técnicas de aprendizado de máquina para reconhecimento de ECs e CVSs. Os resultados também evidenciam os desafios do Processamento de Linguagem Natural (PLN), especialmente no reconhecimento automático de construções linguísticas. A dificuldade de capturar pequenas diferenças semânticas e

contextuais comprova a necessidade contínua de aprimoramento dos modelos de reconhecimento.

Para trabalhos futuros, sugere-se o aprimoramento das TLGs, de modo a abranger melhor as variações semânticas das expressões. Também pretende-se utilizar a mesma metodologia (construir GPs e gerar as GLs) para outras TLGs disponíveis na literatura.

Além disso, uma possibilidade de melhoria é combinar a aplicação dos dicionários do Unitex com uma ferramenta que realize o *POS-Tagging* para evitar atribuir diversas classificações a uma mesma palavra, desconsiderando o contexto em que ela está inserida. Ademais, torna-se relevante explorar uma abordagem híbrida, combinando regras linguísticas com técnicas de aprendizado de máquina, visando aumentar o reconhecimento e reduzir a ocorrência de falso-positivos.

Referências

CALCIA, Nathalia Perussi. “Este foi outro aspecto que sofreu uma avaliação positiva”: as construções conversas fazer-sofrer. **Estudos Linguísticos**, São Paulo, v. 51, n. 2, p. 579-594, ago. 2022. Disponível em: <https://hal.science/hal-04114700v1>. Acesso em: 07 out. 2025.

CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe. (Org.). **Processamento de Linguagem Natural:** Conceitos, Técnicas e Aplicações em Português. 3.ed. São Carlos: BPLN, 2024. Disponível em: <https://brasileiraspln.com/livro-pln/3a-edicao/>. Acesso em: 01 mar. 2025.

CAVALCANTI, Tatiana et al. Os limites da palavra e da sentença no processamento automático de textos. **Revista Brasileira de Iniciação Científica**, Itapetininga, v. 8, p. 1-21, 2021. Disponível em: https://periodicoscientificos.itp.ifsp.edu.br/index.php/rbic/article/view/348. Acesso em: 8 out. 2025.

GROSS, Maurice. **Méthodes en syntaxe:** Régime des constructions complétives. Paris: Hermann, 1975.

GROSS, Maurice. The Construction of Local Grammars. In: ROCHE, E; SCHABÈS, Y. (Org.) Finite-State Language Processing. Cambridge. **MIT Press**. 1997. p. 329-354. Disponível em: <https://shs.hal.science/halshs-00278316/PDF/MIT.pdf>. Acesso em: 05 jun. 2025.

GUILLAUME, Bruno et al. Morpho-syntactically annotated corpora provided for the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions (edition 1.2). LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University,2020. Disponível em: <http://hdl.handle.net/11234/1-3416>. Acesso em: 21 out. 2025.

PARSEME. PARSEME: PARSing and Multi-word expressions. 2025. Disponível em: <https://typo.uni-konstanz.de/parseme/index.php>. Acesso em: 21 out. 2025.

PAUMIER, Sébastien. **Unitex 3.3 User Manual**. 2021. Disponível em: <https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf>. Acesso em: 21 out. 2025.

PICOLI, Larissa. **Contínuo e limite entre Expressão Cristalizada e Construção com Verbo-suporte à luz do Léxico-Gramática.** Tese (Doutorado em Linguística). Universidade Federal de São Carlos, São Carlos, 2020.

SANTIAGO, Daniel Hand. Gramáticas Locais para Reconhecimento de Expressões Cristalizadas em Português. In: JORNADA DE INICIAÇÃO CIENTÍFICA DA UFES, 13., 2022, Vitória. **JIC.** 2022. Volume 13. Disponível em: <https://anaisjornadaic.sappg.ufes.br/desc.php?&id=19144>. Acesso em: 27 fev. 2025.

SAVARY, Agata et al. PARSEME multilingual corpus of verbal multiword expressions. In: MARKANTONATOU, Stella et al. (Org.). **Multiword expressions at length and in depth:** extended papers from the MWEs 2017 workshop. Berlin: Language Science Press, 2018, v. 2.

SILVA, Kailany Alves; SILVA, Thiago Tonelli da; PIROVANI, Juliana Pinheiro Campos. Gramáticas Locais para Expressões Cristalizadas e Construções com Verbo-Suporte. In: ENCONTRO DE INICIAÇÃO CIENTÍFICA, 28., 2024, São José dos Campos. **INIC**. 2024. p.1 – 6. Disponível em: <https://www.inicepg.univap.br/cd/INIC_2024/anais/arquivos/RE_0652_0480_01.pdf>. Acesso em: 07 out. 2025.

SILVA, Messias Gomes da. Reconhecimento de Entidades Nomeadas em Documentos de Editais de Compras utilizando Aprendizado Profundo. Dissertação (Mestrado em Computação Aplicada). Instituto Federal do Espírito Santo, Serra, 2022.

TAYLOR, Petroc. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028. **Statista**, 2025. Disponível em: <https://www.statista.com/statistics/871513/worldwide-data-created/?__sso_cookie_checker=failed>. Acesso em: 10 ago. 2025.

VALE, Oto Araújo. Expressões Cristalizadas: transparência e opacidade. **Revista Signótica**, Goiânia, v. 11, n. 1, p. 163-172, jan./dez. 1999. Disponível em: <https://www.revistas.ufg.br/sig/article/view/7282/5153>. Acesso em: 3 jun. 2025.

VEREAU, Luis Enrique Santos Prado; PIROVANI, Juliana Pinheiro Campos. Gramáticas Locais para Reconhecimento de Construções com Verbo Suporte em Português. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 14., 2023, Belo Horizonte. **STIL**. 2023. p. 347-351. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25469>. Acesso em: 21 out. 2025.