ÁRVORE DE CLASSIFICAÇÕES PARA O ESTUDO DE CORRELAÇÃO ENTRE VARIÁVEIS EM DADOS DE USO LINGUÍSTICO: CONTRIBUIÇÕES DO *SOFTWARE ORANGE DATA MINING*

CLASSIFICATION TREE FOR THE STUDY OF CORRELATION BETWEEN VARIABLES IN LINGUISTIC USAGE DATA: CONSTRIBUTIONS OF ORANGE DATA MINING SOFTWARE

Monclar Guimarães Lopes (UFF)[1]

Resumo: Há uma prerrogativa básica tanto na perspectiva sociolinguística quanto na funcionalista de que os fenômenos em variação e mudança não ocorrem aleatoriamente, mas, sim, motivados ou condicionados por fatores de natureza social, estrutural ou cognitiva. Neste artigo, temos o objetivo de mostrar as contribuições potenciais do *widget classification tree* (árvore de classificações), disponível no *Software Orange Data Mining.* A ferramenta, que recorre a cálculos de regressão logística, serve como método preditivo para a descrição hierárquica e gráfica de como as variáveis preditoras condicionam potencialmente as variáveis de desfecho. Esse recurso estatístico – motivado pelo modelo de *árvore de regressão por inferência condicional* (cf. Hothorn; Hornik; Zeileis, 2006; Speybroeck, 2012) – apresenta uma interface mais amigável ao usuário no *Orange Data Mining* em comparação à interface do Programa *R.* Como ilustração da contribuição desse recurso para pesquisas em língua em uso, descrevemos o impacto de três variáveis preditoras na instanciação de duas construções linguísticas, ambas formadas pela mesma sequência de elementos – preposição *sem* + verbo *dicendi* –, a saber: a oração hipotática adverbial modal ou condicional negativa (*e.g.:* ele saiu *sem falar* com ninguém); o marcador estruturante do discurso de acréscimo (*e.g.:* ele faltou hoje. *Sem falar que,* quando vem, sempre chega atrasado).
Palavras-chave: Árvore de classificações. Variáveis preditoras e de desfecho. Língua em uso.

Abstract: There is a basic prerogative from both the Sociolinguistic and Functionalist perspectives that phenomena undergoing variation and change do not occur randomly, but rather are motivated or conditioned by factors of a social, structural or cognitive nature. In this paper, we aim to show the potential contributions of the *classification tree widget*, available in the Orange Data Mining Software. The tool serves as a predictive method and, to this end, uses logistic regression calculations for the hierarchical description of how predictor variables potentially condition outcome variables. This statistical resource – also known by the term *conditional inference regression tree* (cf. Hothorn; Hornik; Zeileis, 2006; Speybroeck, 2012) – presents a user-friendly interface

[1]Professor de Língua Portuguesa na Universidade Federal Fluminense (UFF). E-mail: monclarlopes@id.uff.br

in Orange Data Mining, compared to the R language. As an illustration of the contribution of this tool to research in language usage, we describe the impact of three predictor variables on the instatiation of two linguistic constructions, both formed by the same sequence of elements – preposition *sem* + verb *dicenci* –, namely: the hypotactic adverbial negative modal or conditional clause (*e.g.:* ele saiu *sem falar* com ninguém); the discourse structuring marker of addition (*e.g.:* ele faltou hoje. *Sem falar que,* quando vem, sempre chega atrasado).
Keywords: Classification tree. Predictor and outcome variables. Language Usage.

Introdução

Por apresentarem natureza dinâmica e heterogênea, são inerentes às línguas naturais a variação e a mudança, fenômenos condicionados por fatores de diferentes ordens, sejam sociais, estruturais e/ou cognitivos. Às perspectivas linguísticas que buscam descrever as línguas sob a ótica de sua variação e mudança – como a Sociolinguística e o Funcionalismo, por exemplo –, cabe não apenas detectar o fenômeno variável em si – isto é, as variáveis de desfecho (ou dependentes) –, mas também identificar as motivações subjacentes a esses processos – isto é, suas variáveis preditoras (ou independentes).

Na literatura linguística, a descrição dos fatores que motivam o fenômeno da variação ou da mudança tem sido feita com base em métodos qualiquantitativos de análise. Sob essa ótica, a face qualitativa se dá, entre outros aspectos, pela identificação e interpretação das variáveis preditoras presentes em cada ocorrência; a quantitativa, pela verificação da extensibilidade desses fatores às demais ocorrências do *corpus* em análise. No que tange a essa última face, a quantificação, em algumas abordagens, como a Funcionalista, por exemplo, tem sido mais frequentemente interpretada por meio da descrição de frequência absoluta (quantidade de vezes que o mesmo fator condiciona a escolha por uma variável de desfecho no total de ocorrências) ou relativa (quantidade de vezes que esse fator condiciona a escolha por uma variável de desfecho em termos percentuais).

À descrição de frequência, podem-se juntar métodos estatísticos ou computacionais mais robustos, que possibilitam ao pesquisador identificar a força de associação entre variáveis preditoras e de desfecho ou, ainda, realizar previsões para além do *corpus*. A depender da quantidade de dados analisados e do método estatístico empregado, pode-se interpretar que estamos diante de fatores generalizáveis, de modo

que encontraremos resultados semelhantes se extrapolarmos a amostra previamente analisada.

A extensibilidade dos dados para além da amostra pode ser tratada por meio de regressão logística, um método estatístico que estima a probabilidade de ocorrência de uma variável de desfecho com base em um determinado conjunto de dados de variáveis preditoras. Quando temos fatores múltiplos, a regressão logística também possibilita, desde que atendidas as condições necessárias quanto à padronização das variáveis preditoras, avaliarmos o grau de associação entre as variáveis, em termos hierárquicos, isto é, identificar qual(is) variável(is) preditora(s) impacta(m) mais fortemente a seleção de uma ou mais variáveis de desfecho.

Tradicionalmente, a regressão logística é feita por meio de medidas estatísticas, que exigem, por parte do pesquisador e do leitor, o domínio de conceitos matemáticos como *distribuição qui-quadrado, medidas de* R^2*, graus de liberdade* etc. Por sua vez, Hothorn, Hornik e Zeileis (2006), bem como Speybroeck (2012), desenvolveram um modelo gráfico de regressão logística intitulado *árvore de regressão por inferência condicional*, de visualização e interpretação mais simples para quem tem pouco ou nenhum conhecimento estatístico[2]. Originalmente, esse gráfico é gerado no Programa *R,* em pacotes como o *partykit*, o que exige do pesquisador o domínio da linguagem de programação da ferramenta. Mais recentemente, no entanto, o recurso foi incorporado ao *Software Orange Data Mining* (ODM)*,* por meio do *widget Classification tree* (árvore de classificações). Como veremos, o ODM apresenta uma interface mais amigável e intuitiva a usuários que têm conhecimentos básicos de informática e pouco ou nenhum conhecimento de linguagem R.

