Revista brasileira de iniciação científica
ISSN: 2359-232X

Gramáticas Locais para o Reconhecimento de Entidades Nomeadas em Bulas de Medicamentos e Relatos de Casos Clínicos

Local Grammars for Named Entity Recognition in Package Inserts and Clinical Case Reports

Gramáticas Locales para el Reconocimiento de Entidades Nombradas en Prospectos de Medicamentos y Reportes de Casos Clínicos

Thiago Tonelli da Silva[1]
Juliana Campos Pirovani[2]

Resumo: Este trabalho teve como objetivo o Reconhecimento de Entidades Nomeadas (REN) em bulas de medicamentos e relatos de casos clínicos, com a finalidade de acelerar a compreensão de informações fundamentais nestes textos, elucidar dúvidas e possibilitar o uso dessas entidades em outras aplicações de Processamento de Linguagem Natural (PLN). Para isso, a pesquisa adotou uma abordagem linguística a fim de analisar sua efetividade na execução da tarefa proposta. Foram criadas 16 Gramáticas Locais (GLs) que foram aplicadas em dois *corpus* por meio de *shell scripts*. Os resultados iniciais mostraram-se promissores, especialmente no *corpus* em que foi realizado o estudo linguístico. Por exemplo, a categoria denominada COISA alcançou precisão de 80%. Contudo, será necessário aprimorar as GLs e incluir novas regras para aumentar a abrangência, bem como realizar um estudo mais detalhado da categoria ABSTRAÇÃO.
Palavras-chave: Gramáticas Locais. PLN. Entidades Nomeadas.

Abstract: This study aimed to perform Named Entity Recognition (NER) in drug package inserts and clinical case reports in order to accelerate the understanding of essential information in these texts, clarify doubts, and enable the use of such entities in other Natural Language Processing (NLP) applications. To this end, the research adopted a linguistic approach to evaluate its effectiveness in carrying out the proposed task. Sixteen Local Grammars (LGs) were developed and applied to two corpora using shell scripts. Initial results proved promising, particularly in the corpus in which the linguistic analysis was conducted. For example, the category denominated COISA achieved a precision rate of 80%. However, further refinement of the LGs and the inclusion of new rules will be necessary to expand coverage, as well as a more detailed study of the ABSTRAÇÃO category.
Keywords: Local Grammars. NLP. Named Entities

Resumen: Este trabajo tuvo como objetivo el Reconocimiento de Entidades Nombradas (REN) en prospectos de medicamentos y reportes de casos clínicos con el fin de agilizar la comprensión de información fundamental en estos textos, aclarar dudas y posibilitar el uso de dichas entidades en otras aplicaciones de Procesamiento de Lenguaje Natural (PLN). Para ello, la investigación adoptó un enfoque lingüístico con el propósito de evaluar su efectividad en la ejecución de la tarea propuesta. Se desarrollaron 16 Gramáticas Locales (GLs), que fueron aplicadas a dos corpus mediante shell scripts. Los resultados iniciales resultaron prometedores, especialmente en el corpus en el que se llevó a cabo el análisis lingüístico. Por ejemplo, la categoría denominada COISA alcanzó una precisión del 80%. No obstante, será necesario perfeccionar las GLs e incorporar nuevas reglas para ampliar la cobertura, así como realizar un estudio más detallado de la categoría ABSTRAÇÃO.
Palabras clave: Gramáticas Locales. PLN. Entidades Nombradas.

Submetido 22/09/2025 **Aceito 31/03/2026** **Publicado 28/05/2026**

[1] Graduando em Ciência da Computação. Universidade Federal do Espírito Santo – UFES, Campus Alegre. ORCID: https://orcid.org/0009-0000-3352-8723. E-mail: thiago.silva.66@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

Considerações iniciais

O Processamento de Linguagem Natural (PLN), uma área interdisciplinar entre a Ciência da Computação e a Linguística, dedica-se ao estudo da geração e compreensão da linguagem humana. Nesse contexto, destaca-se o Reconhecimento de Entidades Nomeadas (REN), tarefa responsável pela identificação e classificação automática de entidades em textos de escrita livre, como nomes de pessoas, locais e organizações (Pirovani, 2019). A identificação e a categorização dessas entidades são fundamentais para aprimorar a compreensão do texto, favorecendo a interpretação e a análise textual.

No contexto da saúde, o uso do REN torna-se particularmente relevante. A identificação automática de princípios ativos, sintomas e doenças pode facilitar a compreensão das bulas de medicamentos por pacientes e profissionais de saúde. Com a identificação automática dessas entidades é possível desenvolver diversos mecanismos de apoio ao médico e aos pacientes, como por exemplo, uma ferramenta que apresenta a bula com entidades destacadas, facilitando a legibilidade e identificação das suas palavras-chave; um chatbot assistente para responder perguntas básicas sobre a bula ou relato clínico. Por exemplo, um paciente poderia perguntar quais são os efeitos de determinado medicamento e o sistema identificaria automaticamente, no texto, as entidades previamente categorizadas como efeitos adversos ou reações colaterais. A partir dessa identificação, o sistema organizaria essas informações e apresentaria uma resposta objetiva.

No Brasil, a Agência Nacional de Vigilância Sanitária (ANVISA) é o órgão responsável por regulamentar a elaboração de bulas medicamentosas. Atualmente, três Resoluções da Diretoria Colegiada (RDCs) são fundamentais nesse processo. A RDC nº 47, de 6 de setembro de 2009, estabelece as regras para a elaboração, harmonização, atualização, publicação e disponibilização de bulas, tanto para pacientes, quanto para profissionais de saúde (Brasil, 2009a). Já a RDC nº 60, de 26 de novembro de 2012, dispõe sobre os procedimentos internos da ANVISA para as alterações nos textos das bulas de medicamentos (Brasil, 2012). Por fim, a RDC nº 770, de 20 de dezembro de 2022, detalha as frases de alerta obrigatórias para bulas e rótulos de medicamentos (Brasil, 2022).

Apesar dessas regulamentações, as bulas medicamentosas, responsáveis por fornecer informações detalhadas sobre composição, administração, efeitos colaterais e interações

medicamentosas, nem sempre apresentam instruções claras. Essa limitação pode comprometer a compreensão por parte dos pacientes (Silva *et al*., 2000). Nesse cenário, o REN mostra-se uma abordagem promissora para garantir maior clareza e acessibilidade da informação ao público-alvo.

Assim como as bulas, os relatos de casos clínicos também desempenham um papel relevante na área da saúde. Esses documentos permitem que os profissionais de saúde documentem de maneira clara e prática o desenvolvimento e o tratamento de uma condição médica. Esses relatos podem ser apresentados de diferentes maneiras, pois cada profissional insere as informações que considera válidas e descreve o tratamento de acordo com sua experiência, abordagem pessoal e a complexidade do caso. A extração automática dessas informações é importante, pois os relatos de casos clínicos são fontes essenciais de conhecimento no campo biomédico, ajudando a entender apresentações de doenças raras e os benefícios de tratamentos não convencionais (Pineda-Leguízamo *et al.*, 2018).

As abordagens empregadas na construção de sistemas de REN abrangem a linguística, o aprendizado de máquina e as abordagens híbridas (Pirovani, 2019). Um exemplo de abordagem linguística é o uso de Gramáticas Locais (GLs), propostas por Gross (1997), que utilizam regras escritas à mão para identificar as entidades de interesse.

No estudo de Pirovani (2019), foram desenvolvidos *scripts* e GLs para as dez categorias do HAREM[3], uma análise coletiva para o REN, na língua portuguesa. Nessa avaliação, foram determinadas as seguintes categorias de Entidades Nomeadas (ENs): ABSTRACCAO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, OUTRO, PESSOA, TEMPO e VALOR. Os *corpora* anotados utilizados no Primeiro e no Segundo HAREM, denominados como *Gold Collections* (GCs), tornaram-se a referência padrão-ouro para avaliação de sistemas de REN, em português (Pirovani, 2021).

