

DESENVOLVIMENTO E APRIMORAMENTO DE GRAMÁTICAS LOCAIS PARA A VERIFICAÇÃO GRAMATICAL DO PORTUGUÊS

Marcelo Vieira Gomes, Juliana Pinheiro Campos Pirovani

Universidade Federal do Espírito Santo/Departamento de Computação, Alto Universitário, S/N, Guararema - 29500-000 - Alegre-ES, Brasil, marcelo.v.gomes@edu.ufes.br, juliana.campos@ufes.br.

Resumo

Este trabalho apresenta a revisão e o aprimoramento de Gramáticas Locais (GLs) para a correção de erros de vírgula e crase na língua portuguesa, implementadas na ferramenta Unitex. Foram analisadas GLs construídas com essa finalidade, sendo identificadas fragilidades principalmente no critério de precisão, devido ao elevado número de falsos positivos. As reformulações buscaram restringir as regras com base na regência nominal e verbal, bem como em contextos sintáticos específicos, de modo a reduzir sinalizações indevidas. Os experimentos foram realizados em dois *corpus*: um *corpus* revisado, proveniente do trabalho de referência, e um *corpus* jornalístico do portal aTribuna. Os resultados demonstraram ganhos significativos em precisão, especialmente em textos do *corpus* jornalístico, ainda que com redução na abrangência. Conclui-se que as GLs reformuladas representam um avanço em termos de confiabilidade para a detecção automática de erros, ao mesmo tempo em que evidenciam a necessidade de futuras melhorias para ampliar sua cobertura.

Palavras-chave: Processamento de Linguagem Natural. Gramáticas Locais. Correção Gramatical. Língua Portuguesa.

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

Introdução

O Processamento de Linguagem Natural (PLN) tem se consolidado como uma área estratégica da Computação, com aplicações que vão desde a tradução automática até a correção gramatical. Em língua portuguesa, os desafios são ainda mais evidentes devido à complexidade da morfossintaxe e à variedade de fenômenos normativos que impactam diretamente a clareza e a correção textual. Entre esses fenômenos, destacam-se o uso da vírgula e da crase, cujas regras, apesar de bem estabelecidas na gramática normativa, são de difícil formalização em sistemas computacionais.

Diferentes abordagens têm sido propostas para o tratamento desses problemas. Modelos de aprendizagem de máquina, como os baseados em redes neurais de grande porte, vêm apresentando resultados promissores, mas ainda enfrentam dificuldades em manter precisão elevada na correção de erros gramaticais mais sutis. Penteado e Perez (2023), por exemplo, ao avaliarem o desempenho do GPT-3.5 e do GPT-4 na tarefa de correção gramatical em português, constataram que esses modelos apresentam alta cobertura, mas tendem a apontar falsos positivos, comprometendo a confiabilidade em aplicações práticas. Esse cenário reforça a importância de abordagens simbólicas mais controladas, capazes de lidar explicitamente com regras linguísticas.

Nesse contexto, destacam-se as Gramáticas Locais (GLs), implementadas por meio da ferramenta Unitex (PAUMIER, 2021), que permitem representar formalmente padrões da língua natural por meio de grafos. As GLs já têm sido utilizadas com sucesso em fenômenos como construções com verbo-suporte e expressões cristalizadas (VEREAU; PIROVANI, 2023; SANTIAGO, 2022), demonstrando sua versatilidade para lidar com estruturas sintáticas complexas. Aplicadas à vírgula e à crase, as GLs oferecem uma forma de sistematizar a norma culta em grafos processáveis, unindo teoria linguística e implementação computacional.

Este trabalho insere-se nesse esforço, partindo da proposta inicial de Lapa, Santos e Pirovani (2024), que desenvolveram GLs para correção de vírgula e crase em português. A análise desse estudo mostrou que, embora eficaz na identificação de alguns padrões, a abordagem apresentava baixa precisão devido ao excesso de sinalizações incorretas.

Assim, o presente artigo tem como objetivo revisar e aprimorar essas Gramáticas Locais, buscando reduzir falsos positivos e aumentar a confiabilidade das correções. Para tanto, foram implementados ajustes baseados em regência nominal e verbal, além de refinamentos no contexto sintático de aplicação das regras. A avaliação foi conduzida em *corpus* revisados e jornalísticos, permitindo observar tanto a evolução obtida quanto as limitações que ainda persistem.

