Um analisador mórfico-morfêmico: aplicação web para verbos regulares do português

Wagner Ferreira Lima
Departamento de Letras Vernáculas e Clássicas
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
wflima@uel.br

Cinthyan Renata Sachs C. de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
cinthyan@uel.br

RESUMO

O Processamento da Linguagem Natural pode apoiar o ensino da morfologia dos verbos regulares do português com analisadores morfológicos. Contudo, os analisadores baseados em modelos de linguagem são muito generalistas, tal que se mostram insuficientes para o cumprimento de tal tarefa. Assim, propõe-se um analisador coerente com a abordagem da morfologia estrutural. O objetivo deste artigo é apresentar o "Analisador mórfico-morfêmico", uma aplicação web que recebe verbos conjugados como entrada e retorna uma série de informações metalinguísticas sobre eles. Ele consegue cobrir a infinidade de verbos de 1ª, 2ª e 3ª conjugações. Conclui-se que essa ferramenta é mais adequada ao ensino formal da morfologia.

PALAVRAS-CHAVE

Analisador mórfico-morfêmica; aplicação web; morfologia estrutural.

ABSTRACT

Natural Language Processing can support the teaching of the morphology of regular Portuguese verbs with morphological analyzers. However, the analyzers based on language models are very general, so they are not adequate to fulfil this task. Therefore, an analyzer that is coherent with the structural morphology approach is proposed. The aim of this paper is to present the "Morphic-Morphemic Analyzer", a web application that takes conjugated verbs as input and provides a range of metalinguistic information about them. This tool can cover an unlimited 1st, 2nd and 3rd conjugation verbs. The conclusion is that this is the most suitable tool for formal morphology teaching.

CCS CONCEPTS

• Human-centered computing • Human computing interaction (HCI) • Interactive systems and tools

KEYWORDS

Morphic-morphemic analyzer; web application; structural morphology.

1 Introdução

Com o advento da Aprendizagem de Máquina, é esperado que as tecnologias em Inteligência Artificial (IA) possam colaborar com o ensino da linguagem humana. Atualmente, o Processamento da Linguagem Natural (PLN), um ramo da IA que usa ferramentas computacionais na análise e produção da linguagem verbal [1] [2]; oferece muitos recursos e ferramentas para o tratamento computacional da linguagem.

Esses recursos compreendem, entre outras coisas [2], os chamados "modelos de línguas" (ML), algoritmos treinados em *corpora* (plural de *corpus*) linguísticos, usando métodos e estratégias de aprendizagem de máquina, como as redes neurais artificiais multicamadas [2]. *Corpus* é uma coleção de documentos ou textos, que pode ser processada por computadores [3]. Por isso, ela precisa ser representativa, balanceada e informativa do comportamento linguístico que se quer apreender [4].

Já as referidas ferramentas incluem aos diferentes comandos para pré-processamento de dados linguísticos, como os "tokenizadores" (segmentadores de textos em *tokens* – sequência de caracteres delimitados a princípio por espaço em branco), os "*stemmers*" (extratores da raiz de uma palavra), os "lematizadores" (transformadores de uma palavra em lema) etc.; os quais preparam os dados brutos para posterior processamento [2]. Portanto, essas ferramentas servem à produção das aplicações práticas propriamente ditas, tais como sumarização de textos, mineração de dados, tradução automática, agentes conversacionais, para citar apenas algumas.

É notável a importância de tais aplicações no setor de serviços. Problemas práticos como reconhecimento de spams em e-mails, atendimentos eletrônicos de bancos, sistemas de busca na web, enfim, serviços *online* em geral, podem ser solucionados com base nos conhecimentos produzidos em PLN.

Já no tocante à educação, especificamente no campo do ensino do vernáculo, não é comum encontrar esse estado de coisas, provavelmente porque os problemas de ensino/aprendizagem do vernáculo requerem soluções mais complexas do que aquelas que a IA pode oferecer.

Um exemplo dessas limitações é a descrição morfológica fornecida por ML, a qual é tipicamente desprovida de representações que capturem a organização interna da palavra considerada. Em geral, os *tokens* obtidos a partir da segmentação dos textos que constituem os *corpora* são classificados em termos morfossintáticos e definidos de acordo com as suas características

morfológicas [5]. Essa forma de descrição, conquanto útil para as tarefas práticas envolvendo PLN, é inadequada ao entendimento da morfologia estrutural do português. Essa última requer uma análise mais especializada da morfologia da Língua Portuguesa, a qual se assenta sobre uma concepção estrutural dos fatos linguísticos, como será mostrado a seguir.

Nas situações em que os verbos estão envolvidos, a descrição verbal baseada em Aprendizagem de Máquina suscita questões sem respostas tais como: "Quais são as unidades mínimas e significativas da palavra verbal, ou seja, os morfemas constitutivos do verbo? Como elas se instanciam como segmentos fônicos ou ortográficos, ou seja, como alomorfes? De que maneira essas unidades se organizam formando uma palavra, e não uma mera sequência de caracteres? Entre outras questões.

Uma possível forma de se superar a limitação de analisadores morfológicos baseados em ML é construir aplicações baseadas em regras, ao invés de algoritmos treinados em *corpora*. Em vista disso, uma aplicação web foi criada que analisa os verbos regulares do português, em consonância com a morfologia estrutural [6]. Essa forma de analisar os verbos é baseada especialmente na visão de Câmara Jr. [7], um dos pioneiros no Brasil da análise morfológica, e de Laroca, uma pesquisadora da morfologia brasileira [8]

O objetivo deste trabalho é, então, descrever a construção da referida aplicação e as possíveis contribuições dela para o ensino do vernáculo em nível acadêmico. Chamar-se-á doravante essa aplicação de "Analisador mórfico-morfêmico", porque, como se verá, ela admite uma análise em nível tanto dos morfemas quanto dos alomorfes.