Colombo e Oliveira (2022) investigaram a extração de informações em bulas de medicamentos e relatos de casos clínicos por meio da identificação de ENs. Os autores utilizaram uma abordagem híbrida (CRF+LG), combinando o modelo estatístico *Conditional Random Fields* (CRF) com GLs (Pirovani, 2019). A partir da análise, observou-se que algumas

[3] https://www.linguateca.pt/harem/

entidades do âmbito da saúde, como “omeprazol” (substância) e “taquicardia” (sintoma), foram identificadas e classificadas nas categorias COISA e ABSTRAÇÃO (ABSTRACCAO) do HAREM, respectivamente, mesmo sem a utilização de uma gramática específica para isso. Os autores observaram que uma Gramática Local (GL) construída especificamente para esse propósito poderia melhorar os resultados dessa e de outras abordagens de aprendizado de máquina.

Dessa forma, este trabalho teve como objetivo desenvolver GLs especificamente para reconhecer entidades na área da saúde, pertencentes às categorias como sintomas, doenças, princípios ativos e excipientes. Este trabalho busca responder à seguinte pergunta de pesquisa: é possível reconhecer automaticamente entidades nomeadas em textos da área da saúde utilizando apenas Gramáticas Locais (GLs)?

Metodologia

Esta pesquisa é de natureza aplicada, com abordagem mista, combinando procedimentos qualitativos e quantitativos. Quanto aos objetivos, classifica-se como descritiva e utiliza procedimento experimental. Utilizou-se uma abordagem linguística, na qual regras são elaboradas manualmente para reconhecer o contexto em que uma Entidade Nomeada (EN) ocorre. Apesar de simples, a abordagem linguística é capaz de reconhecer entidades que podem não ser notadas por outras estratégias (Zhou; Su, 2002). Ela pode ser utilizada quando não existe *corpus* para treinar modelos de aprendizado de máquina ou o processamento para realizar o treinamento é muito alto.

Empregar GLs para o REN se justifica tanto por razões práticas, quanto pelo contexto textual em questão. Em textos altamente padronizados observa-se uma regularidade lexical e sintática que favorece a construção de regras gramaticais específicas para a extração de entidades. Essas estruturas linguísticas previsíveis podem ser eficientemente exploradas por GLs, oferecendo resultados sem a necessidade de grandes volumes de dados de treinamento.

Um dos grandes desafios na abordagem linguística é quando o contexto não é claro o suficiente ou quando o padrão de identificação não é preciso para reconhecer corretamente uma EN. Por isso, um estudo linguístico aprofundado torna-se essencial para analisar o texto e

Revista brasileira de iniciação científica
ISSN: 2359-232X

examinar a construção das sentenças, buscando identificar padrões recorrentes na escrita que possam indicar a presença de uma EN, tanto antes quanto depois de uma palavra.

A pesquisa foi desenvolvida em etapas que contemplaram: a escolha dos *corpus*; o estudo linguístico, destinado à identificação de padrões lexicais e sintáticos relevantes; a elaboração manual de regras linguísticas para o REN; e, por fim, a validação experimental, com a aplicação das regras ao *corpus* e a avaliação da precisão e abrangência dos resultados obtidos.

Inicialmente, o *corpus* escolhido para o estudo e para a criação das GLs foi uma fração do *corpus* elaborado e anotado manualmente por Colombo e Oliveira (2022), com o uso da ferramenta Etiquet(H)arem. Essa amostra é composta por 20 documentos, sendo 10 relatos médicos da SciELO e 10 bulas de medicamentos (Amoxicilina, Esogastro, Cloridrato de Ranitidina, Gastrium, Label, Iniparet, Laflugi, Pyloripac, Omepramix e Ziprol). Para ilustrar o formato textual das bulas utilizadas, a Figura 1 apresenta um exemplo e a Figura 2 apresenta o mesmo trecho anotado por Colombo e Oliveira (2022).

Figura 1 – Trecho da bula sem anotação

```
2. COMO ESTE MEDICAMENTO FUNCIONA?
Este medicamento contém uma penicilina chamada amoxicilina como ingrediente ativo. A
amoxicilina pertence ao grupo dos antibióticos penicilânicos. A amoxicilina é usada
no tratamento de uma gama de infecções causadas por bactérias, que podem manifestar-
se nos pulmões (pneumonia e bronquite), nas amígdalas (amigdalite), nos seios da face
(sinusite), no trato urinário e genital, na pele e nas mucosas. A amoxicilina atua
destruindo as bactérias que causam essas infecções.

3. QUANDO NÃO DEVO USAR ESTE MEDICAMENTO?
Este medicamento não pode ser usado por pessoas alérgicas à amoxicilina, a outros
antibióticos penicilínicos ou antibióticos similares, chamados cefalosporinas. Se
você já teve uma reação alérgica (como erupções da pele) ao tomar um antibiótico,
deve conversar com seu médico antes de usar amoxicilina.
```

Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 2 – Bula da amoxicilina anotada

```
2. COMO ESTE <EM ID="amoxicilina-31" CATEG="COISA">MEDICAMENTO</EM> FUNCIONA?
Este <EM ID="amoxicilina-32" CATEG="COISA">medicamento</EM> contém uma <EM ID="amoxicilina-33"
CATEG="COISA">penicilina</EM> chamada <EM ID="amoxicilina-34" CATEG="COISA">amoxicilina</EM>
como <EM ID="amoxicilina-161" CATEG="COISA" TIPO="SUBSTANCIA">ingrediente ativo</EM>. A <EM
ID="amoxicilina-35" CATEG="COISA">amoxicilina</EM> pertence ao grupo dos <EM
ID="amoxicilina-36" CATEG="COISA">antibióticos penicilânicos</EM>. A <EM ID="amoxicilina-37"
CATEG="COISA">amoxicilina</EM> é usada no tratamento de uma gama de <EM ID="amoxicilina-38"
CATEG="ABSTRACCAO">infecções causadas por bactérias</EM>, que podem manifestar-se nos pulmões
(<EM ID="amoxicilina-39" CATEG="ABSTRACCAO">pneumonia</EM> e <EM ID="amoxicilina-40"
CATEG="ABSTRACCAO">bronquite</EM>), nas amígdalas (<EM ID="amoxicilina-41"
CATEG="ABSTRACCAO">amigdalite</EM>), nos seios da face (<EM ID="amoxicilina-42"
CATEG="ABSTRACCAO">sinusite</EM>), no trato urinário e genital, na pele e nas mucosas. A <EM
ID="amoxicilina-43" CATEG="COISA">amoxicilina</EM> atua destruindo as bactérias que causam
essas <EM ID="amoxicilina-44" CATEG="ABSTRACCAO">infecções</EM>.

3. QUANDO NÃO DEVO USAR ESTE <EM ID="amoxicilina-45" CATEG="COISA">MEDICAMENTO</EM>?
Este <EM ID="amoxicilina-46" CATEG="COISA">medicamento</EM> não pode ser usado por <EM
ID="amoxicilina-47" CATEG="PESSOA">pessoas alérgicas</EM> à <EM ID="amoxicilina-48"
CATEG="COISA">amoxicilina</EM>, a outros <EM ID="amoxicilina-49" CATEG="COISA">antibióticos
penicilínicos</EM> ou <EM ID="amoxicilina-50" CATEG="COISA">antibióticos similares</EM>,
chamados <EM ID="amoxicilina-51" CATEG="COISA">cefalosporinas</EM>. Se <EM ID="amoxicilina-52"
CATEG="PESSOA">você</EM> já teve uma <EM ID="amoxicilina-53" CATEG="ABSTRACCAO">reação
alérgica</EM> (como <EM ID="amoxicilina-54" CATEG="ABSTRACCAO">erupções da pele</EM>) ao tomar
um <EM ID="amoxicilina-55" CATEG="COISA">antibiótico</EM>, deve conversar com seu <EM
ID="amoxicilina-56" CATEG="PESSOA">médico</EM> antes de usar <EM ID="amoxicilina-57"
CATEG="COISA">amoxicilina</EM>.
```

Fonte: o autor (2025)

É fundamental perceber que essas anotações cobrem um escopo mais amplo do que o domínio considerado para o presente estudo. O *corpus* anotado inclui todas as categorias do HAREM e não apenas as entidades de interesse para a nossa análise (como ABSTRAÇÃO e COISA); um exemplo é a categoria PESSOA, que designa um indivíduo.