O objetivo deste artigo é mostrar a contribuição potencial do recurso *árvore de classificações* do ODM para a análise da correlação entre variáveis preditoras e variáveis de desfecho na análise de dados linguísticos. Como ilustração didática, recorremos à descrição de duas construções formadas pela sequência de palavras *sem* + verbo *dicendi*, como *sem falar, sem contar, sem dizer* etc. Como veremos mais adiante, essa sequência pode ser recrutada tanto por uma oração hipotática adverbial modal ou condicional negativa – e.g.: ele saiu *sem falar* com ninguém – quanto por um marcador

[2] Cabe esclarecer que o analista que trabalha com métodos quantitativos precisa ter um domínio mínimo de estatística para a interpretação e a descrição de seus dados. Quando falamos, aqui, em "nenhum conhecimento estatístico", referimo-nos, sobretudo, ao leitor final da divulgação científica, que será capaz de compreender o grau de associação entre variáveis mesmo que não conheça as medidas estatísticas previstas por cada tipo de teste.

estruturante do discurso de acréscimo (cf. Traugott, 2022) – e.g.: ele faltou hoje. *Sem falar que*, quando vem, chega atrasado.

Da regressão logística binária à árvore de regressão por inferência condicional

Segundo Hosmer e Lemeshow (2000), a regressão logística binária é um modelo de análise estatística que tem como objetivo prever, a partir de um conjunto de variáveis preditoras explicativas de natureza contínua e/ou binária, o desfecho de uma variável categórica nominal dicotômica. Mais especificamente, assim como os outros modelos de regressão, busca determinar o coeficiente de correlação entre uma variável de desfecho (ou dependente) e uma ou mais variáveis preditoras (ou independentes).

Trata-se de um modelo aplicável à Linguística, mas que exige, por parte do analista e do leitor, o domínio de alguns conceitos e algumas medidas estatísticas. Como ilustração, apresentamos o Quadro 1, que representa um conjunto de tabelas geradas no *software JASP*, que buscam identificar a existência de correlação entre três variáveis preditoras – (1) anteposição do pronome demonstrativo *isso;* (2) posposição de conjunção *que*; (3) posição supraoracional – e a instanciação ou não de uma variável de desfecho: o marcador estruturante do discurso de acréscimo (MED) formado por *sem* + verbo *dicendi* (*sem falar, sem contar, sem dizer* etc.) no lugar da oração hipotática adverbial modal ou condicional negativa (OHA), formada por esses mesmos elementos[3].

[3] Cabe frisar que os dados estatísticos apresentados no Quadro 1, neste momento do texto, tem como objetivo tão-somente ilustrar o raciocínio empregado por parte da interpretação desse teste estatístico, e não de servir como ilustração de resultados de pesquisa. O fenômeno será descrito posteriormente, na seção em que ilustramos a aplicação da ferramenta.

Quadro 1 - Conjunto de tabelas com dados da regressão logística binária realizada no JASP:

Model Summary – CLASSE

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	5739.683	5741.683	5748.132	4670						
H_1	3270.428	3278.428	3304.224	4667	2469.256	< .001	0.430	0.580	0.471	0.411

Coefficients ▼

					Wald Test			95% Confidence interval (odds ratio scale)	
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p	Lower bound	Upper bound
(Intercept)	−1.471	0.075	0.230	−19.550	382.203	1	< .001	0.198	0.266
PRONOME	3.535	0.136	34.286	25.918	671.752	1	< .001	26.244	44.792
CONJUNÇÃO	1.455	0.093	4.285	15.565	242.270	1	< .001	3.568	5.147
SUPRAORACIONAL	3.062	0.121	21.371	25.258	637.949	1	< .001	16.851	27.103

Note. CLASSE level 'MED' coded as class 1.

Performance Diagnostics

Performance metrics

	Value
Accuracy	0.830

Fonte: Elaboração própria.

No quadro acima, algumas medidas são relevantes para a intepretação da existência ou não de correlação entre as três variáveis preditoras – *pronome, conjunção* e *supraoracional* – e a variável de desfecho, que é a classe *MED*:

a) Na primeira tabela, observamos os valores relativos à hipótese alternativa (H_1), que revelam que a medida obtida por qui-quadrado (X^2), 2469.256, é estatisticamente significativa ($p < 0{,}001$). A medida de Nagelkerke R^2 evidencia, em termos percentuais – 0,58 (58%) – o quanto o modelo com as três variáveis preditoras inclusas explica a variável de desfecho em comparação à hipótese nula (H_0);

b) Na segunda tabela, a coluna *odds ratio* (razão de chances) indica a magnitude das chances, em número de vezes, de a variável condicionar a instanciação de MED. No caso, temos: anteposição por pronome demonstrativo *isso* (34,3 vezes mais em relação à referência), posição supraoracional (21,3 vezes a mais em relação à referência), posposição de conjunção *que* (4,2 vezes a mais em relação à referência). A precisão dessas medidas é atestada pelo teste de *Wald*, cujo valor de *p* foi considerado estatisticamente significativo para as três variáveis ($p < 0{,}001$)[4].

[4] O valor da *Odds Ratio* (Razão das Chances) representa a transformação da escala logarítmica (log-odds) presente na coluna *Estimate*.

c) Na tabela 3, temos a precisão de todo o modelo, medido em 83% de acurácia. Isso indica que, se aplicarmos as mesmas variáveis em outras amostras, devemos encontrar, na maior parte das vezes, um resultado equivalente ou aproximado.

Por sua vez, a árvore de regressão por inferência condicional é uma abordagem estatística exploratória que emprega testes de significância a partir da inter-relação de um conjunto de variáveis, assim como a regressão logística binária. Não obstante, recebe esse nome porque é organizado em estruturas arbóreas de classificação e regressão, com nódulos e ramos (semelhantes às folhas e aos caules de uma árvore). O modelo identifica o impacto das variáveis preditoras, hierarquizando-as, por meio de uma sequência de decisões: da variável mais relevante à menos relevante, mostrando de que modo a convergência entre elas caminham para um desfecho.