Vale lembrar que o fato de a análise estrutural dos verbos ser necessário não significa, contudo, que ela seja facilmente assimilável. Pelo contrário, trata-se de um processo pouco intuitivo e, *ipso facto*, muito demandante em termos cognitivos. Assim, essa aplicação se justifica também por ser uma ferramenta pedagógica. E, porque essa ferramenta funciona como um meio a auxiliar o ensino da morfologia, sua aplicação dispensa a princípio uma validação empírica.

Este texto está organizado como segue. Em "Análise mórfico-morfêmica" é exposta a concepção teórico-metodológica da morfologia verbal de um ponto de vista estrutural. Os requisitos que satisfazem essa abordagem são apresentados que podem ser usados para avaliar a abordagem baseada em ML. Em "Modelos de linguagem" é brevemente definida e comentada a noção de ML. Será evidenciado como a descrição inerente ao modelo de linguagem é insuficiente para o ensino da morfologia verbal. "Trabalhos correlatos" apresenta os comentários acerca de trabalhos acadêmicos que também objetivaram a análise da morfologia portuguesa.

"Materiais e métodos" apresenta o Flask, o *framework* usado na construção da referida aplicação. Esse *framework* é combinado com outras duas tecnologias de desenvolvimento web, a saber, HTML e CSS. Em "Resultados" são mostradas e comentadas as funcionalidades do analisador mórfico-morfêmico. "Discussão geral", por sua vez, traz uma discussão que compara e avalia as duas abordagens e enfatiza a importância do referido analisador para o ensino da morfologia em nível acadêmico. "Considerações finais", por fim, apresenta os limites da aplicação e caminhos futuros.

2 **Análise mórfico-morfêmica**

Tradicionalmente, morfologia é a parte da gramática que trata da flexão e derivação das palavras, mas o estudo científico dos fatos morfológicos só apareceu no início do século XX, com o advento da linguística estrutural. Bloomfield [6] [9] estabeleceu os constituintes das palavras, os morfemas e o método para a sua identificação, a segmentação e a comutação [6], duas operações interdependentes.

Assim, a estrutura de um verbo é explicitada pela divisão do todo em seus morfemas constituintes (por exemplo, "cantávamos" segmenta-se em "cant/á/va/mos") e pela substituição de cada morfema por outro da mesma classe (a comutação de "cant-" por outra raiz, como "fal-", dá lugar a "falávamos"). Entretanto, a expressão das estruturas morfológicas raramente é uniforme ao longo das conjugações.

A irregularidade morfológica decorrente dessa variação é denominada alomorfia e o produto dela é o alomorfe. Por exemplo, a mesma noção de "imperfeito do indicativo" se expressa mediante dois alomorfes distintos: "-va" em "(ele) cantava" e "-ia" em "(ele) comia".

Na década de 60, a morfologia passou a ser tratada de acordo com dois planos de análise: o mórfico (de morfe) e o morfêmico (de morfema) [10] [11]. Segundo essa visão, os morfemas são as unidades formais que permitem a identificação das palavras. Devido à sua particularidade formal, eles são indicados na descrição por meio de chaves "{ }". Já os morfes são os segmentos efetivos, seja fônico ou ortográfico, pelos quais os morfemas se expressam na palavra (Figura 1). A seguir será priorizada a expressão ortográfica dos morfemas.

Vale lembrar, contudo, que, devido à natureza formal do morfema, a expressão dele por meio dos morfes não é necessária. É o caso da noção de "pretérito perfeito do indicativo", cuja compreensão não depende da presença de um morfe correspondente específico (salvo na 6ª pessoa): "(eu) cant-e-0-i" *vs* "(eles) cant-a-ra-m".

Tal análise em dois planos permite evidenciar também que um mesmo morfema pode admitir morfes alternativos, denominados alomorfes: {1ª pessoa} = "-o" em "(eu) canto" *vs* "-i" em "eu cantei".

Em sua análise, Laroca [8] considera, ademais, os processos morfofonêmicos, tais como a "neutralização morfológica e crase"

(p. 55-56): A oposição de conjugação (p.ex., "vender" *vs* "partir") se neutraliza na "1ª pessoa do pretérito perfeito", e os seus respectivos alomorfes sofrem crase: "vend - i - 0 - i" e "part - i - 0 - i"; para citar apenas um exemplo. Cognitivamente, essa análise é mais demandante, porém, em compensação, é reveladora da estrutura interna dos verbos.

A melhor representação da morfologia estruturalista para o português é, sem dúvida, os trabalhos de Câmara Jr [7]. Mesmo não operando com a teoria de dois planos, o autor é a principal referência dos estudos morfológicos no Brasil [8]. Com efeito, Câmara Jr. propôs um esquema estrutural para os verbos regulares que, incrementado com a teoria dos dois planos, permite a uma análise mais contundente com os fatos morfológicos do português. A título de exemplo, a Figura 1 mostra esse esquema para o verbo "(nós) cantaremos"[1].

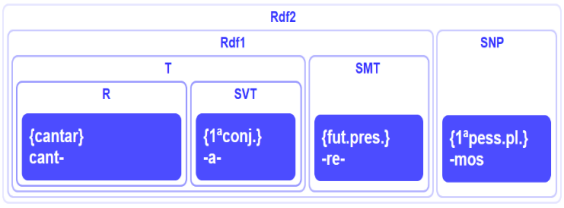

Figura 1: Esquema estrutural verbo "cantaremos". Os alomorfes com seus respectivos morfemas são mostrados dentro das caixas azuis. (Fonte: Autores)

Em suma, depois de tudo é possível estabelecer uma síntese dos requisitos de uma análise estrutural dos verbos e, a partir disso, comparar com as análises oferecidas pelos modelos de linguagem.