As anotações de ENs no *corpus* seguem uma estrutura padronizada (observe a Figura 2), onde o termo identificado é encapsulado por *tags* <EM> e </EM>. A *tag* de abertura contém atributos essenciais para a identificação e categorização da entidade. O atributo “ID” fornece um identificador único, composto pelo nome do documento de origem (ex: amoxicilina) e um número sequencial que indica a ordem daquela anotação específica dentro do documento (ex: 31). Já o atributo “CATEG” designa a categoria semântica da entidade, classificando o tipo de informação que ela representa (ex: COISA para "MEDICAMENTO").

Selecionou-se um *corpus* diferente do utilizado no estudo linguístico para testar as GLs. Para isso, foi escolhido o SemClinBr[4], o primeiro *corpus* clínico semanticamente anotado em

[4] https://github.com/HAILab-PUCPR/SemClinBr

Português (do Brasil), formado por 1000 relatos clínicos de diferentes situações e especialidades médicas (Oliveira *et al.,* 2022). Segundo Oliveira *et al*. (2022), os dados foram obtidos de um grupo de hospitais do Brasil (gerados entre 2013 e 2018) e um hospital universitário, com base nas entradas no período entre 2002 e 2007. Os autores observaram diversas características nos relatos, como o alto uso de siglas e jargão médico, erros ortográficos, problemas de pontuação e letras minúsculas e maiúsculas incorretas.

É fundamental destacar que os relatos do SemClinBr apresentam características linguísticas distintas dos relatos da SciELO utilizados para construção das GLs. Por exemplo, o alto uso de siglas, jargão médico e erros ortográficos, como citado anteriormente. Além disso, seu padrão de anotação diverge do HAREM, o que resultou em desafios significativos para adaptá-lo. A Figura 3 apresenta um relato do *corpus* SemClinBr.

Figura 3 – Relato do *corpus* SemClinBr

```
<?xml version='1.0' encoding='UTF-8'?>
<ANNOTATIONS>
<TEXT>03:03 Retornou do CC por volta das 21:00 horas ,  POI de tto cx por fx de punho esquerdo . Apresenta curativo + tala
gessada em MSE , referindo algia moderada , edema distal , mobilidade diminuida , apresentou 1 episodio de emese , sendo
medicada , diurese presente , segue cuidados . FRATURA DA EXTREMIDADE DISTAL DO RADIO</TEXT>
<TAGS>
<annotation id="18446" tag="Health Care Related Organization|Abbreviation" start="18" end="20" text="CC" abbr="Centro
Cirúrgico" />
<annotation id="18447" tag="Abbreviation|Temporal Concept" start="50" end="53" text="POI" abbr="" />
<annotation id="18448" tag="Medical Device|Therapeutic or Preventive Procedure" start="101" end="109" text="curativo"
abbr="" />
<annotation id="18449" tag="Abbreviation|Body Location or Region" start="128" end="131" text="MSE" abbr="" />
<annotation id="18450" tag="Quantitative Concept" start="210" end="211" text="1" abbr="" />
<annotation id="18451" tag="Therapeutic or Preventive Procedure|Health Care Activity" start="238" end="246" text="medicada"
abbr="" />
<annotation id="18452" tag="Health Care Activity" start="274" end="282" text="cuidados" abbr="" />
<annotation id="18453" tag="Abbreviation|Therapeutic or Preventive Procedure" start="57" end="63" text="tto cx" />
<annotation id="18454" tag="Abbreviation|Injury or Poisoning" start="68" end="88" text="fx de punho esquerdo" />
<annotation id="18455" tag="Medical Device|Therapeutic or Preventive Procedure" start="112" end="124" text="tala gessada" />
<annotation id="18456" tag="Sign or Symptom" start="144" end="158" text="algia moderada" />
<annotation id="18457" tag="Pathologic Function|Sign or Symptom" start="161" end="173" text="edema distal" />
<annotation id="18458" tag="Sign or Symptom" start="176" end="196" text="mobilidade diminuida" />
<annotation id="18459" tag="Sign or Symptom" start="212" end="229" text="episodio de emese" />
<annotation id="18460" tag="Finding|Physiologic Function" start="249" end="265" text="diurese presente" />
<annotation id="18461" tag="Injury or Poisoning" start="285" end="323" text="FRATURA DA EXTREMIDADE DISTAL DO RADIO" />
</TAGS>
<RELATIONS>
<rel annotation1="18447" annotation2="18453" reltype="associated_with" />
<rel annotation1="18453" annotation2="18454" reltype="associated_with" />
</RELATIONS>
</ANNOTATIONS>
```

Fonte: o autor (2025)

Nos relatos do *corpus* SemClinBr (Figura 3), é possível ver as características já citadas anteriormente, o que dificulta a compreensão e o reconhecimento automático das entidades. Já os relatos provenientes da SciELO (Figura 4) apresentam maior clareza textual e padrões mais definidos.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 4 – Relato da SciELO

```
<DOC DOCID="relato_002">
<EM ID="relato_002-80" CATEG="PESSOA">Paciente do gênero feminino</EM>, <EM ID="relato_002-81" CATEG="VALOR">79 anos</EM>, <EM
ID="relato_002-82" CATEG="VALOR">60kg</EM>, <EM ID="relato_002-83" CATEG="VALOR">1.60m</EM>, portadora de <EM ID="relato_002-84"
CATEG="ABSTRACCAO">depressão</EM> e <EM ID="relato_002-85" CATEG="ABSTRACCAO">labirintite</EM>, em uso regular de <EM ID="relato_002-86"
CATEG="COISA">lorazepam</EM>, <EM ID="relato_002-87" CATEG="COISA">labirin</EM>, <EM ID="relato_002-88" CATEG="COISA">ranitidina</EM>,
<EM ID="relato_002-89" CATEG="COISA">citalopram</EM> e <EM ID="relato_002-90" CATEG="COISA">risperidona</EM>. <EM ID="relato_002-91"
CATEG="TEMPO">Após duas semanas</EM> da realização de artroplastia total de quadril, apresentou queda, com luxação da prótese, sendo
submetida a cirurgia de emergência.

Usava as medicações habituais, acrescidas de <EM ID="relato_002-92" CATEG="COISA">xarelto</EM>, o que motivou a opção <EM
ID="relato_002-45" CATEG="ABSTRACCAO" TIPO="ESTADO">pela anestesia geral</EM>. Administrados <EM ID="relato_002-93"
CATEG="COISA">cefazolina</EM> <EM ID="relato_002-94" CATEG="VALOR">2g</EM> IV no <EM ID="relato_002-97" CATEG="COISA">SF</EM> <EM
ID="relato_002-95" CATEG="VALOR">0,9%</EM> <EM ID="relato_002-96" CATEG="VALOR">500 ml</EM> e <EM ID="relato_002-98"
CATEG="COISA">palonosetrona</EM> <EM ID="relato_002-99" CATEG="VALOR">0,75 mg</EM> IV diluído em <EM ID="relato_002-100" CATEG="VALOR">20
mL</EM> de <EM ID="relato_002-101" CATEG="COISA">SF</EM> <EM ID="relato_002-102" CATEG="VALOR">0.9%</EM>.

Após a indução, realizada com <EM ID="relato_002-103" CATEG="COISA">fentanil</EM> <EM ID="relato_002-104" CATEG="VALOR">200 micro g</EM>,
<EM ID="relato_002-105" CATEG="COISA">lidocaína</EM> <EM ID="relato_002-106" CATEG="VALOR">2%</EM> <EM ID="relato_002-107"
CATEG="VALOR">60 mg</EM>, <EM ID="relato_002-108" CATEG="COISA">propofol</EM> <EM ID="relato_002-109" CATEG="VALOR">100 mg</EM> e <EM
ID="relato_002-110" CATEG="COISA">cisatracúrio</EM> <EM ID="relato_002-111" CATEG="VALOR">60 mg</EM>, <EM ID="relato_002-112"
CATEG="PESSOA">a paciente</EM> evoluiu com <EM ID="relato_002-113" CATEG="ABSTRACCAO">bradicardia sinusal</EM> e <EM ID="relato_002-114"
CATEG="ABSTRACCAO">hipotensão</EM>. Não houve resposta a <EM ID="relato_002-115" CATEG="VALOR">0,75 mg</EM> de <EM ID="relato_002-116"
CATEG="COISA">atropina</EM> e <EM ID="relato_002-117" CATEG="COISA">efedrina</EM> <EM ID="relato_002-118" CATEG="VALOR">30 mg</EM> ;
ocorreu rápida progressão para <EM ID="relato_002-119" CATEG="ABSTRACCAO">parada cardiorrespiratória</EM> em <EM ID="relato_002-120"
CATEG="ABSTRACCAO">assistolia</EM>, revertida em <EM ID="relato_002-121" CATEG="TEMPO">3 minutos</EM> de manobras de <EM
ID="relato_002-46" CATEG="ABSTRACCAO" TIPO="ESTADO">ressuscitação cardiorrespiratória</EM>. Foi extubada após descurarização, lúcida e
hemodinamicamente estável.
</DOC>
```

Fonte: o autor (2025)

Para categorização das entidades, este trabalho manteve os padrões de categorias estabelecidos pelo HAREM, que reconheceram entidades no contexto considerado (COISA e ABSTRAÇÃO), acrescentando como tipo as entidades de interesse: doença, sintoma, princípio e excipiente. A manutenção dessas categorias tem como objetivo facilitar a reutilização dos *scripts* desenvolvidos por Pirovani (2019) e, posteriormente, viabilizar a aplicação das GLs construídas na abordagem híbrida CRF+LG. A Tabela 1 apresenta as categorias e seus respectivos tipos, acompanhados de uma breve descrição e exemplos.