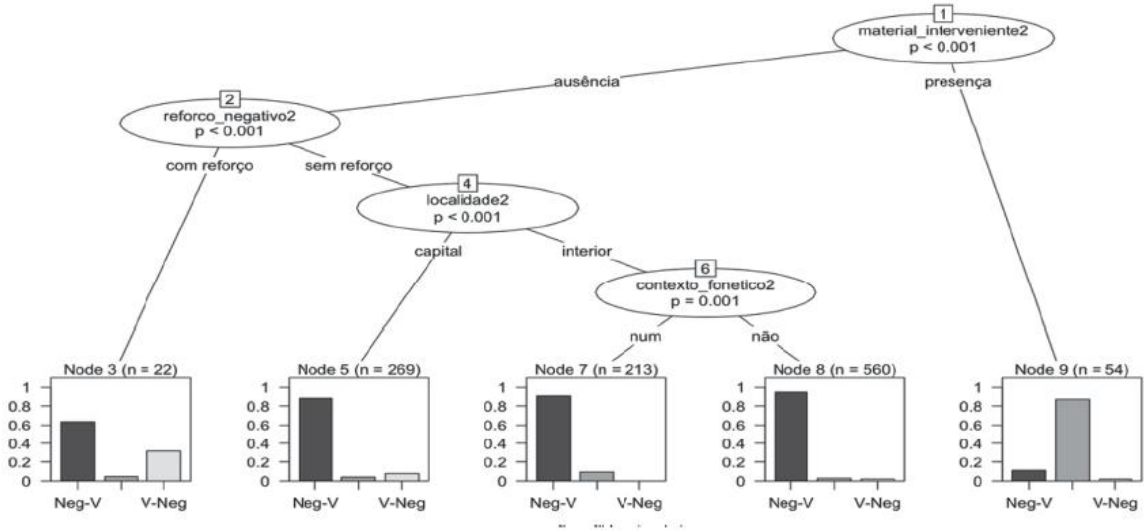
O modelo foi desenvolvido por Hothorn, Hornik e Zeileis (2006) e também por Speybroeck (2012) e tem aplicação em diversas áreas do conhecimento. Para ilustrar a aplicação desse método em dados linguísticos, partimos de Freitag e Pinheiro (2020), que o exemplificam a partir de uma série de estudos sobre a negação no português (cf. Furtado da Cunha, 2001; Santana; Nascimento, 2011; Yakovenco; Nascimento, 2016). Em resumo, no português brasileiro, a negação ocorre por meio de três variantes: neg-V ~ neg-V-neg ~ V-neg. Abaixo, segue uma ocorrência de cada uma delas:

> (a) Negação canônica (neg-V)
> (falando sobre amizade) “em geral... eu não tenho problema com ninguém... eu procuro sempre ser sociável.” (cam_fem_22).
>
> (b) Negação dupla (neg-V-neg)
> (falando de violência no bairro) “lá tem muito usuário [de droga] mas sabe se controlar... não é desse tipo de sair roubando não”. (ema_mas_23)
>
> (c) Negação final (neg-V)
> doc: seus amigos pensam em mudar de escola? (sobre educação) (ali_mas_17): “pensam não...”
>
> (Freitag e Pinheiro, 2020, p. 330)

Para a análise do fenômeno, os autores fazem uso do pacote *partykit*, disponível no Programa R. As análises apontam a presença de quatro variáveis preditoras: a) presença de material interveniente; b) presença ou ausência de material de reforço negativo; c) variedade rural ou urbana; d) contexto fonético favorável (presença da

forma átona *não* – isto é, *num* no cotexto prévio). Tais variáveis têm pesos diferentes na seleção das variantes, aspectos que estão representados na árvore abaixo:

Gráfico 1 - Regressão em função da estrutura morfossintática da negação:



Fonte: Freitag e Pinheiro, 2020, p. 330.

O Gráfico 1 apresenta a interação entre as variáveis por meio de uma sequência de seis decisões tomadas pelo falante – não necessariamente de forma consciente. O fator preponderante, identificado em (1), é a presença ou não de material interveniente. Se positivo, os falantes preferirão majoritariamente a variante neg-V-neg; se negativo, haverá um conjunto de outras decisões em sequência: (2) há presença de material de reforço negativo (*num, nada* etc.)? Se positivo, preferência por Neg-V; se negativo, há mais uma condicionante: (4) o falante é da variedade urbana ou rural? Se urbana, predileção por Neg-V; se rural, ainda haverá uma última decisão: (6) presença de contexto fonético favorável. Nos dois contextos possíveis, há predileção por Neg-V, mas, no caso, de existência de contexto fonético favorável, esse uso ainda concorre com a variante neg-V-neg.

Cabe mencionar que os resultados apresentados no Gráfico 1 são gerados a partir de cálculos de regressão logística semelhantes àqueles apresentados no Quadro 1. A diferença é que o recurso hierarquiza as variáveis em um modelo binário de entrada e saída, mostrando de que modo cada uma das variáveis preditoras impactam, sozinhas ou em relação com outras, uma determinada variável de desfecho.

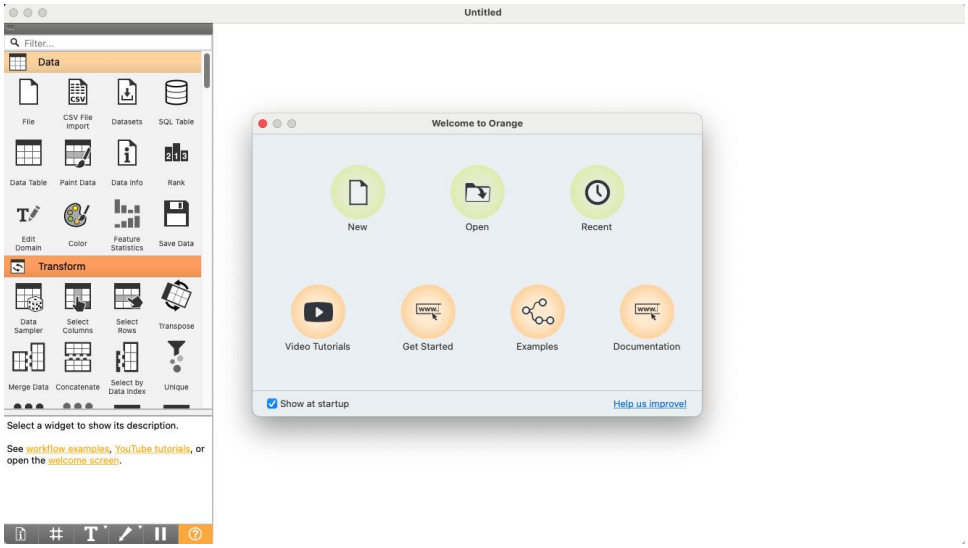
No programa R, onde esse gráfico foi gerado, os resultados são gerados algoritmicamente por sistemas estatísticos que correlacionam as variáveis informadas

em um *dataframe,* submetidos a um *script* do pacote *partykit*. A realização dessa operação, como é de se esperar, exige conhecimentos básicos de linguagem *R*. O que queremos mostrar, neste texto, é de que modo esse procedimento é mais simples e amigável no *Orange Data Mining*, podendo ser realizado por qualquer pessoa com conhecimentos básicos de informática e sem conhecimentos profundos de estatística.