(1) Decompor as palavras em seus segmentos, ou alomorfes, constituintes (p.ex., "cantaremos" = "cant-/-a-/-re-/-mos").

(2) Correlacionar tais alomorfes com seus respectivos morfemas (p.ex., "cant-" = "{cantar}").

(3) Classificar os morfemas identificados com as classes morfológicas a que eles pertencem (p.ex., "ação de cantar" pertence a categoria "R" (raiz)).

(4) Plotar um digrama da organização interna da palavra, ou seja, sua estrutura, de acordo com o esquema descrito acima (p.ex., diagrama mostrado na Figura 1).

Após essa descrição teórica, apresentando a concepção da morfologia estrutural, uma pergunta que deve ser respondida é se os analisadores existentes baseados em Aprendizagem de Máquina são capazes de realizar uma análise que seja adequada com a estrutura morfológica dos verbos. A seguir, pretende-se responder a essa questão, esclarecendo os ML e evidenciando suas limitações como ferramenta para análise verbal.

3 Modelos de Linguagem

Embora possa decorrer de um sistema de regras simbólicas explicitamente declaradas [2], um Modelo de Linguagem (ML) é tipicamente considerado em PLN como um recurso[2] que deriva de um algoritmo de Aprendizado de Máquina (AM) treinado em *corpora* [2] [12]. É nessa acepção *stricto sensu* de aprendizado de máquina que o termo é empregado neste trabalho.

Um ML pode ser criado no interior de três paradigmas: o simbólico, o estatístico e o neural [2] [12].

O simbólico é o paradigma inicial, que teve início na década de 50. Ele é baseado na explicitação de regras de como o computador deve processar os dados linguísticos, dispensando assim o treinamento em *corpora*. O analisador a ser descrito, mesmo não sendo tratado como um ML *stricto sensu*, representa esse paradigma simbólico.

Já o estatístico, anos 1980 a 2010, caracteriza-se pelo uso de estatística e probabilidade aplicadas a *corpora* [12]. É propriamente o início da chamada era do Aprendizado de Máquina. O aprendizado em questão é do tipo supervisionado, da maneira que os dados linguísticos de treinamento devem obrigatoriamente ser anotados. Desde 2010, com o advento das redes neurais artificiais, esse paradigma vem perdendo lugar para o paradigma neural.

O neural é assim conhecido porque baseia-se em algoritmos de redes neurais artificiais multicamadas, como a rede *Transformer*, os quais aprendem a reconhecer padrões a partir de grandes quantidades de dados e sem a declaração explícita de quais padrões linguísticos encontrar [2] [13]. A rede *Transformer* é uma arquitetura de rede neural multicamada, proposta em 2017, que faz uso do mecanismo de atenção, entre outros componentes conhecidos de redes neurais [12] [13].

Nesta apresentação, não serão descritos a lógica e os cálculos subjacentes a essa rede[3], uma vez que o objetivo é mais precisamente considerar o que ele pode oferecer em termos de analisador morfológico. Contudo, vale destacar que muitos ML poderosos atuais advêm dessa rede, como é o caso do BERT e dos seus congêneres (DistilBERT, MobileBERT, Funnel Transformers e MPNET). Diferente dos ML inerentes ao paradigma estatístico, BERT usa representações numéricas de tokens que são contextualizadas [13]; de modo a capturar mais propriedades (linguísticas e comunicativas) dos textos.

Quanto aos analisadores morfológicos, até é possível saber, inexiste um analisador morfológico, baseado em AM, que modele os fatos morfológicos do português em termos da concepção linguística descrita na seção anterior.

[1] Lista de símbolos da Figura 1, adaptada de [8]: (R (raiz), ST (sufixo temático), T (tema), SMT (sufixo modo-temporal), SNP (sufixo número-pessoal), Rdf1 (radical flexional 1) e Rdf2 (radical flexional 2).

[2] Outros recursos são o léxico, o *corpus*, o dicionário etc. [2], que não supõem treinamento e podem servir de dados de treinamento para os ML.

[3] Para um panorama do assunto, sugere-se a leitura de [12].

Entretanto, já existem tokenizadores com potencial para cumprir essa tarefa. O *WordPiece* é um algoritmo de tokens de subpalavras da rede *transformer* que é usado para aliviar o problema das palavras que estariam fora de um vocabulário pré-treinado [13] [14]. Ele foi desenvolvido pelo Google para pré-treinar o BERT e seus similares. Seja como for, essa é uma proposta intrigante para o aprimoramento dos analisadores baseados em ML *stricto sensu*.

A Figura 2 mostra a tokenização dos verbos (nós) "cantaremos", (vós) "cantareis", (eles) "cantarão", realizada pelo *WordPiece*, após treinamento em dados dos *corpora* Wikipedia ("./wikitext/wikitext-103-raw-v1.zip").

```
1 output = tokenizer.encode("cantaremos cantareis cantarão")
2 print(output.tokens)
```

```
['cant', '##are', '##mo', '##s', 'cant', '##are', '##is', 'cant', '##ar', '##ão']
```

Figura 2: Análise fornecida pelo tokenizador de subpalavras. O símbolo "#" duplicado iniciando um token indica a posição de um prefixo.