Tabela 1 – Descrição do padrão utilizado

CATEGORIA	TIPO	DESCRIÇÃO	EXEMPLO
ABSTRAÇÃO	DOENÇA	Condição patológica ou distúrbio	Alzheimer
ABSTRAÇÃO	SINTOMA	Manifestação da doença	Tontura
COISA	EXCIPIENTE	Substâncias não ativas	Glicerina
COISA	PRINCÍPIO	Princípios ativos	Amoxicilina

Fonte: o autor (2025)

A adoção desses padrões também possibilita, em trabalhos futuros, a realização de análises comparativas com os resultados apresentados por Colombo e Oliveira (2022). No

entanto, essa comparação ainda não é possível, uma vez que o *corpus* utilizado encontra-se divergente, o que inviabiliza uma análise consistente.

No que se refere ao processo de anotação, as entidades reconhecidas foram anotadas utilizando *tags*, com o padrão <CATEGORIA_TIPO>EN</CATEGORIA_TIPO>. Um exemplo de anotação é <COISA_PRINCIPIO>amoxicilina</COISA_PRINCIPIO>.

Em virtude das distinções entre as categorias presentes nos dois *corpus*, no processo de adaptar o SemClinBr para o padrão HAREM, foi fundamental definir o que se enquadraria como COISA e ABSTRAÇÃO. Para isso, avaliaram-se todas as categorias presentes no *corpus* SemClinBr, e as que se mostraram mais compatíveis com o escopo do presente trabalho foram selecionadas.

Para a categoria ABSTRAÇÃO do HAREM, as seguintes categorias do SemClinBr foram selecionadas: *Sign or Symptom* (manifestações clínicas), *Disease or Syndrome* (condições médicas específicas), *Finding* (observações clínicas), *Cell or Molecular Dysfunction* (alterações celulares ou moleculares), *Injury or Poisoning* (danos físicos ou tóxicos), *Pathologic Function* (funcionamento anormal do organismo) e *Laboratory or Test Result* (resultados que indicam um estado ou valor).

Para a categoria COISA do HAREM, as seguintes categorias do SemClinBr foram selecionadas: *Antibiotic* (substâncias medicamentosas que combatem infecções), *Pharmacologic Substance|Hormone|Amino Acid, Peptide, or Protein* (substâncias químicas, como medicamentos, hormônios e componentes biológicos) e *Body Substance* (fluidos ou materiais biológicos do corpo).

Com objetivo de avaliar o desempenho das GLs, foram empregados dois cenários distintos. Primeiramente, foram realizados testes no próprio *corpus* onde o estudo linguístico foi conduzido, permitindo comparar as anotações geradas automaticamente com as referências anotadas manualmente. Essa avaliação utilizou os *scripts* do Segundo HAREM, que calculam métricas amplamente reconhecidas no REN: precisão (proporção de acertos em relação ao total de ENs identificadas), abrangência (proporção de acertos em relação ao total de ENs existentes) e medida-F (média harmônica entre precisão e abrangência) (Mota; Santos, 2008).

Por fim, para aferir o desempenho do trabalho em um *corpus* alternativo, aplicou-se o mesmo processo ao SemClinBr. No entanto, foi necessário desenvolver *scripts* específicos para

Revista brasileira de iniciação científica
ISSN: 2359-232X

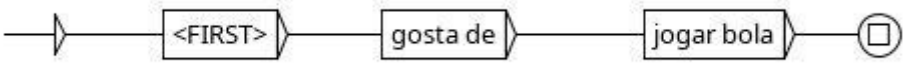
converter as anotações do SemClinBr para o padrão HAREM, possibilitando assim o uso dos *scripts* de avaliação do Segundo HAREM e o cálculo das mesmas métricas de desempenho.

Unitex

A partir da análise de regras, foram construídas GLs utilizando o Unitex[5], um conjunto de *softwares* livres para o PLN, que fornece diversos recursos para sua criação, utilizando um formato de grafo. Sua representação no Unitex possui uma estrutura bem definida com entradas, transições e saída.

A Figura 5 exibe uma GL que reconhece a sentença "João gosta de jogar bola" ou, alternativamente, outro nome próprio no lugar de "João". Um aspecto fundamental para o funcionamento das GLs no Unitex é o seu sistema de *tags* lexicais, que são etiquetas representativas de características ou categorias gramaticais das palavras, como verbos, substantivos ou adjetivos. A *tag* lexical <FIRST> representa qualquer palavra cuja primeira letra esteja em maiúsculo, o que possibilita o reconhecimento de uma ampla variedade de construções que seguem esse padrão, otimizando a generalização das regras.

Figura 5 – Exemplo de GL



Fonte: o autor (2025)

Para garantir maior precisão no reconhecimento de entidades da categoria COISA, especialmente substâncias, adotou-se uma estratégia específica que envolveu o uso da *tag* lexical <!DIC>. Essa *tag* foi crucial para o aprimoramento do processo devido à sua assertividade. É importante ressaltar que o *software* Unitex dispõe de uma vasta gama de dicionários pré-existentes para diferentes idiomas e a *tag* lexical <DIC> é empregada para reconhecer palavras presentes nesses léxicos. No entanto, uma limitação observada é que muitos termos, especialmente os específicos ou técnicos, de determinados domínios, não estão

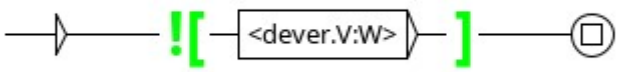
[5] https://unitexgramlab.org/pt

incluídos nesses dicionários padrão. Consequentemente, uma parcela significativa das substâncias presentes em bulas de medicamentos, por exemplo, não pertence ao dicionário convencional do Unitex. Dessa forma, para representar esse conjunto de termos que não estão nos dicionários padrão, foi utilizada a *tag* lexical <!DIC>, permitindo a identificação precisa dessas ocorrências.

Outro conceito importante do Unitex é o contexto. Basicamente, contexto é tudo aquilo que aparece ao redor de uma palavra ou expressão e que ajuda a determinar seu significado ou sua função no texto. No Unitex, o contexto serve para controlar quando uma palavra ou estrutura será reconhecida por uma GL. Isso significa que uma expressão só será aceita se estiver dentro de uma situação específica, previamente definida. Caso contrário, ela será ignorada.

Em uma GL, um nó pode ser configurado para aceitar diversas flexões de uma palavra. Por exemplo, se o objetivo é excluir a forma do verbo “dever”, no infinitivo, o Unitex oferece mecanismos para especificar esse contexto de exclusão. Isso significa que, caso a entrada seja “dever”, o sistema não seguirá essa transição, indicando uma rejeição. No Unitex a representação de um contexto é um nó envolto por colchetes, e para a negação do contexto adiciona-se um sinal de exclamação antes do colchete de abertura. A Figura 6 demonstra o exemplo discutido.