O Orange Data Mining e o *Widget Classification Tree*

O Orange Data Mining (ODM), segundo sua própria plataforma, é um *software* de livre acesso que constitui uma caixa de ferramentas (*toolbox*) de mineração de dados, *machine learning* e visualização de dados. Diferentemente dos softwares estatísticos comuns (como JASP, SPSS) ou, ainda, a linguagem R, o ODM é construído sob uma interface visual e interativa, em um ambiente CANVAS[5].

Figura 1 - Tela inicial do Orange Data Mining:



Fonte: Orange Data Mining Software.

O programa é constituído por uma série de *widgets* que interagem entre si, a partir de uma base de dados. O sistema permite a realização de uma série de testes estatísticos, sobretudo aqueles que resultam em gráficos, como é o caso da árvore de

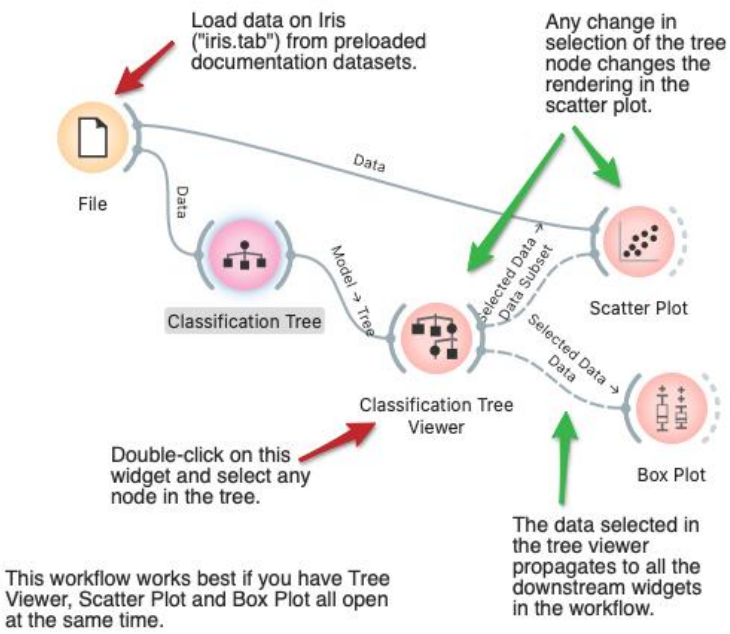
[5] CANVAS é um modelo de representação visual. No ODM, o CANVAS é representado por uma tela em branco em que o usuário seleciona um ou mais *widgets* e estabelece uma série de relações entre eles, de acordo com seu objetivo de pesquisa.

regressão por inferência condicional, ou simplesmente, árvore de classificações no ODM.

A geração de uma árvore de classificações, dentro do ODM, é relativamente simples. Eis o passo a passo:

a) Na tela de abertura (cf. *Figura 1*), o usuário deve clicar na opção *Examples* e, em sequência, na opção *Classification Tree.* Como resultado, o programa abrirá automaticamente uma série de *Widgets* interconectados, conforme indicado na Figura 2.

Figura 2 - Classification Tree – Modelo:



Fonte: Orange Data Mining Software.

b) Em sequência, o usuário deve clicar sobre o *Widget* File e carregar seu arquivo de dados anotados. Para a interpretação do sistema, é imprescindível que haja uma única variável nominal categórica (que será sua variável de desfecho). As variáveis preditoras, nesse caso, devem ser lidas como fatores. No exemplo em tela, estamos lidando com fatores binários, representados por 0 e 1. Veja a Tabela 1, como ilustração desse procedimento:

Tabela 1. Dados anotados: variável de desfecho e variáveis preditoras:

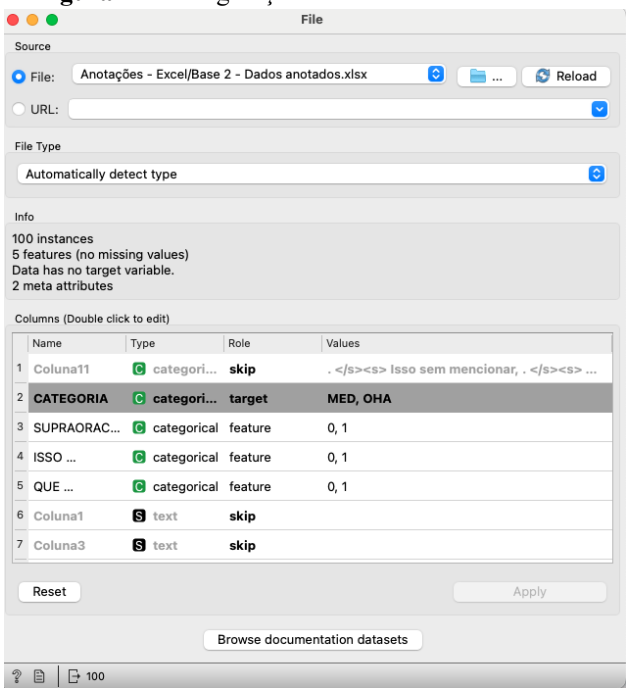
	CATEGORIA	SUPRAORACIONAL	ISSO ANTEPOSTO	QUE POSPOSTO
contém ainda o	OHA	0	0	0
	MED	1	0	0
artir da	MED	0	0	0
ıs, Médiuns,	MED	0	1	1
essoas de bem,	OHA	0	0	0
e dê a impressão de ser uma	OHA	0	0	0
SARS-Cov-1,	MED	1	1	0
uage Acquisition and Language	OHA	0	0	0
ando	MED	0	0	0
pelos	MED	1	1	0
Se tudo isso	MED	0	1	1
	MED	0	0	1

Fonte: Elaboração própria.

c) A tabela indica, na coluna "categoria", a variável de desfecho, que será representada como a única variável categórica nominal na coluna. As colunas das categorias *supraoracional, isso anteposto* e *que posposto*, que são as variáveis preditoras, são representadas de forma numérica, sendo 0 a representação do "não"; 1 do "sim".

d)

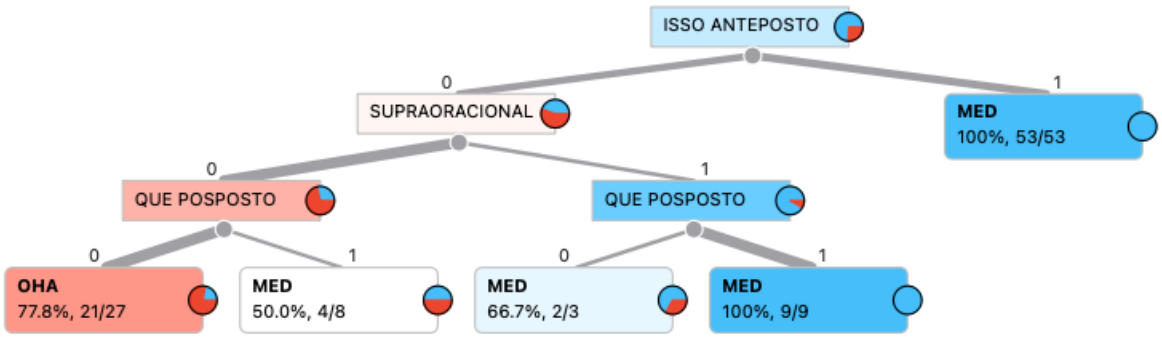
Figura 4 - Configuração da Tabela dentro do ODM:



Fonte: Orange Data Mining Software.

e) Assim que carregar o arquivo para dentro do CANVAS, o sistema pedirá que o usuário realize algumas configurações. A variável de desfecho deve ser classificada como *target* (alvo), na coluna *role*; as variáveis preditoras, por sua vez, como *feature* (propriedade); as demais informações, se houver, como ocorre na Tabela 1, serão desconsideradas e, por isso, indicamos o atributo *skip* nessa mesma coluna. Em seguida, o usuário deve clicar no botão *apply* e fechar a caixa. Por fim, deve clicar no *widget classification tree viewer,* no ambiente do CANVAS. Como resultado, visualizará um gráfico de estrutura arbórea semelhante a este:

Gráfico 2 - Árvore de regressão por inferência condicional para as construções de OHA e MED[6]:



Fonte: Elaboração própria.

Exemplo de aplicação

Nesta última seção, temos o objetivo de ilustrar, por meio de análise empírica de dados de uso linguístico, de que modo uma ou mais variáveis preditoras podem impactar a instanciação de uma variável de desfecho. Para isso, adotamos o modelo de árvore de classificações do ODM, baseado no modelo de árvore de regressão por inferência condicional (cf. Hothorn; Hornik; Zeileis, 2006; Speybroeck, 2012), na medida em que ele permite prever de que modo as variáveis preditoras potencialmente ajudam a prever uma ou mais variáveis de desfecho a partir de um conjunto de decisões, hierarquicamente estruturadas.

[6] O Gráfico 2 será explorado e descrito na seção subsequente.

Caracterização do fenômeno

Num primeiro momento, faz-se imprescindível caracterizar nosso objeto de pesquisa, que é uma construção formada pela preposição *sem*, seguida de verbo *dicendi.* A esse respeito, cabe mencionar que, segundo a literatura gramatical e linguística, *sem* apresenta o sentido básico de subtração, ausência ou desacompanhamento (Cunha e Cintra, 2001; Rocha Lima, 1972), que é mobilizado, na relação hipotática adverbial, para a introdução da noção de modo ou condição negativa, conforme indica Neves (2018, p. 791):

> a) Modais:
> (i) Será que está me seguindo há algum tempo, *sem* eu perceber?
> (ii) Lorenzo tinha deixado o fundo da sala, *sem que* eu percebesse.
>
> b) Condicionais:
> (iii) Por isso eu não poderia ir embora *sem* dizer a você, enquanto é tempo, alguma coisa sem sentidos ocultos.
> (iv) "*Sem que* se produzam fatos que comprometam a democracia, não podemos deixar que os trabalhadores sem-terra sejam donos do país", afirmou o senador.

Além desses usos, já classificados e disseminados pela literatura gramatical e linguística, detectamos um outro no português contemporâneo, em que *sem,* justaposto a um verbo *dicendi*, como *contar, falar, dizer, mencionar* etc., recategoriza-se como um conector de acréscimo – aqui classificado como um *marcador estruturante do discurso* – ou MED – (cf. Traugott, 2022). Como ilustração, segue duas ocorrências desse uso:

> (01) A psicóloga e psicoterapeuta Olga Tessari aponta que o problema da mania de limpeza está na pessoa, pois, por mais que se busque a perfeição, "jamais chegaremos a ela, pois a perfeição é uma ilusão". No máximo, o que ela irá conseguir é ficar ainda mais nervosa, pois, para manter a casa arrumada e limpa sempre, ninguém haveria de morar nela, *sem contar* o problema de a poeira tomar conta após algum tempo sem ninguém para tirar o pó.

> (02) Hoje vamos falar de uma casta muito famosa e bem conhecida no mundo todo, especialmente, por fazer parte de um corte (blend/assemblage) de um vinho icônico e poderoso de nome Châteauneuf-du-Pape. Mas o interessante, que a origem desta uva não é a França, mas sim a Espanha. Na França a chamam de Mourvedre e na Espanha de Monastrell. Esta uva é cultivada, principalmente, na costa mediterrânica da Espanha, cujo clima é o maior aliado desta variedade. É nesta região que também ficam as principais praias espanholas e é um dos roteiros turísticos mais lindos da Europa. *Sem falar* na gastronomia, que é um show à parte!

Em (01), *sem contar* pertence a uma categoria funcional distinta das apresentadas por Neves (2018) no grupo de ocorrências (a) e (b). A despeito de suas feições hipotáticas – uma preposição seguida de verbo no infinitivo, estrutura típica de orações adverbiais reduzidas –, a construção não escopa um verbo (ou algum elemento) presente em uma oração matriz, atribuindo-lhe uma noção circunstancial de modo ou de condição. Na verdade, argumentamos que *sem contar* não faz referência a nenhum elemento particular do co-texto, diferentemente do que se deveria esperar de uma estrutura prototipicamente hipotática.

A noção de acréscimo atribuída a *sem contar,* em (01), é construída discursivamente, na medida em que o enunciador enumera dois fatos convergentes, ligados por essa construção, que levariam uma pessoa maníaca por limpeza a um estado de nervosismo: (1) manter a casa limpa e arrumada sempre exigiria que ninguém morasse nela; (2) mesmo sem ninguém morando, a poeira naturalmente tomaria conta da casa depois de algum tempo. A plausibilidade de conferir a *sem contar* uma noção de acréscimo pode ser atestada por meio de um teste de substituição, no qual é possível alternar *sem contar* por uma construção canônica de valor de acréscimo em que as condições de verdade sejam mantidas. Como evidência, em (01'), propomos a substituição de *sem contar* por *além de*[7]:

> (01') No máximo, o que ela iria conseguir é ficar ainda mais nervosa, pois, para manter a casa arrumada e limpa sempre, ninguém haveria de morar nela, *além do* problema da poeira tomar conta após algum tempo sem ninguém para tirar o pó.

Defendemos que a ideia de acréscimo construída discursivamente é o resultado da ação da intersubjetividade mobilizada na cena comunicativa (cf. Traugott e Dasher, 2002). Haja vista que o falante *conta* o que efetivamente não iria *contar* ao enunciar a frase "sem contar o problema de a poeira tomar conta após algum tempo sem ninguém para tirar o pó" –, o ouvinte é levado a inferir um novo significado para aquele uso, dado que as noções de ausência e negação, próprias da semântica do *sem*, tornaram-se opacas. No lugar, atribui-se à construção um valor de acréscimo, que equivaleria a dizer [D1 e ainda D2]. Sob essa ótica, a replicação desses contextos e desse mesmo tipo de inferência teria resultado em um novo uso convencionalizado, cuja frequência de

[7] A despeito da manutenção das condições de verdade no teste de substituição proposto, há diferenças discursivo-pragmáticas entre *sem contar* e *além de*. Esse aspecto, no entanto, foge ao escopo deste artigo.

ocorrência pode ser atestada em uma comunidade linguística de falantes, como acontece no caso de [D1 sem contar D2].