Metodologia

Inicialmente, realizou-se uma revisão bibliográfica voltada ao Processamento de Linguagem Natural (PLN) em língua portuguesa, com ênfase na utilização do software Unitex¹ para construção e aplicação de Gramáticas Locais. Foram consultados trabalhos clássicos sobre Léxico-Gramática (GROSS, 1975; 1997), além de manuais técnicos do Unitex (PAUMIER, 2021) e pesquisas recentes que aplicam GLs em fenômenos linguísticos variados (VEREAU; PIROVANI, 2023; SANTIAGO, 2022).

Paralelamente, foi necessário recorrer a obras de referência da gramática normativa da língua portuguesa para fundamentar as regras implementadas nas Gramáticas Locais, especialmente no que se refere ao uso da vírgula e da crase. Para isso, consultou-se a Moderna Gramática Portuguesa de Evanildo Bechara (2019), que apresenta de forma sistemática e detalhada as normas de regência, acentuação e pontuação, servindo como base teórica para a elaboração e revisão das regras utilizadas neste trabalho.

Além da revisão bibliográfica, este trabalho tomou como ponto de partida a pesquisa de Lapa, Santos e Pirovani (2024), que desenvolveram Gramáticas Locais voltadas para a identificação de erros de crase e vírgula utilizando o Unitex. No total, foram elaboradas seis GLs para o tratamento da crase, nomeadas de 1 a 6, e oito GLs destinadas ao uso da vírgula, nomeadas de v1 a v8.

A análise desse estudo foi fundamental para compreender a estrutura inicial das GLs e suas limitações. Durante os testes realizados, observou-se que, embora a proposta conseguisse identificar determinados padrões de erros, apresentava fragilidades significativas no critério de precisão, uma vez que parte considerável das frases sinalizadas era corrigida de forma inadequada. Essa constatação orientou a revisão e o aprimoramento das regras, servindo de base metodológica para o desenvolvimento das novas GLs apresentadas nesta pesquisa.

As Gramáticas Locais reformuladas foram avaliadas em dois conjuntos de dados. O primeiro corresponde ao *corpus* utilizado no trabalho de Lapa, Santos e Pirovani (2024), composto por um conjunto de frases selecionadas, todas construídas de modo a apresentar erros que demandavam correção, seja no uso da crase ou da vírgula. Esse *corpus* foi revisado, de modo que algumas sentenças de vírgula que não se adequavam ao escopo foram descartadas, resultando em 38 frases ao invés das 40 originais, enquanto o *corpus* de crase permaneceu com 40 sentenças. Esse material revisado permitiu uma comparação direta com os resultados anteriores, mas com maior rigor na seleção das estruturas linguísticas analisadas.

Além desse *corpus* revisado, também foi utilizado um segundo conjunto de dados, constituído por 100 documentos jornalísticos do portal aTribuna². Esse material foi selecionado com o objetivo de ampliar a avaliação das Gramáticas Locais em cenários de maior variabilidade linguística e complexidade textual, distinto das frases controladas do primeiro *corpus*. Como textos jornalísticos tendem a passar por revisão gramatical antes da publicação, a expectativa era que apresentassem poucos ou nenhum erro de crase. Dessa forma, esse *corpus* serviu principalmente para verificar a robustez das GLs frente a dados reais, avaliando a ocorrência de falsos positivos e a capacidade das regras em manter a precisão em contextos de escrita mais extensa e natural.

Para a avaliação, foram utilizadas as métricas de precisão, abrangência e medida-F, amplamente adotadas em Processamento de Linguagem Natural. A precisão corresponde à proporção de sinalizações corretas em relação ao total de casos identificados; a abrangência mede a quantidade de sinalizações corretas em comparação ao total de casos existente no *corpus*; e a medida-F combina os dois indicadores, oferecendo uma visão equilibrada entre precisão e cobertura. A aplicação dessas métricas nos dois *corpus* permitiu analisar de forma comparativa o desempenho das GLs reformuladas em relação às originais, bem como levantar aspectos que indicam a necessidade de ajustes futuros.