Figura 6 – Negação de contexto



Fonte: o autor (2025)

Essa capacidade de definir contextos é crucial para evitar ambiguidades e aumentar a precisão da análise sintática. Ela permite criar regras que não apenas identificam padrões lexicais, mas também permitem controlar propriedades morfológicas e gramaticais das palavras, como tempo verbal, número, gênero, ou a forma da flexão de um verbo. Ao refinar as

condições de aceitação ou rejeição em cada etapa do grafo, o Unitex possibilita a construção de gramáticas mais robustas e capazes de lidar com as complexidades e nuances da linguagem natural.

Por fim, é importante citar que o Unitex permite uma arquitetura modular através da incorporação de subGLs em nós de uma GL principal. Essa capacidade favorece o reuso de componentes gramaticais e a construção de estruturas complexas de forma organizada. Complementarmente, a ferramenta dispõe de recursos para anexar saídas a textos, função crucial para a anotação de entidades. Esse processo permite a demarcação automática de sentenças ou a marcação de entidades específicas diretamente no texto.

Análise dos dados e resultados

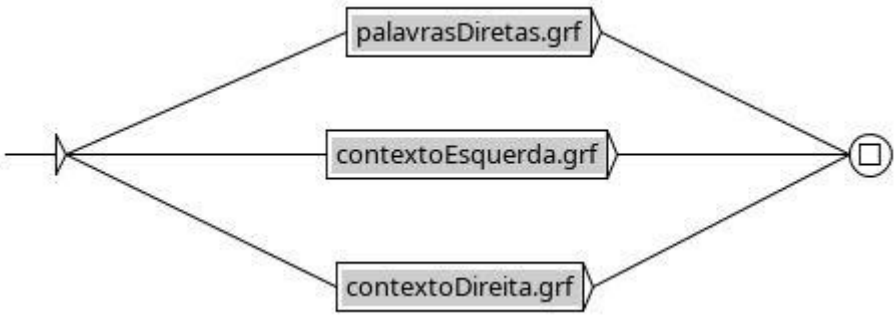
Foi necessário construir uma GL principal denominada “Main.grf” para acionar as duas subGLs: “Coisa.grf” e “Abstraccao.grf”, responsáveis por reconhecer, respectivamente, COISA e ABSTRAÇÃO. Dentro de cada uma delas, foram elaboradas outras três subGLs, para agrupar os contextos à esquerda, os contextos à direita e as palavras diretas.

GL para a categoria COISA

Inicialmente, serão exibidas as regras para identificar entidades da categoria COISA (Figura 7). Para justificar a estrutura considerada, é importante destacar que a nomeação dos arquivos tem o propósito de indicar o lado predominante dos contextos, como no caso de “contextoEsquerda.grf”. No entanto, isso não significa que nessa GL todos os padrões estarão exclusivamente ao lado esquerdo, pois algumas regras podem ocasionalmente aparecer no lado oposto. Isso ocorre porque, em certos casos, as entidades de interesse podem conter simultaneamente padrões comuns à esquerda e à direita, o que é benéfico, pois resulta em maior precisão.

Revista brasileira de iniciação científica
ISSN: 2359-232X

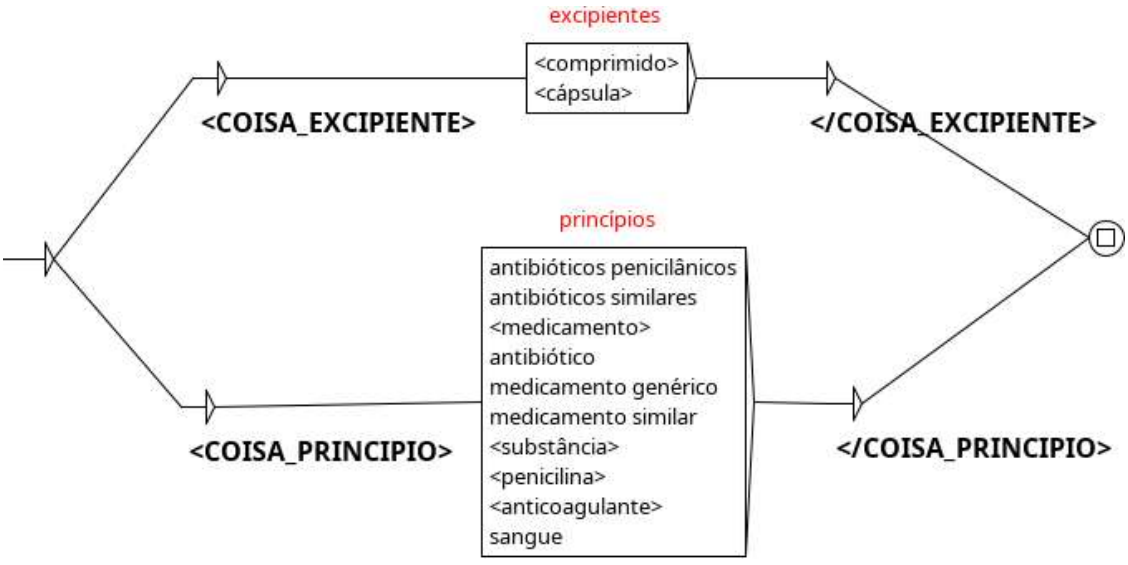
Figura 7 – GL “Coisa.grf”



Fonte: o autor (2025)

Pode-se afirmar que a anotação de uma palavra específica é um processo mais simples e preciso. Nesse sentido, é possível observar que, em bulas de medicamentos, a palavra “medicamento” ocorre com frequência. Ou seja, há entidades de interesse que aparecem abundantemente nos textos e não precisam de contexto para serem identificadas. Uma abordagem eficaz para lidar com essa questão é inserir diretamente as palavras desejadas como regra. Com isso, foi construída a GL “palavrasDiretas.grf” (Figura 8), onde essas palavras podem ser visualizadas.

Figura 8 – GL “palavrasDiretas.grf”



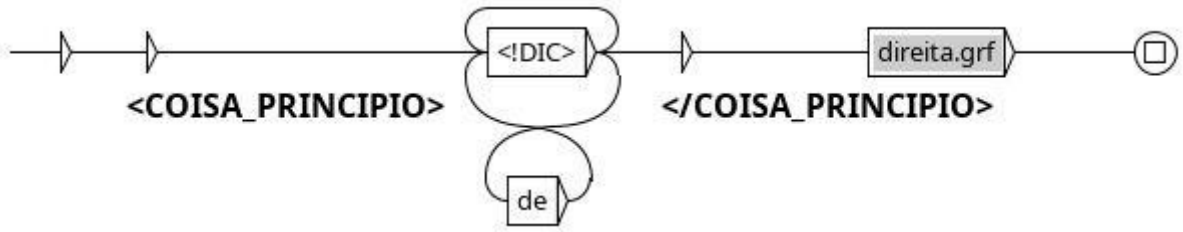
Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Em contrapartida, existem as GLs que utilizam contextos para o reconhecimento de entidades, uma vez que esse contexto pode ser suficiente para reconhecer entidades distintas. Por exemplo, a sentença “tratamento com” indica que a próxima palavra possivelmente será um medicamento, o que significa que qualquer termo pode surgir nesse contexto. Alguns exemplos são: tratamento com amoxicilina e tratamento com dipirona, entre outros.

A GL “contextoDireita.grf” (Figura 9) possui entre os rótulos de marcação a *tag* <!DIC>. O nó dessa *tag* contém um ciclo, o que permite múltiplas ocorrências e possibilita o reconhecimento de palavras compostas como “dipirona monoidratada”. Além disso, há um desvio condicional para o nó “de”, permitindo a identificação de expressões como “cloridrato de sertralina”.