O que o *WordPiece* entrega está longe de ser um analisador morfológico. Isso porque ele foi criado para ser um tokenizador e apenas isso; o que não descarta, como já se sugeriu acima, a possibilidade de ele ser um componente de um analisador morfológico baseado em AM.

Seja como for, os analisadores existentes se aplicam a *tokens* como palavras e, em geral, procedem atribuindo características gramaticais aos segmentos de textos. Um exemplo disso pode ser encontrado examinando o analisador morfológico disponibilizado pelo spaCy [5], uma biblioteca para PLN escrito em Python.

A Figura 3 mostra o resultado da análise de um conjunto de verbos flexionados, a saber, (nós) "cantaremos", (vós) "cantareis" e (eles) "cantarão". Note-se que a análise consiste em atribuir aos *tokens* identificados etiquetas morfossintáticas (*POS tags*) (segunda coluna) e características (*features*) morfológicas (terceira coluna).

```
nós        | PRON  | Case=Nom|Number=Plur|Person=1|PronType=Prs
cantaremos | VERB  | Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin
vós        | ADV   |
cantareis  | ADJ   | Gender=Fem|Number=Plur
eles       | PRON  | Case=Nom|Gender=Masc|Number=Plur|Person=3|PronType=Prs
cantarão   | VERB  | Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin
```

Figura 3: Resultados do analisador morfológico do spaCy, com base no ML "pt_core_news_sm".

Esse formato de descrição, que é um padrão nos trabalhos em PLN, existe para apreender as propriedades gramaticais dos tokens como palavras e permitir que tarefas práticas, como o Reconhecimento de Entidades Nomeadas [15], sejam possíveis. Por isso, representações gramaticais mais específicas, como as de morfemas e categorias morfológicas, úteis para o entendimento da estrutura das palavras, está fora de discussão até o momento.

Uma medida para lidar com essa restrição é, além de criar analisadores de subpalavras mais completos, revisitar o paradigma simbólico/linguístico e implementar analisadores baseados em regras de morfologia estrutural, como é o caso do analisador mórfico-morfêmico aqui apresentado.

Antes, porém, vale saber se há analisadores morfológicos visando exclusivamente o português e se eles cumprem a função de esclarecer a estrutura dos verbos.

4 Trabalhos correlatos

Uma busca no Google Acadêmico pelo termo-chave "analisador morfológico" retornou uma série de trabalhos. Quando essa busca foi filtrada por termos-chave mais específicos, como "trabalhos nacionais", "publicados a partir de 2000" e "seguindo uma abordagem estruturalista", apenas três trabalhos sobressaíram, os quais são brevemente comentados a seguir.

Alencar [16] apresentou o transdutor LEXPOR, protótipo de um componente morfológico do português, que é capaz de segmentar e classificar os constituintes de vocábulos formados por derivação. A característica central desse transdutor é analisar neologismos criados a partir de bases não lexicalizadas, tomadas de empréstimo de outras línguas.

Já Galhardi *et al.* [17] criaram uma analisador léxico-morfológico visando o estudo das redações do ENEM; os vocábulos verbais sendo o foco do trabalho. Com isso, foi possível mostrar as variantes sobre cada palavra e algumas informações morfológicas, como o lema, a raiz, sufixo etc., das palavras.

Vasilévski e Araújo [18], por seu turno, propuseram um analisador morfológico automáticos dos verbos do português que é baseado em processamento das regras do sistema verbal e no tratamento de ambiguidades geradas. Trata-se de um programa muito completo, treinado em corpus, que é capaz de resolver alomorfias consultando o contexto da conjugação verbal.

A qualidade desses trabalhos é indubitável. E, como seria de esperar, os comentários só poderiam ser no sentido de apontar a limitação deles em oferecer *prima facie* um entendimento didático da morfologia verbal. Contudo, essa limitação não será bem explanada até que o analisador mórfico-morfêmico seja descrito e uma base de comparação mais sólida seja oferecida.

5 Materiais e Métodos

A aplicação web foi implementada mediante o uso de três tecnologias: Flask, HTML e CSS. Flask é um micro-framework de aplicação web leve escrito em Python. Flask é um *framewok* em Python para desenvolvimento web [19]. Trata-se de uma estrutura leve e flexível quando comparada com seus congêneres

[20]. Com efeito, as extensões que ele admite não são pré-fixadas, mas sim adicionadas à medida que aplicação se torna mais complexa [19].

Já HTML (Linguagem de Marcação de Hipertexto) foi empregada para definir a estrutura das páginas. Essa linguagem oferece uma diversidade de *tags* que indicam qual a função dos elementos na página, vale dizer, texto, botão, formulário, cabeçalho, rodapé etc. Os arquivos de *scripts* HTML são tecnicamente denominados "*templates*" e são implementados mediante a função "*render_template*" do Flask.

Contudo, HTML *per se* não é suficiente para a criação de uma interface amigável, uma vez que é destituído de aspectos estéticos. É nesse ponto que CSS (Folha de Estilo em Cascata) é necessário. CSS é uma linguagem de estilização que atribui aos elementos da página HTML cores, formas, estilo, espaçamento, efeitos visuais etc. Os comandos de CSS são tipicamente escritos em um arquivo à parte e importados diretamente dentro das *tags*.

Pensando na inclusão digital, empregou-se neste analisador web caracteres grandes e contrastantes com o fundo. Além disso, optou-se por uma escala monocromática de azul claro com o objetivo de criar uma visão suave das páginas e promover identidade visual entre elas.