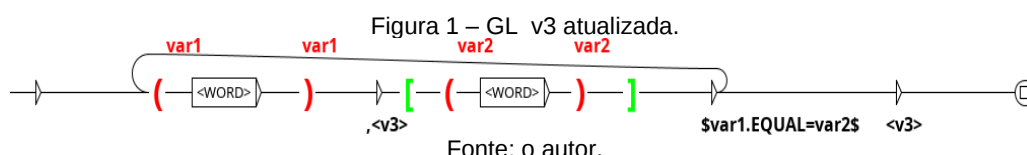
¹ <https://unitexgramlab.org/pt>

² <https://tribunaonline.com.br>

Resultados

Em relação às GLs que identificam erros na colocação da vírgula, as primeiras modificações relevantes ocorreram na GL v3, responsável por lidar com repetições de palavras e enumerações. A versão inicial corrigia apenas pares de repetições, mas não capturava sequências mais longas, o que resultava em saídas incorretas como em “A vitória será minha minha e de mais ninguém!” que era corrigida como “A vitória será minha, minha minha e de mais ninguém!” ao invés de “A vitória será minha, minha, minha e de mais ninguém!”. Após a reformulação, a regra passou a reconhecer corretamente ocorrências desse tipo, evitando sinalizações excessivas e garantindo maior consistência.

O grafo apresentado na Figura 1 implementa essa solução. Nele, a tag lexical <WORD> representa qualquer palavra, e a equação entre variáveis (var1.EQUAL=var2) garante que apenas repetições idênticas sejam reconhecidas. O uso do recurso de contexto (colchetes verdes) junto com o loop retornando ao início do grafo permite que a GL identifique não apenas pares, mas também sequências maiores de palavras iguais.



Outro ponto de destaque foi a GL v8, voltada para conjunções adversativas deslocadas, que passou a exigir a presença de um sintagma nominal antes da conjunção, de modo a indicar a necessidade de vírgulas antes e depois do conector, como em “Os alunos, contudo, não estudaram para o teste”. Essa modificação eliminou falsos positivos e trouxe ganhos expressivos na precisão das correções.

Também foram feitos ajustes pontuais em outras regras, como na v6, em que se acrescentou maior contexto para contornar casos em que a vírgula era inserida incorretamente, e na v7, que foi modificada para evitar o uso da vírgula no início de frases.

Com essas alterações, as GLs de vírgula identificaram corretamente 17 ocorrências em um total de 38 frases avaliadas, alcançando precisão de 100%, abrangência de 44,7% e um ganho de 34,4% em medida-F em relação à GL original. A Tabela 1 mostra a comparação entre as Gramáticas Locais para a verificação de vírgula.

Tabela 1 – Comparação das GLs para vírgula.

Modelo	Precisão	Abrangência	Medida-F
GL inicial	53,8%	18,4%	27,5%
GL final	100%	44,7%	61,9%

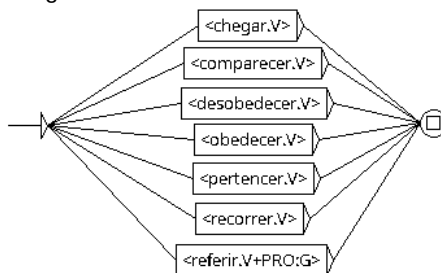
Fonte: elaborado pelo próprio autor.

A análise das Gramáticas Locais voltadas para a correção de crase revelou que este foi o fenômeno mais problemático, em grande parte devido à amplitude excessiva de algumas regras. As versões iniciais das GLs 1 e 5 partiam da definição Verbo + a + Substantivo Feminino, o que se mostrou genérico demais e resultava em falsos positivos. Em casos como “A funcionária apenas conseguiu a promoção”, a regra indicava crase de forma indevida, já que o verbo conseguir é transitivo direto. Para corrigir esse problema, as regras foram reformuladas com base na regência verbal, apoiadas por uma pequena lista de verbos transitivos indiretos (Figura 2) que sempre, ou quase sempre, exigem a preposição “a”.

No Unitex, a marcação <VERBO.V> indica que os itens correspondem a verbos e que o sistema deve reconhecer todas as suas flexões registradas no dicionário. Assim, ao utilizar “chegar.V”, por exemplo, a GL passa a identificar automaticamente formas como chego, chegava, chegaram, entre outras.