Figura 9 – GL “contextoDireita.grf”



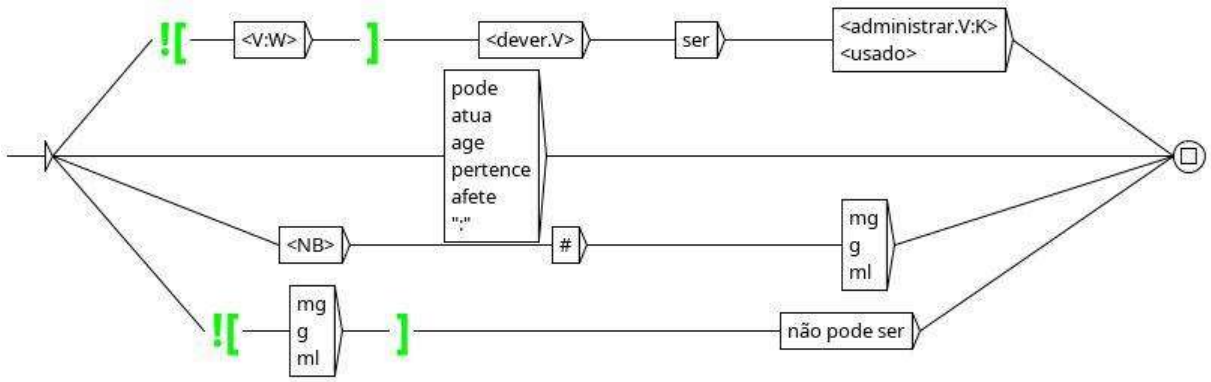
Fonte: o autor (2025)

Avançando para o próximo nó do grafo, encontra-se a subGL “direita.grf” (Figura 10), responsável por armazenar os padrões observados à direita das entidades de interesse. Observou-se que após a aparição de substâncias, frequentemente surgem verbos que indicam o funcionamento do remédio como “age”, “atua”, “pode”, “pertence” e “afete”.

Além disso, no estudo linguístico verificou-se que os verbos que sucedem uma entidade de interesse nunca aparecem no infinitivo. Para garantir essa regra, utilizou-se a negação de contexto. Neste caso, a negação de contexto atua rejeitando imediatamente qualquer verbo no infinitivo (<V:W>). Após essa verificação, outras regras são aplicadas para garantir a assertividade.

Revista brasileira de iniciação científica
ISSN: 2359-232X

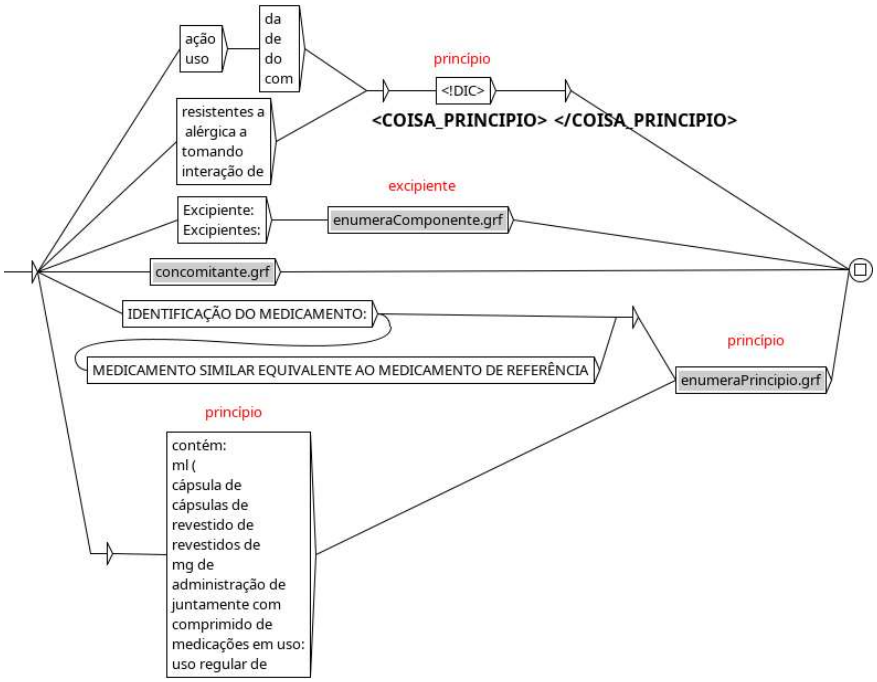
Figura 10 – subGL “direita.grf”



Fonte: o autor (2025)

A GL responsável por agrupar os contextos à esquerda demonstrou-se significativamente maior que as demais, resultado da observação de múltiplos padrões linguísticos durante a análise do *corpus*. A Figura 11 exibe a GL “contextoEsquerda.grf”.

Figura 11 – GL “contextoEsquerda.grf”



Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

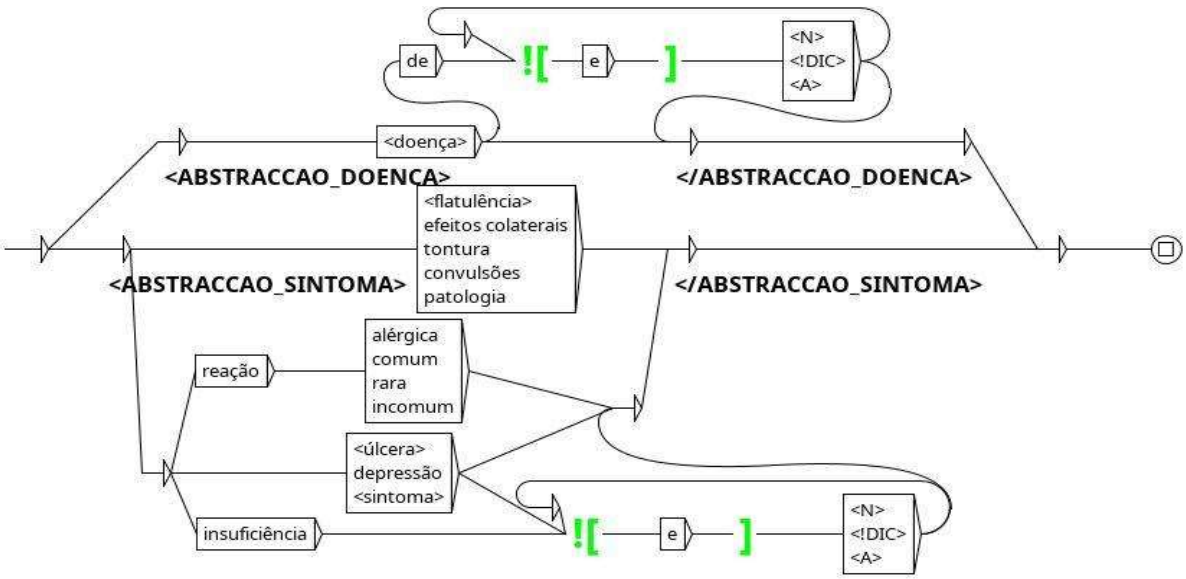
Devido ao tamanho abundante de nós e transições, houve a necessidade de construir subGLs para agrupar as regras. A subGL “concomitante.grf”, por exemplo, apresenta diversas especificações em torno da palavra “concomitante”, pois ela se mostrou recorrente à esquerda de “substâncias”.

Além disso, existem outras duas subGLs utilizadas para enumerar entidades, uma vez que alguns contextos podem indicar uma sequência de entidades que precisam ser reconhecidas. Em suma, os arquivos “enumeraComponente.grf” e “enumeraPrincipio.grf” são semelhantes, diferenciando-se apenas por suas anotações: uma trata da enumeração de princípios ativos e anexa a saída correspondente, enquanto a outra trata dos excipientes.

GL para a categoria ABSTRAÇÃO

A seguir serão apresentadas as regras construídas para capturar entidades da categoria ABSTRAÇÃO. Inicialmente, seguindo a mesma estrutura, foi elaborada uma GL para anotar palavras específicas. No entanto, nesse contexto as entidades desejadas apresentam diversas possibilidades de composição. O termo “úlcera”, por exemplo, pode ser especificado como “úlcera duodenal”. Diante dessas mudanças, foi necessário adicionar regras às palavras para garantir a abrangência e a precisão (Figura 12).

Figura 12 – GL “palavrasDiretasABS.grf”



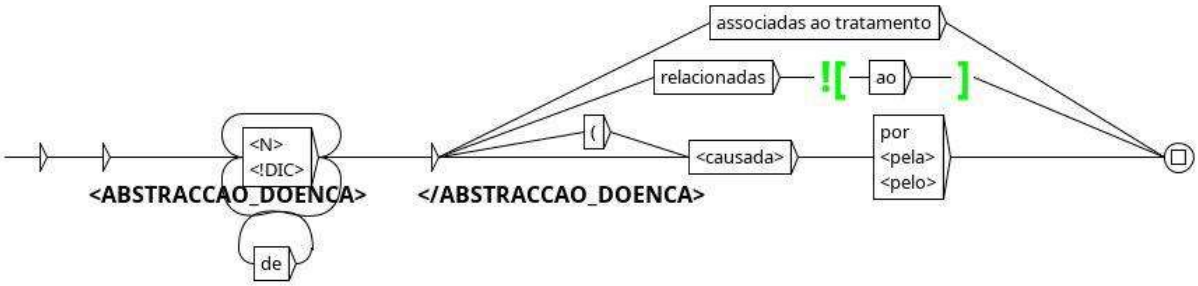
Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Adiante, nas próximas GLs, será possível observar que a *tag* <!DIC> não será mais suficiente, pois as palavras relevantes para a categoria estão, em sua maioria, presentes no dicionário do Unitex. Por isso, a *tag* complementar será <N>, que representa substantivos.