Os comandos centrais do Flask permitiram a criação da inteligência operando por trás da interface estética do site. As tarefas de análise que estavam encapsuladas na função elaborada originalmente [21] foram divididas em arquivos Python individuais e, posteriormente, importadas em um novo arquivo de criação de páginas web para análise dos dados de entrada. Assim, foi possível colocar cada tarefa específica de análise (os requisitos descritos na "Introdução") em páginas distintas.

Em síntese, procedeu-se da seguinte maneira: as saídas de cada arquivo do Flask foram renderizadas em uma página web, mediante a transformação delas em um arquivo HTML, e sua posterior estilização, usando os comandos do CSS. O resultado foi um site por meio do qual se pode navegar clicando em botões distribuídos segundo duas classes de funcionalidades: (a) oferecer informações teórico-metodológicas sobre a morfologia e (b) realizar análises mórfico-morfêmicas dos verbos dados como entrada.

(a) Botões de informação redirecionam o usuário para páginas contendo uma explicação dos temas da morfologia: Morfologia, Morfema, Alomorfe etc. Tais botões servem para informar acerca da teoria e método empregados na descrição da morfologia dos verbos. Eles são localizados na barra de navegação no cabeçalho do site (Figura 4).

(b) Já os botões de análise redirecionam o usuário para as páginas apresentando a saída dos dados processados. São oito botões dispostos em coluna no corpo da página, cada um dos quais com uma funcionalidade analítica específica (Figura 4).

Esses botões permitem a realização da análise mórfico-morfêmica dos verbos regulares e, *ipso facto*, satisfazem os requisitos esperados de uma análise morfológica estrutural. Na seção seguinte são apresentados os resultados dessa aplicação.

Figura 4: Página principal da aplicação. Botões em linha no cabeçalho redirecionam para informações teórico-metodológicas. Botões em coluna no corpo da página realizam e plotam análises morfológicas.

6 Resultados

Ao entrar no site, o usuário se depara com uma explicação da ferramenta, bem como as opções de conjugar o verbo no modo indicativo ou no modo subjuntivo. Qualquer que seja a escolha feita, ele será redirecionado para a página com o formulário para entrada dos dados (Figura 4). Uma vez inseridos os dados, o verbo flexionado nas seis pessoas, esses serão analisados de acordo com a metodologia da análise estrutural e retornados na forma de uma descrição linguística. Os botões de análise são indicados por rótulo da informação metalinguística que se deseja obter (Figura 4). São elas:

(1) "Paradigma de conjugação verbal": informa se o verbo de entrada pertence à 1ª conjugação (final "-ar": "cantar"), à 2ª conjugação (final "-er": "beber") ou à 3ª conjugação (final "-ir": "sair").

(2) "Flexão verbal": plota o verbo conjugado nas seis pessoas e define o tempo-modo em que foi flexionado (p.ex., "cantarei", "cantarás", "cantará" etc.: "Futuro do presente do indicativo").

(3) "Alomorfes de raiz": retorna os morfes pelos quais o morfema de R (raiz) do verbo se expressa. Por se tratar de verbos regulares, que por definição são invariáveis, os alomorfes apresentam a mesma forma ortográfica: no exemplo do verbo "cantar", "cant-" para toda flexão.

(4) "Alomorfes de sufixo temático": apresenta os alomorfes de vogal temática (SVT) que aparecem nos verbos flexionados dados como entrada. Porque os morfemas temáticos sofrem alomorfia, os alomorfes podem ser distintos: "-a-" para "(eu) cantarei"; "-e-" para "(eu) cantei", "-0-" (zero) para "(eu) canto".

(5) "Alomorfes de sufixo modo-temporal": plota os morfes que ocupam a posição de SMT (sufixo modo-temporal) de acordo com esquema acima. Essa página mostra que, por exemplo, o morfema de "futuro do presente do indicativo" assume alomorfes

distintos: “-ra-” em “(tu) cantarás” e “(ele) cantará”; “-re-” em “(eu) cantarei”, “(nós) cantaremos” e “(vós) cantareis”; e finalmente “-rã-” em “(eles) cantarão”.

(6) “Alomorfe de sufixos número-pessoais”: retorna o alomorfe para cada morfemas de número-pessoa do verbo. Por exemplo, “-i” para “1ª pessoa do singular” ((eu) “cantarei”); “-s” para “2ª pessoa do singular” ((tu) “cantarás”); “-0” para “3ª pessoa do singular” ((ele) “cantará”); “-mos” para “1ª pessoa do plural” ((nós) “cantaremos”); “-is” para “2ª pessoa do plural” ((vós) “cantareis”); e por fim “-o” para “3ª pessoa do plural” ((eles) “cantarão”).

(7) “Morfemas do verbo”: essa página mostra os morfemas que constituem os verbos conjugados. Se os alomorfes podem ser vários, os morfemas, por outro lado, só podem ser um único, porque são eles que conferem identidade aos segmentos que formam as palavras. A Figura 5 deixa isso mais explícito.

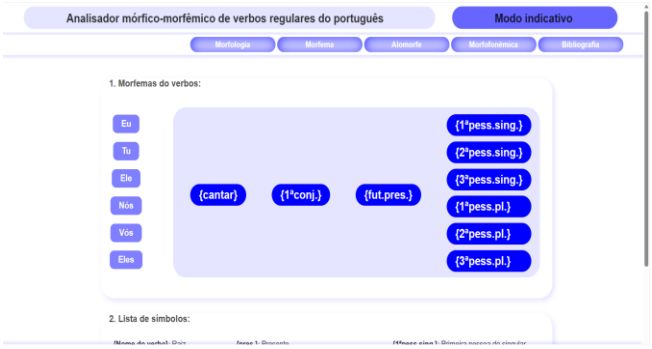

Figura 5: Os morfemas são mostrados ocupando a estrutura do verbo, linearmente dispostos em R-ST-SMT-SNP, ocupando a posição das categorias a que pertencem.