A Ciência do NANO e seu impacto transformador no MACRO

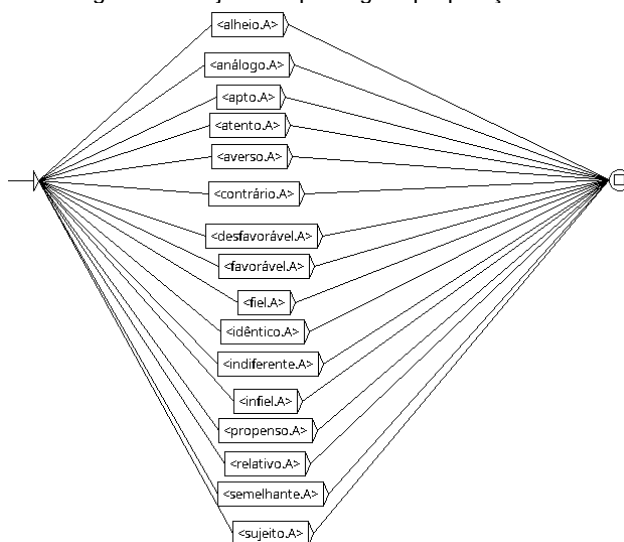
Figura 2 – Verbos transitivos indiretos.



Fonte: o autor.

Problema semelhante foi observado na GL 2, que identificava indiscriminadamente estruturas do tipo Adjetivo Feminino + a + Substantivo Feminino. A reformulação restringiu o padrão para adjetivos que efetivamente regem a preposição “a”, complementados por uma lista de adjetivos levantada durante o trabalho (Figura 3). A GL 3 também foi ampliada, passando a contemplar os pronomes demonstrativos aquele e aquela, além de integrar o grafo de verbos transitivos indiretos.

Figura 3 – Adjetivos que regem preposição “a”.



Fonte: o autor.

Nos 40 itens do *corpus* revisado de crase, a GL reformulada identificou 8 ocorrências, das quais 7 foram corretas e 1 incorreta, resultando em precisão de 87,5%, abrangência de 17,5% e medida-F de 28,8%. A GL original desenvolvida por Lapa, Santos e Pirovani (2024) havia identificado 24 ocorrências, com 16 acertos e 8 erros, alcançando precisão de 66,7%, abrangência de 40,0% e medida-F de 50,0%. Apesar de a GL reformulada ter identificado um número menor de ocorrências no *corpus* revisado de crase, sua maior precisão confirma uma abordagem mais conservadora e confiável. A Tabela 2 mostra a comparação das GLs para a verificação do uso da crase.

Tabela 2 – Comparação das GLs para crase.

Modelo	Precisão	Abrangência	Medida-F
GL inicial	66,7%	40,0%	50,0%
GL final	87,5%	17,5%	28,8%

Fonte: elaborado pelo próprio autor.

A Ciência do **NANO** e seu impacto transformador no **MACRO**

No *corpus* jornalístico do portal aTribuna, os testes foram realizados em dois momentos. Em um primeiro documento de menor extensão, a GL reformulada não produziu sinalizações de correção, o que era esperado em um texto revisado e com baixo índice de desvios gramaticais. Já a GL original indicou sete frases, todas incorretas, demonstrando baixa precisão nesse cenário. Em seguida, ampliou-se a análise para um conjunto de 100 documentos jornalísticos. Nesse caso, a GL reformulada sinalizou 8 frases, das quais 4 estavam corretas e 4 incorretas, mostrando que, apesar da tendência conservadora, ainda há espaço para aprimoramento. A GL original, por sua vez, sinalizou 1060 frases.

Devido ao elevado número de ocorrências, não foi possível realizar a validação manual de todas elas. Contudo, a inspeção preliminar de uma amostra sugere que a maioria corresponde a falsos positivos. Além disso, a natureza do *corpus*, textos jornalísticos revisados e a amplitude excessiva das regras reforçam essa interpretação. Esse contraste evidencia como as reformulações priorizaram a precisão e reduziram drasticamente o número de sinalizações indevidas, em oposição à abordagem original, que manteve abrangência ampla, mas com custo elevado em termos de confiabilidade.

Esses resultados demonstram que a estratégia de estreitamento das regras por regência foi eficaz na eliminação de falsos positivos, elevando significativamente a precisão, ainda que com redução na abrangência. Tal comportamento era esperado, já que o aumento da restrição das regras tende a reduzir a cobertura.