Referente ao lado direito das entidades almejadas, foi construída a GL "contextoDireitaABS.grf" (Figura 13), que se baseou em sentenças observadas como "infecção micótica (causada por fungos)", "toxicidades relacionadas ao metotrexato".

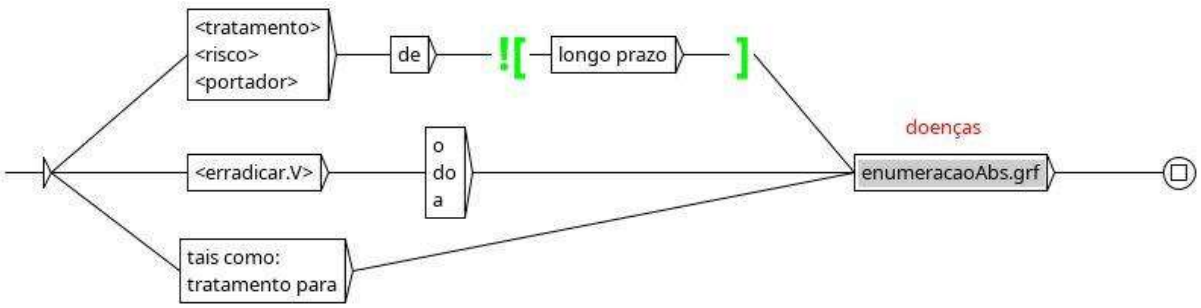
Figura 13 – GL "contextoDireitaABS.grf"



Fonte: o autor (2025)

Por fim, é apresentado o arquivo "contextoEsquerdaABS.grf" (Figura 14), que unificou as regras observadas à esquerda de entidades relacionadas à ABSTRAÇÃO. Algumas sentenças reconhecidas são "tratamento de otite", "Erradicação da bactéria Helicobacter pylori", "tais como: pirose, azia". É possível notar que também foi construída uma subGL para capturar enumerações.

Figura 14 – GL "contextoEsquerdaABS.grf"



Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Com a implementação e explicação dos padrões apresentados, a etapa seguinte consiste na avaliação dos resultados obtidos após a aplicação das GLs nos dois *corpus* que foram descritos anteriormente.

Obtenção das Métricas

Ao todo, foram elaboradas 16 GLs com base no estudo linguístico realizado. De acordo com os critérios e métricas previamente estabelecidos, aplicou-se o *script* de avaliação do Segundo HAREM para a geração dos resultados, considerando os cenários COISA e ABSTRAÇÃO, sendo as duas categorias analisadas simultaneamente. A consistência no desempenho das GLs é crucial para avaliações conjuntas, uma vez que os erros de uma categoria influenciam diretamente os resultados da outra. A Tabela 2 apresenta os resultados obtidos com a aplicação dos *scripts* sobre o *corpus* utilizado no estudo linguístico.

Tabela 2 – Resultados

PRECISÃO (%)	ABRANGÊNCIA (%)	MEDIDA-F (%)
66	18	29

Fonte: o autor (2025)

Com o objetivo de avaliar individualmente as GLs de COISA e ABSTRAÇÃO, também foram gerados resultados avaliando os cenários correspondentes a cada categoria. A comparação é exibida na Tabela 3.

Tabela 3 – Resultados individuais

CENÁRIO	PRECISÃO (%)	ABRANGÊNCIA (%)	MEDIDA-F (%)
COISA	80	36	50
ABSTRAÇÃO	16	3	5

Fonte: o autor (2025)

Revista brasileira de iniciação científica
ISSN: 2359-232X

Similarmente, os mesmos testes foram efetuados para o *corpus* SemClinBr. A Tabela 4 apresenta os resultados para a avaliação conjunta e a individual.

Tabela 4 – Resultados para o *corpus* SemClinBr

CENÁRIO	PRECISÃO (%)	ABRANGÊNCIA (%)	MEDIDA-F (%)
COISA;ABSTRAÇÃO	62	2	4
COISA	73	7	13
ABSTRAÇÃO	25	0,4	0,8

Fonte: o autor (2025)

As GLs reconhecedoras da categoria COISA no geral apresentaram melhor desempenho, como visto nas Tabelas 3 e 4, justificado pelo uso da *tag* lexical <!DIC>, que simplificou e auxiliou o reconhecimento das entidades corretas, exceto por alguns contextos à direita, que demandaram alguns ajustes para evitar falso-positivos. Devido à estratégia utilizada, essas GLs apresentaram uma boa precisão.

Observou-se que a anotação direta de termos contribuiu para ampliar a abrangência de entidades reconhecidas. Como exemplo, no *corpus* de Colombo e Oliveira (2022), o termo "medicamento" apresentou 400 ocorrências dentro de um total de 2.182 ENs classificadas como a categoria COISA. O aprimoramento dos contextos à esquerda que antecedem enumerações representou um avanço significativo, permitindo o reconhecimento de múltiplas ENs. A Figura 15 ilustra as entidades identificadas a partir do contexto "Excipientes:".

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 15 – Trecho da saída anotada pela GL

```
Excipientes:<EM ID="gastrium-6" CATEG="COISA" TIPO="EXCIPIENTE">sacarose</EM>,
<EM ID="gastrium-7" CATEG="COISA" TIPO="EXCIPIENTE">manitol</EM>,
<EM ID="gastrium-8" CATEG="COISA" TIPO="EXCIPIENTE">carbonato de cálcio</EM>,
<EM ID="gastrium-9" CATEG="COISA" TIPO="EXCIPIENTE">fosfato de sódio dibásico</EM>,
<EM ID="gastrium-10" CATEG="COISA" TIPO="EXCIPIENTE">laurilsulfato de sódio</EM>,
<EM ID="gastrium-11" CATEG="COISA" TIPO="EXCIPIENTE">metilparabeno sódico</EM>,
<EM ID="gastrium-12" CATEG="COISA" TIPO="EXCIPIENTE">propilparabeno</EM>,
<EM ID="gastrium-13" CATEG="COISA" TIPO="EXCIPIENTE">povidona</EM>,
<EM ID="gastrium-14" CATEG="COISA" TIPO="EXCIPIENTE">hipromelose</EM>,
<EM ID="gastrium-15" CATEG="COISA" TIPO="EXCIPIENTE">polimetacrilicocopolialato de etila</EM>,
<EM ID="gastrium-16" CATEG="COISA" TIPO="EXCIPIENTE">dietilftalato</EM>,
<EM ID="gastrium-17" CATEG="COISA" TIPO="EXCIPIENTE">dióxido de titânio</EM>,
<EM ID="gastrium-18" CATEG="COISA" TIPO="EXCIPIENTE">talco</EM>,
<EM ID="gastrium-19" CATEG="COISA" TIPO="EXCIPIENTE">polissorbato 80</EM>,
<EM ID="gastrium-20" CATEG="COISA" TIPO="EXCIPIENTE">hidróxido de sódio</EM>.
```

Fonte: o autor (2025)

No caso da categoria COISA, a precisão indica que 80% das entidades foram identificadas corretamente para o *corpus* utilizado no estudo linguístico e 73% para o *corpus* SemClinBr. Apesar disso, é necessário construir novas GLs para aumentar a abrangência em ambos os casos.

Para a categoria ABSTRAÇÃO, os resultados não foram tão expressivos devido à baixa precisão e pouca abrangência em ambos os *corpus*. As baixas métricas dessa categoria impactaram negativamente os resultados da avaliação conjunta para as duas categorias, como exposto nas Tabelas 2 e 4. Visando melhorar esses resultados, torna-se essencial o aprimoramento das GLs para a categoria ABSTRAÇÃO.