(8) “Estrutura do verbo”: apresenta uma síntese das informações metalinguísticas anteriores na forma de um diagrama na forma de blocos aninhados. Longe de ser arbitrário, esse diagrama reflete a organização interna subjacente a cada flexão verbal (Figura 6).

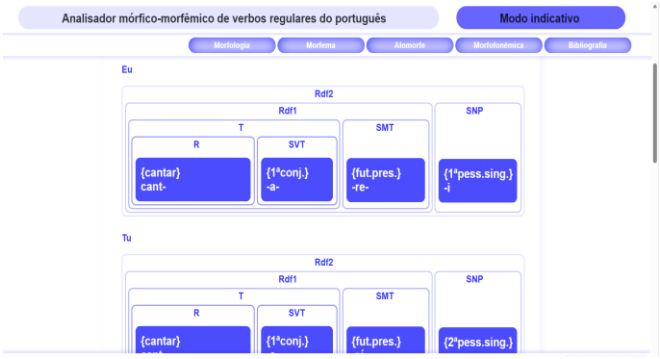

Figura 6: Diagramas em caixas aninhadas evidenciando a organização interna das palavras verbais.

Cada bloco encaixado representa uma categoria morfológica. O rotulado na parte superior de cada um indica a que categoria representada. Rdf2 (Radical flexional 2) é o bloco maior que compreende todos os demais e indica o limite da estrutura do verbo flexionado. Os outros subordinados hierarquicamente são: Rdf1 (Radical flexional 2), SNP (sufixo número-pessoal), T (Tema), SMT (Sufixo modo-temporal), R (Raiz), SVT (Sufixo de vogal temática). Os alomorfes com seus respectivos morfemas são indicados no bloco preenchido na base de cada diagrama.

Vale lembrar que cada página traz também uma “lista de símbolos” informando o significado dos rótulos de notação empregados. Assim, o usuário não precisa voltar à teoria para entender o que está se passando.

Em suma, até onde foi possível mostrar, essas funcionalidades permitem uma análise estrutural dos verbos regulares que é mais contundente com uma teoria linguística da morfologia verbal (requisitos 1-4, em “Análise mórfico-morfêmica”). Que isso é o caso pode ser visto na seção seguinte, onde uma análise das descrições fornecidas por modelos de língua e pelo analisador mórfico-morfêmico será realizada.

7 Discussão geral

Esta apresentação iniciou-se com um questionamento acerca da capacidade dos analisadores baseados em modelos de linguagem de realizar uma análise morfológica em termos estruturais. Esse tipo de análise e descrição é central para o ensino/aprendizado da morfologia especialmente em cursos de graduação em Letras, em que os formandos precisam adquirir um conhecimento não apenas dos aspectos funcionais-comunicativos do vernáculo, como também da estrutura inerente às expressões da língua.

Como foi dito, ML são, em síntese, algoritmos treinados em base de dados linguísticos (*corpora*) tal que, uma fez construídos, possam fazer previsões sobre novos comportamentos linguísticos. Aqui se podem destacar duas limitações de modelos de linguagem: (a) as previsões que eles permitem dependem de comportamentos presentes nos *corpora* de treinamento; e (b), no tocante às descrições verbais, eles se resumem a fornecer etiquetagem morfossintática e caracterização morfológica dos verbos.

Um problema decorrente de (a) pode ser verificado nos dados da Figura 3, que mostra uma análise equivocada dos verbos (nós) “cantaremos” e (vós) “cantareis”. No primeiro caso, o verbo é definido como “presente” em vez de “futuro” (“Tense=Pres”) e, no segundo, como “adjetivo” (“ADJ”) em vez de “verbo” (“VERB”). Obviamente, por se tratar de um modelo de linguagem, os *corpora* podem ser retreinados e essa análise corrigida; este portanto sendo o menor dos problemas.

Já um problema relativo a (b) é que a concepção analítica por trás desses modelos não satisfaz os requisitos de uma análise morfológica estrutural nos termos das pesquisas linguísticas. Isso

pode ser observado na caracterização morfológica dos verbos, mostrada na Figura 3. Veja-se "(eles) cantarão": "VERB | Mood=Ind | Number=Plur | Person=3 |Tense=Fut | VerbForm=Fin" (VERBO | Modo=Indicativo | Número = Plural | Pessoa= 3 | Tempo= Futuro | Forma do verbo = Finito).

Diante dessa descrição, uma pessoa interessada no entendimento da morfologia verbal pode sensatamente questionar: Que segmentos das palavras expressam essas noções gramaticais? O que faz os segmentos encontrados indicarem essas noções e não outras, sendo, portanto, considerados alomorfes? Como os referidos segmentos constituem uma "palavra" do português e não uma sequência aleatória de caracteres? Entre outras questões.

O analisador mórfico-morfêmico é para superar as limitações, indicadas por tais questões, no campo de PLN. Obviamente, sabe-se que os modelos de linguagem existem para promover generalização dos dados linguísticos; o que justifica sua forma de descrever as estruturas da língua. Porém, para o ensino da morfologia verbal, as representações que eles fornecem não são suficientes.

Em contrapartida, como evidenciado nas Figuras 4, 5 e 6, o analisador mórfico-morfêmico proposto realiza o tipo de descrição morfológica previsto nos manuais de morfologia portuguesa. Essa descrição se apoia na concepção de língua como estrutura, tal que cada verbo flexionado contém uma estrutura subjacente que a análise é capaz de revelar. Além disso, essa concepção assume que as unidades mínimas significativas da língua, os morfemas, sendo de natureza essencialmente formal, não se confundem necessariamente com a substância, fônica ou ortográfica, que os expressam na palavra, os alomorfes.