Em (02), temos a demonstração da produtividade do esquema [D1 sem $V_{dicendi}$ D2], na medida em que outros verbos de natureza *dicendi* passam a instanciá-lo, como *falar*, por exemplo. Como é possível notar na respectiva ocorrência, o falante tece elogios à costa mediterrânica da Espanha, categorizando-a como: (1) uma das regiões com praias mais lindas da Espanha; (2) um dos roteiros mais lindos da Europa; (3) um ótimo lugar para ter experiências gastronômicas. No caso em tela, *sem falar* é empregado especificamente para a introdução do último argumento (01), ao qual podemos propor um teste de substituição análogo ao que realizamos em (01'):

> (02') É nesta região que também ficam as principais praias espanholas e é um dos roteiros turísticos mais lindos da Europa. *Além da* gastronomia, que é um show à parte![8]

Um dos aspectos que nos leva também a argumentar favoravelmente à ideia de que MEDs como *sem contar, sem falar* se distanciam dos usos hipotáticos prototípicos é a frequência com que ele ocorre em segmentos supraoracionais, como podemos verificar em (02), em que D1 e D2 encontram-se em períodos distintos. Em nossa pesquisa, por exemplo, identificamos que a posição supraoracional é muito produtiva para MED, mas não para OHA. Em 400 ocorrências da sequência *sem falar* em posição supraoracional, observou-se que todas elas eram classificadas como MED, e não como OHA.

Hipóteses e Metodologia

Durante análises pretéritas das sequências *sem* + verbo *dicendi,* observamos que era muito frequente a instanciação de MED em contextos linguísticos específicos: a) quando havia anteposição de *isso* à construção; b) quando havia posposição de conjunção *que*; c) quando a construção ocorria em posição supraoracional.

Segundo nossas hipóteses, havia, portanto, três variáveis preditoras potenciais para a instanciação de MED: (1) No que tange ao pronome demonstrativo *isso*, que é recorrentemente um encapsulador de predicações, a construção de MED se vincularia a

[8] O teste proposto busca verificar se são mantidas as condições de verdade na substituição, ou seja, se há equivalência funcional entre *sem contar* e *além de* nesse contexto de uso. Não estamos, aqui, comprometidos com um critério normativo, que tende a avaliar negativamente a ausência de verbo no início da estrutura.

esse elemento, e não mais a um verbo em uma oração matriz, tal qual tende a ocorrer com as circunstâncias de modo ou condição. (2) A posição supraoracional, por sua vez, quebraria o vínculo sintático mais forte entre oração matriz e oração hipotática. Isso decorreria do fato de a construção evocar uma dimensão mais textual (de enumeração, introdução de novos argumentos) do que circunstanciação de um estado de coisas. (3) Por sua vez, a conjunção *que* atuaria como uma das subpartes do esquema $X_{\text{que conect}}$ (cf. Santos; Cezario, 2017; Santos; Silva; Cezario, 2019; Ely; Cezario, 2023), que tende a formar, via analogização, novos elementos procedurais no domínio da conexão no português.

Com o objetivo de verificar a plausibilidade dessas três hipóteses, recorremos ao *Corpus Portuguese Web 2020 (ptTenTen20)*, da plataforma *Sketch Engine*[9]. Trata-se de um *corpus* de 12 bilhões de palavras, com dados de diferentes variedades do português, coletados entre o período de junho e novembro de 2020. O *corpus* permite filtrar as ocorrências por diferentes aspectos, como variedade, gênero, assunto etc.

No que tange a esta pesquisa, restringimos nossas buscas à variedade brasileira do português presente no domínio **.br*, sem outras especificações, como as atinentes a assunto e gênero, por exemplo. Esse *subcorpus* apresenta aproximadamente 8 bilhões de palavras (mais especificamente, 8.010.603.604 *tokens*), dentre as quais 87.369.605 são verbos (.5,868,2 de *tokens* por milhão de palavras). Nossa preocupação inicial foi a de identificar, no *corpus*, os dez verbos *dicendi* que, justapostos à preposição *sem*, apresentavam maior quantidade de *tokens*. Isso nos levou aos dados apresentados na Tabela 2.

Tabela 2 - Frequência *token* da sequência *sem* + verbo *dicendi* no *corpus* ptTenTen20:

Sequência	Frequência Token	Token por milhão de palavras
1. Sem contar	99.091	6,66
2. Sem falar	81.953	5,50
3. Sem dizer	10.437	0,70
4. Sem mencionar	10.185	0,68
5. Sem considerar	13.305	0,89
6. Sem citar	9.484	0,68
7. Sem revelar	4.442	0,28
8. Sem especificar	3.354	0,23
9. Sem explicar	1.954	0,13
10. Sem comentar	1.082	0,07

Fonte: Elaboração própria.

[9] http://sketchengine.eu – Acesso em 08 mai. 2024.

Em virtude de nossas hipóteses, empregamos a estratégia de buscar 100 ocorrências de cada padrão abaixo para os 10 verbos selecionados. Assim sendo, traçamos o objetivo de selecionar 800 ocorrências por verbo, o que perfaria um total de 8.000 dados analisados. O levantamento dessas ocorrências no *corpus* foi realizado por meio da ferramenta *concordance*, jogando-se, no campo de busca, os contextos sintáticos descritos abaixo.

a) . isso sem ($V_{dicendi}$) que
b) . isso sem ($V_{dicendi}$)
c) . sem ($V_{dicendi}$) que
d) . sem ($V_{dicendi}$)
e) isso sem ($V_{dicendi}$) que
f) isso ($V_{dicendi}$)
g) sem ($V_{dicendi}$) que
h) sem ($V_{dicendi}$)

O objetivo dessa notação era o de restringir: (1) as buscas de *a* a *d* aos contextos supraoracionais; (2) as buscas de *e* a *h* aos contextos oracionais. Como é possível notar, os padrões se diferem entre si no que diz respeito às variáveis preditoras, isto é, anteposição ou não de pronome demonstrativo *isso;* posposição ou não de conjunção *que*, além da posição oracional ou supraoracional, como acabamos de mencionar. Cabe esclarecer que alguns padrões se mostraram improdutivos ou pouco produtivos para alguns verbos, conforme podemos observar na Tabela 3:

Tabela 3 - Distribuição das ocorrências no *corpus*, por padrão de uso:

PADRÕES	Contar	Falar	Dizer	Mencionar	Considerar	Citar	Revelar	Especificar	Explicar	Comentar	**Totais**
. Isso sem $V_{dicenci}$ que	100	100	47	100	52	05	0	0	0	09	413
. Isso sem $V_{dicenci}$	100	100	34	100	100	82	0	01	0	31	548
. Sem $V_{dicenci}$ que	100	100	100	100	83	18	11	04	01	11	528
. Sem $V_{dicenci}$	100	100	100	100	100	100	100	100	100	100	1000
isso sem $V_{dicenci}$ que	100	100	36	50	29	03	0	0	01	07	326
isso sem $V_{dicenci}$	100	100	43	100	100	61	13	02	07	28	554
sem $V_{dicenci}$ que	100	100	100	100	100	81	100	89	65	30	865
sem $V_{dicenci}$	100	100	100	100	100	100	100	100	96	100	1000
Totais	800	800	560	750	664	450	324	297	270	316	5.230

Fonte: Elaboração própria

Na Tabela 3, os valores inferiores a 100 representam a totalidade de ocorrências existentes no *corpus*. Conforme é possível observar, os padrões supraoracionais e oracionais iniciados pelo pronome demonstrativo *isso* são menos produtivos para alguns verbos e até improdutivos para outros, como é o caso de *revelar, especificar* e *explicar*.

Já os campos identificados com o número 100 representam os dados selecionados, tendo sido sorteados randomicamente pelo próprio sistema do *corpus*.

As análises da pesquisa foram realizadas por meio do emprego do método misto, que se caracteriza pelo "equacionamento entre a metodologia qualitativa e quantitativa" (Lacerda, 2016, p. 85). Para a análise proposta para este texto, especificamente, buscamos descrever tão-somente o impacto das três variáveis indicadas em nossas hipóteses – anteposição de pronome demonstrativo *isso,* posposição de conjunção *que* é posição supraoracional – na instanciação ou não de MED. Portanto, somente esses aspectos serão mobilizados na subseção dos resultados, haja vista que o objetivo deste artigo não é fazer a caracterização dos MEDs, mas, sim, ilustrar a aplicação do modelo de árvore de classificação e mostrar suas contribuições potenciais para a descrição da correlação entre variáveis preditoras e de desfecho na análise de dados de uso linguístico.

Resultados

No intuito de verificar, por meio do emprego do método de árvore de classificações, o eventual ou potencial impacto das três variáveis preditoras na instanciação de MED, criamos um *dataframe* no Excel com os dados quantitativos das construções de OHA e de MED, com a inclusão das três variáveis em forma de fatores (sendo 0 para ausência do atributo; 1 para sua presença). O *dataframe* foi constituído de 4.671 linhas (tendo sido excluídas 649 linhas relativas a ocorrências descartadas). Na Tabela 4, ilustramos como ele foi preenchido.

Tabela 4 - Distribuição das ocorrências de OHA e de MED por variável:

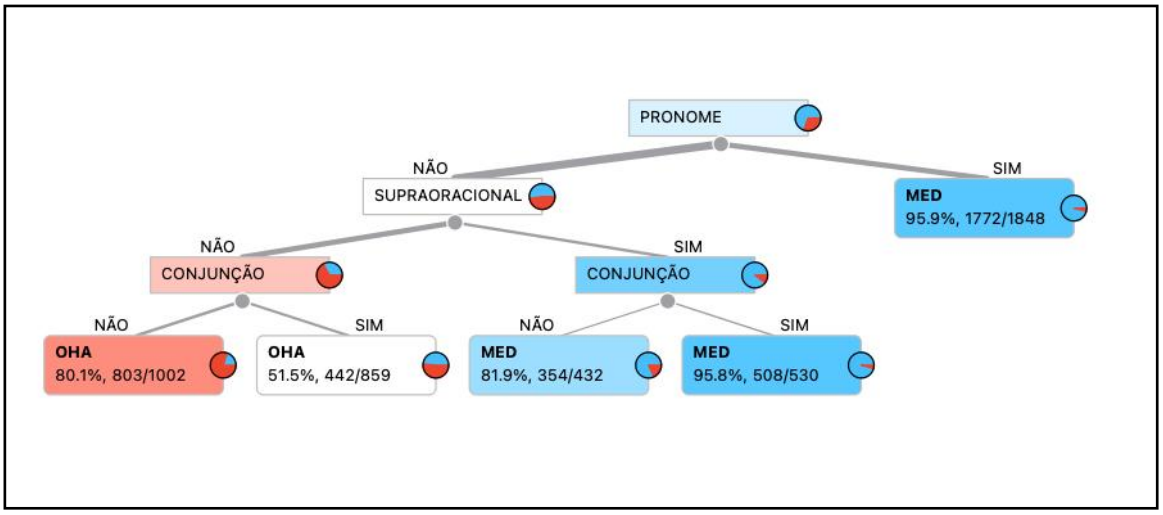
	Lexema	**Categoria**	**Isso**	**Que**	**Supraoracional**
1	Contar	MED	1	1	1
2	Contar	MED	1	1	1
(...)	(...)	(...)	(...)	(...)	(...)
2538	Mencionar	MED	1	0	0
(...)	(...)	(...)	(...)	(...)	(...)
4671	comentar	OHA	0	0	0

Fonte: Elaboração própria

O *dataframe* foi criado no Excel e importado para o software *Orange*. Dentro do sistema, recorremos aos *Widgets Files*, *Predictions, Tree* e *Tree Viewer* e geramos duas árvores de regressão. Na primeira, buscamos ver somente a atuação das três variáveis – a) anteposição do pronome demonstrativo *isso,* b) posposição da conjunção *que*, c) posição *supraoracional* – em relação à classe OHA ou MED[10]. O sistema retornou a seguinte regressão logística:

[10] Na segunda árvore de classificações, buscamos verificar como essas mesmas variáveis interagem com os dez diferentes lexemas que ocupam o *slot* do verbo *dicendi* na construção. No entanto, não trouxemos esses dados para este artigo, cujo objetivo primário é discorrer sobre as contribuições da ferramenta e mostrar os procedimentos necessários para sua operacionalização.

Gráfico 1- Árvore de regressão por inferência condicional para as construções de OHA e MED:



Fonte: Elaboração própria

Sem considerar os lexemas especificamente, observamos a seguinte hierarquia das variáveis: o que mais importa para a construção de MED, no *corpus*, é a presença do pronome demonstrativo *isso*. Dessa maneira, havendo o pronome *isso*, o lexema entra na construção por meio da instanciação de uma construção de MED em 95.9% das vezes (isto é, das 1.848 ocorrências em que isso ocorreu no contexto dos fenômenos analisados, 1.772 eram MED). Caso não haja o pronome demonstrativo *isso*, a segunda variável relevante é a posição supraoracional. Em estando a construção em posição supraoracional e estando seguida da conjunção *que*, ela é uma construção de MED 95.8% (508 das 530 ocorrências) e, sem a conjunção *que*, 81.9% das vezes (354 das 432 ocorrências). Por sua vez, se a construção ocorre em contexto oracional e sem posposição da conjunção *que*, são quase sempre OHA (80.1% das vezes – 803, de 1002 ocorrências), mas dividem a média entre OHA e MED se houver conjunção (51.5% de OHA, 442 de 859 ocorrências).