Uma possível causa para a baixa precisão dessa categoria é que, para abranger todas as possíveis ENs, utilizou-se a *tag* <N>, que representa substantivos. Contudo, essa abordagem resultou em falso-positivos, como por exemplo na sentença “tratamento de suporte”, que foi erroneamente identificada como entidade, visto que “suporte” é classificado como substantivo, assim como outras palavras que deveriam ser reconhecidas corretamente, como em “tratamento de artrite”. Neste contexto, torna-se essencial a definição de outras regras que possam contribuir para o melhor reconhecimento dessas entidades. O uso de funcionalidades avançadas do Unitex pode auxiliar na elaboração de regras mais precisas.

No *corpus* SemClinBr, o total de entidades reconhecidas foi de 861, sendo que 632 pertencem à categoria COISA, e 229, à categoria ABSTRAÇÃO. Os resultados obtidos nesse *corpus* não puderam ser comparados com os de outros estudos, uma vez que as categorias das

entidades divergem das originalmente propostas pelos autores do *corpus*.

A análise dessas entidades revelou a ocorrência de alguns falso-positivos. Dentre eles, para categoria COISA, destacaram-se palavras comuns como “tipoia”, que foi reconhecida erroneamente 4 vezes. Isso ocorre devido à presença da *tag* <!DIC> na GL “contextoEsquerda.grf”, pois por mais que a palavra não seja altamente técnica, o dicionário do Unitex não a reconhece. Para a categoria ABSTRAÇÃO, o termo “doença” impactou negativamente, visto que no *corpus* anotado manualmente essa palavra não possui nenhuma anotação. Logo, as 22 ocorrências foram consideradas falso-positivos.

Por outro lado, observou-se a existência de entidades que foram reconhecidas corretamente, ou seja, verdadeiro-positivos. Dentre elas, estão termos anotados diretamente, como “úlceras”, que ocorreram 36 vezes, sendo todas identificadas corretamente. Ademais, alguns contextos foram eficientes para o reconhecimento de determinadas ENs. Para a categoria COISA, o uso do contexto à esquerda “em uso de” (como em “em uso de LOSARTAN”) resultou em 109 verdadeiros-positivos, demonstrando a alta precisão dessa regra específica.

Os resultados indicam que a utilização isolada de GLs não foi suficiente para garantir elevada cobertura na tarefa de REN neste trabalho. Futuramente pode-se identificar novas regras capazes de melhorar o desempenho obtido. Porém, a baixa abrangência observada sugere que abordagens baseadas exclusivamente em regras tendem a apresentar limitações quanto à generalização, reforçando a necessidade de estratégias complementares, usando modelos estatísticos ou híbridos, como o CRF+LG.

Considerações finais

Este trabalho demonstrou a possibilidade de utilizar Gramáticas Locais (GLs) específicas para o Reconhecimento de Entidades Nomeadas (REN) no contexto da saúde, onde há dificuldades na extração de informações. No total foram construídas 16 GLs para o reconhecimento das entidades de interesse.

Inicialmente, os resultados revelaram-se satisfatórios para o domínio compreendido no estudo linguístico, que foi realizado no *corpus* de Colombo e Oliveira (2022). Teve-se um bom reconhecimento de substâncias no geral (categoria COISA), alcançando 80% de precisão. Entretanto, para o reconhecimento de sintomas e doenças (categoria ABSTRAÇÃO) é

necessário um novo estudo das regras linguísticas para verificar a viabilidade de aprimorar as GLs para essas entidades, devido a falso-positivos apresentados anteriormente.

A mudança de contexto trouxe impactos causados pela diferença da estrutura textual de bulas medicamentosas e relatos clínicos, visto que bulas devem seguir regras de acordo com o órgão regulamentador. Sendo assim, no *corpus* SemClinBr, os relatos tinham um padrão textual diferente, em comparação com os da SciELO, e o uso de jargões médicos e abreviações dificultou o reconhecimento correto das entidades, gerando diversos falso-positivos. Portanto, para esse *corpus* é necessário reconsiderar o uso da *tag* lexical <!DIC> e analisar contextos de jargões frequentemente utilizados por profissionais da saúde.

Os resultados indicam que, para a abordagem linguística ser efetiva, é crucial a identificação de padrões mediante o domínio desejado e considerar as particularidades de cada *corpus*, possibilitando melhorar os resultados com GLs específicas para cada caso.

Como trabalhos futuros, pretende-se aprimorar as GLs para melhorar o desempenho de todas as métricas em geral, utilizando de recursos mais avançados da ferramenta Unitex. Também pretende-se analisar melhor os resultados no *corpus* SemClinBr, para identificar quais regras estão se adaptando bem e quais estão prejudicando os resultados, criando GLs específicas para esse domínio. Além disso, pretende-se usar a GL construída na abordagem CRF+LG para avaliar o seu potencial nessa abordagem e comparar os resultados com trabalhos correlatos.

Financiamento

A pesquisa recebeu financiamento da FAPES, na modalidade de bolsas de Iniciação Científica e Tecnológica (ICT). Edital FAPES nº 28/2022 - Universal.

Referências

BRASIL. Agência Nacional de Vigilância Sanitária. **Resolução da Diretoria Colegiada (RDC) nº 47, de 6 de setembro de 2009**. Brasília, 2009. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2009/rdc0060_26_11_2009.html. Acesso em: 06 jun. 2025.

BRASIL. Agência Nacional de Vigilância Sanitária. **Resolução da Diretoria Colegiada (RDC) nº 60, de 26 de novembro de 2012**. Brasília, 2012. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2012/rdc0060_12_12_2012.html. Acesso em: 06 jun. 2025.

BRASIL. Agência Nacional de Vigilância Sanitária. **Resolução da Diretoria Colegiada (RDC) nº 770, de 20 de dezembro de 2022**. Brasília, 2022. Disponível em: https://anvisalegis.datalegis.net/action/UrlPublicasAction.php?acao=abrirAtoPublico&num_ato=00000770&sgl_tipo=RDC&sgl_orgao=RDC/DC/ANVISA/MS&vlr_ano=2022&seq_ato=000&cod_modulo=134&cod_menu=1696. Acesso em: 06 jun. 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.

COLOMBO, Cristiano da Silveira; OLIVEIRA, Elias Silva de. Intelligent information system for extracting knowledge from pharmaceutical package inserts. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 18., 2022, Curitiba. **SBSI**. 2022. p. 1 - 9. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/21388>. Acesso em: 20 jun. 2025.

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.

MOTA, Cristina; SANTOS, Diana. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas:** O Segundo HAREM. Porto: Linguateca, 2008. Disponível em: <https://comum.rcaap.pt/entities/publication/4a8de0bd-8f7e-47df-b194-472136104564>. Acesso em: 05 jun. 2025.

OLIVEIRA, Lucas Emanuel Silva et al. SemClinBr-a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. **Journal of Biomedical Semantics**, Curitiba, v. 13, n. 1, p. 13, 2022.

PINEDA-LEGUÍZAMO, Ricardo; MIRANDA-NOVALES, Guadalupe; VILLASÍS-KEEVER, Miguel Ángel. La importancia de los reportes de casos clínicos en la investigación. **Revista Alergia México**, Ciudad de México, v. 65, n. 1, p. 92-98, 2018.

PIROVANI, Juliana Pinheiro Campos. **CRF+ LG:** uma abordagem híbrida para o reconhecimento de entidades nomeadas em português. Tese (Doutorado em Ciência da Computação). Universidade Federal do Espírito Santo, Vitória, 2019.

PIROVANI, Juliana Pinheiro Campos; OLIVEIRA, Elias. Studying the adaptation of Portuguese NER for different textual genres. **The Journal of Supercomputing,** Vitória, v. 77, n. 11, p. 13532-13548, 2021.

Revista brasileira de iniciação científica
ISSN: 2359-232X

SILVA, Tatiane da et al. Bulas de medicamentos e a informação adequada ao paciente. **Revista de Saúde Pública**, Porto Alegre, v. 34, n. 2, p. 184-189, 2000.

ZHOU, GuoDong; SU, Jian. Named entity recognition using an HMM-based chunk tagger. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Philadelphia. **ACL**. 2002. p. 473-480. Disponível em: <https://www.researchgate.net/publication/220874212_Named_Entity_Recognition_using_an_HMM-based_Chunk_Tagger>. Acesso em: 05 jun. 2025.