Essa visão, em princípio desnecessária para as pesquisas em PLN, as quais são voltadas para o mercado, é fundamental para as pessoas que precisam conhecer as propriedades formais do objeto que elas vão ensinar. Por isso, o analisador mórfico-morfêmico é mais apropriado para as tarefas de ensino/aprendizagem da morfologia dos verbos regulares do português.

Um ponto a se destacar no analisador apresentado é que a visualização da estrutura dos verbos (Figura 6) permite fazer uma reflexão sobre as causas da variação indicadas pelos alomorfes, os chamados processos morfofonêmicos. Uma descrição completa desses processos para verbos pode ser encontrada em Laroca [8, p. 55-56]. A seguir, tratar-se-á unicamente do caso mostrado no diagrama, a saber, "(eu) cantarei", em resposta à questão de por que o morfema "{fut.pres.}" se instancia como "-re-" e não como "-ra-", como em "(tu) cantarás".

A resposta é que o morfema em consideração sofre harmonização vocálica, que é um processo fonético pelo qual uma vogal varia em direção à articulação de outra linguisticamente contígua [22] [23] [24]. É assim que o "a" (vogal central baixa) teria se transformado em "e" (vogal anterior médio-alta) por influência de "i" (vogal anterior alta).

Os resultados mostrados na Figura 6 ("Estrutura do verbo") obviamente não declaram que as variações observadas na flexão verbal são determinadas pelo processo de harmonização vocálica. Porém, uma vez que explicitam os distintos alomorfes para um mesmo morfema, eles têm o poder de levar os estudantes a inferir as razões de tal alomorfia.

Isso é possível porque, longe de ser aleatórias, as variações linguísticas são fenômenos regulares que se explicam em termos de mecanismos recorrentes e historicamente presentes no sistema linguístico [22]. O estudante, familiarizado com os processos de variação linguística do português, será apto então a fazer a inferência correta. Esse conhecimento será obtido durante as aulas de morfologia. Por isso, o analisador proposto não tem a ambição de ser senão uma ferramenta de apoio ao ensino da morfologia verbal.

7.1 Especificidade do analisador

Por fim, vale perguntar se os trabalhos comentados em "Trabalhos correlatos" podem satisfazer os requisitos da morfologia estrutural descritos na "Introdução". Se se considerar que eles não foram desenhados para esse fim, a resposta é logicamente negativa. Senão veja-se.

O analisador proposto por Alencar [16] realiza uma análise detalhada dos vocábulos gerados por derivação. Trata-se do protótipo de um componente morfológico que é capaz de realizar análises de vocábulos derivados por sufixação de "-ismo," "-iano", "-ês" e "-mente"; bem como de derivados por prefixação com elementos de origem grega ou latina do tipo de "neo"-, "pseudo"-, "semi-", "anti-", "pós-" ou "sub". Essa análise aplica as regras de geração de palavras que encontramos em estudos sobre derivação lexical.

Entre os muitos aspectos relevantes do trabalho ora mencionado, está a riqueza dos detalhes analíticos. Inclusive, ele oferece um esquema arbóreo da estrutura da derivação lexical; o que tem a ver com o diagrama retornada pelo analisador mórfico-morfêmico. Apesar disso, são analisadores bem diferentes, no que o transdutor de Alencar [10] implementa análises de nomes derivados enquanto a aplicação aqui desenvolvida realiza descrições de verbos flexionados.

A aplicação web proposta contrasta também com o analisador léxico-morfológico de Galhardi *et al.* [17] no que diz respeito ao objetivo e escopo do algoritmo. A ferramenta que os autores propuseram faz parte de um dicionário léxico cujas entradas recebem diferentes informações linguísticas, entre as quais informações morfológicas como POS-tag, raiz, sufixo, morfemas e o final que indica as inflexões da palavra como número, sexo, pessoa e tempo. Ou seja, a ferramenta inclui mais dados linguísticos para análise que a aplicação web aqui descrita o faz; e extrai esses dados do Banco de Redações UOL. Assim, o analisador de Galhardi *et al.* [17] pode ser complementado pelo analisador mórfico-morfêmico.

Finalmente, o analisador mórfico-morfêmico cumpre *grosso modo* as tarefas implementadas pelo analisador morfológico proposto por Vasilévski e Araújo [18]. Ambas as ferramentas promovem a análise morfológica automática dos verbos conjugados em seus constituintes mínimos significativos, os morfemas. Inclusive, a ferramenta proposta por esses autores realiza *mutatis mutandis* as operações mórfico-morfêmicas expostas na seção "Análise mórfico-morfêmica", mas as semelhanças não vão além disso.

Enquanto o programa deles processa um *corpus* com verbos conjugados em situações de uso, encontrados em enunciados reais; a aplicação web ora descrita processa dados de entrada descontextualizados: verbos conjugados nas seis pessoas e de acordo com a norma culta.

Em síntese, até foi possível demonstrar, no tocante à descrição estrutural dos verbos, o analisador mórfico-morfêmico é mais especializado do que os analisadores oferecidos por trabalhos correlatos. Assim sendo, ele deve ser considerado como uma ferramenta mais apropriada para o ensino/aprendizagem da morfologia estrutural dos verbos regulares do português.