Para verificar o impacto de cada uma das variáveis preditoras para a seleção de MED é possível recorrer a outros métodos estatísticos. Nesta pesquisa, também foi realizada uma regressão logística binária (cf. Field; Miles & Field, 2012), por meio da aplicação do método *enter* no *software JASP*. Os resultados estão indicados no Quadro 1, apresentado na seção 2 deste artigo. O modelo foi estatisticamente significativo [$X^2(21) = 2.469.256$, $p < 0{,}001$; Tjur $R^2 = 0{,}471$], sendo capaz de prever adequadamente 83% dos casos (conforme *performance metrics* > accuracy).

Todas as variáveis tiveram impacto estatisticamente significativo, com preponderância do *pronome* (Odds Ratio = 34.286 [95% IC: 26.473 – 44.405], seguido

de *posição supraoracional* (Odds Ratio = 21.371 [95% IC: 16.095 – 27.016] e, por fim, *conjunção* (Odds Ratio = 4.285 [95% IC: 3.576 – 5.134].

Por fim, cabe mencionar que o ODM também apresenta *widgets* para a realização de regressão logística, mas, pelo menos na versão utilizada para este artigo, menos potentes e detalhados que outros *softwares* estatísticos, como o JASP e também o SPSS. Dessa maneira, entendemos que o ODM seja mais adequadamente utilizado junto a outros softwares (como o JASP, cuja interface é bastante amigável), sempre que for necessário a recorrer a cálculos mais robustos e detalhados. Logo, a principal vantagem do ODM que se defende neste artigo é a realização de testes de árvores de classificação, que são bem mais fáceis de programar e executar em comparação à interface do Programa R.

Considerações finais

Neste artigo, buscamos demonstrar a aplicação do recurso árvore de classificações (*classification tree*) no *Software Orange Data Mining (ODM)* para a análise de dados do uso linguístico. Trata-se de uma ferramenta que recorre à regressão logística para prever, por meio de uma estrutura hierárquica de decisões, o modo como uma ou mais variáveis preditoras, isoladamente ou em associação com outras, condicionam a instanciação de uma variável de desfecho.

O recurso atua de maneira semelhante à *árvore de regressão por inferência condicional*, desenvolvido em linguagem R no pacote *partykit* por Hothorn, Hornik e Zeileis (2006) e Speybroeck (2012), mas apresenta a vantagem de apresentar uma interface simples e amigável, de fácil utilização e interpretação para usuários com conhecimentos básicos de informática e pouco ou nenhum conhecimento de ferramentas estatísticas.

Defendemos, assim como diversos outros pesquisadores, que a aplicação de métodos estatísticos, como o que apresentamos neste artigo, é uma forma mais segura e eficaz para se chegar a determinadas generalizações linguísticas do que se ater exclusivamente à análise de frequências absolutas ou relativas. Afinal, são esses métodos que nos permitem afirmar, com certa precisão, quando há um forte nível de associação entre duas ou mais variáveis e nos possibilitam também prever se as

conclusões a que chegamos têm chances de serem extensíveis a outras amostras para além do nosso *corpus*. Por fim, cabe frisar que, a despeito da facilidade de operacionalização da ferramenta, um domínio mínimo de conhecimento estatístico em regressão logística por parte do analista continua sendo essencial.

Referências

CUNHA, C. E.; CINTRA, L. F. L. *Nova Gramática do Português Contemporâneo.* 7ª. Ed. Rio de Janeiro: Lexikon, 2001.

ELY, L; CEZARIO, M. M. [Vai que] e a modalidade: uma análise baseada no uso sobre o domínio condicional. *Soletras*, n. 45, p. 151-168, 2023.
FIELD, A.; MILES, J.; FIELD, Z. *Discovering Statistics Using T.* London: Sage Publications Ltd., 2012.

FREITAG, R. M. K.; PINHEIRO, B. F. M. Modelo de árvore de inferência condicional para explicar usos linguísticos variáveis. In: CARVALHO, C. S.; LOPES, N. S.; RODRIGUES, A. (Org.). *Sociolinguística e Funcionalismo.* Vertentes e Interfaces. Salvador: Eduneb, 2020, p. 317-342.

FURTADO DA CUNHA, M. A. O modelo das motivações competidoras no domínio funcional da negação. *Delta,* São Paulo, v. 17, n. 1, p. 1-30, 2001.

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression, 2nd ed.* [S.I.]: New York; Chichester, Wiley, 2000.

HOTHORN, T.; HORNIK, K.; ZEILEIS, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15, p. 651-675, 2006.

LACERDA, P. F. A. C. O papel do método misto na análise de processos de mudança em uma abordagem construcional: reflexões e propostas. *Revista Linguística/Revista do Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro*, Volume Especial, p. 83-101, 2016.

NEVES, M. H. M. *A gramática do português revelada em textos.* São Paulo: UNESP, 2018.
ROCHA LIMA, C. H. *Gramática Normativa da Língua Portuguesa.* 1a. Ed. Rio de Janeiro: José Olympio, 1972.

SANTANA, J. C.D. de.; NASCIMENTO, P. B. S. do. A negação no português falado da Matinha/BA: um estudo sociolinguístico. *Letra Magna,* [S.1], v. 14, p. 1-17, 2011.

SANTOS, M.; CEZARIO, M. M. Estudo cognitivo-funcional da formação da construção [Xque]$_{conect}$ no português. *Gallaecia*. Estudos de linguística portuguesa e galega. Santiago de Compostela, v. 1, p. 959-974, 2017.

SANTOS SILVA, T.; CEZARIO, M. M. Construcionalização e competição de conectores concessivos e concessivo-condicionais instanciados pelo esquema [Xque] em português. *Odisseia*, v. 4, n. especial, p. 132-153, 2019.

SPEYPROECK, N. Classification and regression trees. *International Journal of Public Health,* New York, V. 57, n. 1, p. 243-246, 2012.

TRAUGOTT, E. C. *Discourse Structuring Markers in English.* Philadelphia: John Benjamins, 2022.

TRAUGOTT, E. C.; DASHER, R. *Regularity in Semantic Change.* Cambridge: Cambridge University Press, 2002.

YAKOVENCO, L. C.; NASCIMENTO, C. A. R. A negação no português falado em Vitória/ES. *(Con)Textos Linguísticos,* Vitória, v. 10, n. 17, p. 122-138, 2016.