Discussão

Os resultados obtidos permitem observar diferenças claras entre a GL reformulada e a versão original proposta por Lapa, Santos e Pirovani (2024). No caso da vírgula, as modificações realizadas, em especial nas GLs voltadas para repetições (v3) e conjunções adversativas deslocadas (v8), eliminaram falsos positivos e garantiram que todas as sinalizações realizadas correspondessem a casos efetivamente corretos. A precisão de 100% indica robustez no reconhecimento. Além disso, houve um ganho de 26,3% em abrangência (de 18,4% para 44,7%) e de 34,4% em medida-F (de 27,5% para 61,9%). Apesar disso, a abrangência de 44,7% ainda limita a cobertura dos fenômenos. Os resultados confirmam a melhoria em relação à GL original, reforçam que a estratégia adotada priorizou confiabilidade, o que é essencial em contextos práticos, mas apontam para a necessidade de expandir as regras a fim de capturar um maior número de casos.

Na crase, a diferença entre as duas abordagens foi ainda mais evidente. As regras originais, excessivamente amplas, resultavam em grande número de falsos positivos, enquanto a reformulação baseada em listas de verbos e adjetivos regentes reduziu esse problema de maneira significativa. A GL reformulada alcançou precisão de 87,5%, em contraste com os 66,7% da versão original, embora a abrangência tenha diminuído de 40,0% para 17,5%. Essa redução da abrangência era esperada, já que o estreitamento das regras, ao eliminar sinalizações indevidas, inevitavelmente restringe a cobertura. Assim como nas vírgulas, o desafio futuro está em aumentar a abrangência sem perder a precisão conquistada.

De forma geral, a principal contribuição das reformulações está na elevação da precisão, o que torna a ferramenta mais confiável em cenários reais, como textos jornalísticos. Por outro lado, a cobertura ainda é restrita, revelando a necessidade de ajustes futuros, como a expansão das listas de verbos e adjetivos regentes e a inclusão de novos contextos sintáticos. Essa tensão entre precisão e abrangência, amplamente discutida na literatura de PLN, reaparece aqui como um desafio central para a evolução das Gramáticas Locais aplicadas à norma culta do português.

Conclusão

Este trabalho revisou e aprimorou as Gramáticas Locais para correção de erros de crase e vírgula na língua portuguesa, desenvolvidas no trabalho de Lapa, Santos e Pirovani (2024). As reformulações priorizaram a precisão, o que reduziu significativamente os falsos positivos em comparação à proposta inicial. Nos testes realizados, as GLs reformuladas apresentaram maior confiabilidade, chegando a 100% de precisão nos casos de vírgula e 87,5% nos casos de crase, ainda que com menor abrangência, sobretudo nos casos de crase, cuja taxa foi reduzida de 40% para 17,5%. Em textos jornalísticos, mostraram-se mais adequadas por evitar sinalizações indevidas, reforçando a importância de regras mais restritivas. Apesar dos avanços, permanece como desafio ampliar a cobertura sem comprometer a precisão, apontando para a necessidade de futuras melhorias nas GLs.

Como perspectivas futuras, destaca-se a necessidade de aprimorar ainda mais as GLs, sobretudo na identificação de erros de crase, em que a análise de regência exerce forte influência sobre os resultados. Nesse sentido, mostra-se relevante ampliar as listas de termos regentes e considerar o uso de ferramentas auxiliares de análise sintática, de modo a aumentar a cobertura sem comprometer a precisão.

Referências

BECHARA, Evanildo. **Moderna gramática portuguesa**. 37. ed. rev., ampl. e atual. conforme o novo Acordo Ortográfico. Rio de Janeiro: Nova Fronteira, 2009.

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.

LAPA, C.P.; SANTOS, J.V.M.F.; PIROVANI, J.P.C. **Gramáticas Locais para correção de vírgula e crase em português**. 2024. Trabalho de Iniciação Científica (INIC 2024) – Universidade Federal do Espírito Santo, Alegre, 2024. Disponível em: https://www.inicepg.univap.br/cd/INIC_2024/anais/arquivos/RE_0981_0738_01.pdf. Acesso em: 22 ago. 2025.

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: 22 ago. 2025.

PENTEADO, Maria Carolina; PEREZ, Fábio. **Evaluating GPT-3.5 and GPT-4 on Grammatical Error Correction for Brazilian Portuguese**. arXiv preprint, arXiv:2306.15788, 2023. Disponível em: <https://arxiv.org/abs/2306.15788>. Acesso em: 22 ago. 2025.

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: 22 ago. 2025.

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, Bneveelo Horizonte. Anais... Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 347-351.

Agradecimentos

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

Além disso, expresso minha gratidão à orientadora Juliana pela sua atenção, dedicação, paciência e por ter me proporcionado essa oportunidade.