8 Considerações Finais

Evidenciou-se como uma aplicação web para análise automática de verbos regulares pode ser uma ferramenta de apoio ao ensino/aprendizagem da morfologia verbal. O objetivo com a proposição desta aplicação não foi substituir um programa de ensino de morfologia em nível acadêmico, mas sim, de maneira modesta, oferecer um meio de os estudantes confrontarem e avaliarem suas análises linguísticas de maneira dinâmica e individual.

Contudo, esta aplicação também está sujeita a algumas limitações. Veja-se algumas delas: (a) os verbos de entrada devem ser conjugados de acordo com a norma culta da ortografia portuguesa. Para lidar com isso, um botão de ajuda ("Consultar conjugação") é disponibilizado dentro do formulário (Figura 4). Esse botão redireciona o usuário a uma página web contendo a conjugação dos verbos nos tempos e modos solicitados. Além disso, conjugar os verbos é um exercício pedagógico positivo, pois ajuda na memorização dos paradigmas verbais e, por extensão, na aprendizagem da ortografia padrão;

(b) a aplicação precisa estar conectada a uma internet. Assim, a consulta à análise vai depender de conexão, tal que nem sempre ela vai estar disponível aos estudantes;

(c) o uso da ferramenta requer que os estudantes já tenham tido um contato inicial com a abordagem linguística da morfologia estrutural. Isso significa que ela é, como dito, voltada exclusivamente para o público acadêmico. Os botões de informação podem auxiliar na compreensão dos temas da morfologia ("Morfologia", "Morfema", "Alomorfe" etc.); mas eles não são para substituir as aulas presenciais da disciplina.

Assim sendo, melhorias podem ser implementadas à medida que os estudantes experimentarem o analisador.

REFERÊNCIAS

[1] Daniel Jurafsky and James H. Martin. 2014. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). New Jersey: Upper Saddle River.

[2] Helena M. Caseli, Maria das Graças V. Nunes e Adriano Pagano. O que é PLN. In Caseli, H.M.; Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln.

[3] Cláudia Freitas. Dataset e corpus. In Caseli, H.M.; Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln.

[4] James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

[5] spaCy. 2023. *Industrial-strength Natural Language Processing in Python.* Disponível em: https://spacy.io/

[6] Leonard Bloomfield. 2023 [1926]. *A set of postulates for the science of language*. Disponível em: https://pure.mpg.de/rest/items/item_2282987/component/file_2282986/content

[7] Joaquim M. Câmara Júnior. 2001. *Estrutura da língua portuguesa* (34ª ed.). Rio de Janeiro: Vozes.

[8] Maria N. de C. Laroca. 2003. *Manual de morfologia do português* (3ª ed.). Pontes.

[9] Leonard Bloomfield. 1984[1933]. *Language*. Chicago: University of Chicago Press.

[10] John Lyons. 1987. Gramática, linguagem e linguística: uma introdução. In *Linguagem e Linguística*. Rio de Janeiro: LTC AS. 75-101.

[11] Peter H. Matthews. 1974. *Morphology. An introduction to the theory of word-structure*. Cambridge: Cambridge University Press.

[12] Aline Paes, Daniela Vianna e Jessica Rodrigues. Modelos de linguagem. In Caseli, H.M.; Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln.

[13] Helena Caseli, Cláudia Feitas e Roberta Viola. 2022. Processamento de linguagem natural. *Short courses of the 37th Brasilian Symposium on Data Bases*. Búzios, 2022.

[14] Maria J. B. Finatto, Helena M. Caseli, Luceli Lopes e Amanda Rassi. 2023. Sequência de caracteres e palavras. Modelos de linguagem. In Caseli, H.M.; Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln.

[15] Daniela O. F. do Amaral. *O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa*. Faculdade de Informática da PUCRS, 2013. Dissertação de Mestrado.

[16] Leonel F. de Alencar. 2009. Produtividade morfológica e tecnologia do texto: aspectos da construção de um transdutor lexical do português capaz de analisar neologismos. *Calidoscópio* 7, 3 (Set/Dez. 2009), 199-220. https://doi.org.10.4013/cld.2009.73.04

[17] Lucas B. Galhardi, Cinthyan R. S. C. de Barbosa, João Coelho Neto e Jacques D. Brancher. 2018. Analisador léxico-morfológico de redações de estudantes no estilo do ENEM. *Nuevas Ideas en Informática Educativa*. In *Memorias del Congreso Internacional de Informática Educativa (TISE'18)*, Brasília, 509-513.

[18] Vera Vasilévski e Márcio J. Araújo. 2011. Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português. *LinguaMÁTICA*. 3, 2 (Dez. 2011), 107-118.

[19] Shalabh Aggarwal. 2014. *Flask framework cookbook*. Birmingham-Mumbai. Packt Publishing LTD.

[20] Devndra Ghimire. 2020. *Comparative study on Python wed frameworks: Flask and Django.* Metropolia: University of Applied Sciences. Monografia de bacharelado em Engenharia.

[21] Wagner F. Lima; Cinthyan R. S. C. de Barbosa. Analisador mórfico-morfêmico dos verbos regulares do português. 2023. In *Actas da Conferência Internacional sobre Informática na Educação (TISE'23)*. Natal, 46-54.

[22] Fernando Tarallo. 1986. *A pesquisa sócio-lingüística* (2ª ed.). São Paulo: Ática.

[23] Maria C. D. de Castro e Maria S. de Aguiar. 2008. Reflexões de aspectos morfofonêmicos das vogais do português. *Pesquisa em foco*. 16, 1 (Nov. 2008): 52-61. https://doi.org/10.18817/pef.v16i1.11

[24] Thais C. Silva. 2002. *Fonética e fonologia do português* (6ª ed.). São Paulo: Contexto.