DOI: 10.20396/cel.v64i00.8666995

DESCRIÇÃO LINGUÍSTICA E APRENDIZADO DE MÁQUINA: ANÁLISE DE VERBOS LOCATIVOS DO ESPANHOL

DESCRIPCIÓN LINGÜÍSTICA Y APRENDIZAJE AUTOMÁTICO: ANÁLISIS DE VERBOS LOCATIVOS DEL ESPAÑOL

Roana Rodrigues[1]
Jackson Wilke da Cruz Souza[2]
Roney Lira de Sales Santos[3]

Resumo: Com o intuito de explicitar as relações estabelecidas entre a linguística descritiva e o aprendizado de máquina, este artigo apresenta os resultados de uma pesquisa que analisa um algoritmo gerado com base na classificação humana de construções verbais locativas da língua espanhola. Os dados utilizados na investigação foram retirados de Rodrigues (2019), que apresenta uma análise e uma descrição manuais de 318 instâncias que se constituem por verbos que selecionam, obrigatoriamente, um argumento interpretado como lugar (*poner, salir, entrar, enjaular* etc.), distribuídos em 10 classes distintas, segundo suas propriedades estruturais, distribucionais e transformacionais. Tendo como base o paradigma simbólico e utilizando o *software* Weka, os dados possibilitaram a geração de duas propostas de regras do algoritmo JRip: *sem* e *com* a seleção de atributos. Ambos os procedimentos geraram 10 regras compostas e avaliaram as medidas de precisão, cobertura, medida-f e matriz de confusão dos algoritmos criados. O algoritmo *sem* a seleção de atributos apresentou 100% de acurácia, demonstrando que os dados linguísticos apresentam uma descrição e classificação coerentes. Já o algoritmo *com* a seleção de atributos, possuindo 96,54% de acurácia, possibilitou, além de expor as propriedades linguísticas mais relevantes para fins classificatórios, analisar os casos mais sensíveis para a distinção das classes, culminando no levantamento de seis aspectos descritivos de revisão e/ou refinamento dos dados que devem ser analisados em trabalhos linguísticos futuros. Sendo assim, esta investigação auxiliou, mais especificamente, no aprimoramento da descrição das construções verbais locativas da língua espanhola e demonstrou que a relação *descrição humana* e *aprendizado de máquina* não consiste somente na importância da descrição como *insumo* para a máquina, mas, principalmente, em como é possível utilizar algoritmos (e suas medidas de avaliação) para validar e aperfeiçoar a descrição de diferentes fenômenos das línguas naturais.
Palavras-chave: aprendizado de máquina; sintaxe; verbo locativo.

Resumen: Con el fin de esclarecer las relaciones que se establecen entre la lingüística descriptiva y el aprendizaje automático, este artículo presenta resultados de una investigación que analiza un algoritmo generado a partir de una propuesta de clasificación humana de construcciones verbales locativas de la lengua española. Se utilizaron datos sacados de Rodrigues (2019), que presentan un análisis y descripción de 318 construcciones verbales que seleccionan, de manera obligatoria, un argumento interpretado como lugar (*poner, salir, entrar, enjaular* etc.), organizadas en 10 clases distintas, de acuerdo con sus atributos estructurales, distribucionales y transformacionales. Partiendo del paradigma simbólico y utilizando el *software* Weka, los datos permitieron generar dos propuestas de reglas del algoritmo JRip: *sin* y *con* la selección de atributos. Ambos los procedimientos generaron 10 reglas compuestas y evaluaron las medidas

[1] Universidade Federal de Sergipe (UFS), São Cristóvão, SE, Brasil. r.roanarodrigues@gmail.com
Orcid: https://orcid.org/0000-0002-7748-8716
[2] Universidade Federal da Bahia (UFBA), Camaçari, BA, Brasil. jackcruzsouza@gmail.com
Orcid: https://orcid.org/0000-0003-1881-6780
[3] Universidade Federal do Piauí (UFPI), Teresina, PI, Brasil. roneyleft@gmail.com
Orcid: https://orcid.org/0000-0001-9562-0605

de precisión, exhaustividad, puntuación-f1 y matriz de confusión de los algoritmos creados. El algoritmo *sin* selección de atributos presentó el 100% de desempeño, demostrando que los datos lingüísticos presentan una descripción y clasificación coherentes. Por su vez, el algoritmo *con* selección de atributos, con el 96,54% de desempeño, permitió, además de exponer las propiedades lingüísticas más relevantes con fines de clasificación, analizar los casos más sensibles para distinción entre las clases, culminando en la lista de seis aspectos descriptivos de revisión y/o refinamiento de datos que se deben analizar en investigaciones futuras. Por tanto, esta investigación auxilió, más específicamente, en la mejora de la descripción de las construcciones verbales locativas de la lengua española y demostró que la relación *descripción humana* y *aprendizaje automático* no consiste solamente en la importancia de la descripción como *input* para la máquina, pero, principalmente, sobre cómo es posible utilizar algoritmos (y sus métricas de evaluación) para validar y mejorar la descripción de diferentes fenómenos de las lenguas naturales.
Palabras clave: aprendizaje automático; sintaxis; verbo locativo.

1. INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que estabelece as relações entre cientistas da Linguística e da Computação, a fim de construir sistemas de reconhecimento e produção de informação baseados nas línguas naturais (VIEIRA; STRUBE, 2001). Segundo Finatto, Lopes e Ciulla (2015, p. 58), trabalhos nessa área, que associam linguagem e tecnologia, não são uma opção, mas sim uma necessidade, principalmente ao se considerar a escassez de descrições e recursos linguístico-computacionais no cenário brasileiro. Sendo assim, nesta pesquisa, pretende-se: (i) ressaltar a importância do papel do linguista na construção, revisão e aperfeiçoamento de recursos interpretáveis por máquina; e, sobretudo, (ii) evidenciar as contribuições que os meios computacionais podem oferecer para as reflexões e o refinamento de descrições linguísticas feitas por humanos.

Para discutir essa relação de mão dupla entre os estudos linguísticos e a computação, será interpretado um conjunto de regras gerado por um classificador simbólico do aprendizado de máquina, doravante AM, com base na tipologia descritiva (humana) das construções verbais locativas da língua espanhola, a base de dados Léxico-Gramática dos verbos Locativos do Espanhol – LGLE (RODRIGUES, 2019)[4]. Acredita-se que os conjuntos de regras gerados a partir de algoritmos em AM permitam a extração de informações e interpretações acerca da caracterização linguística da referida tipologia, contribuindo para a avaliação e o aperfeiçoamento dos dados descritivos linguísticos realizados inicialmente de maneira manual.

Ressalta-se que o AM é uma subárea da Inteligência Artificial que desenvolve técnicas computacionais para a aquisição automática de conhecimento a partir de sistemas que tomam "decisões baseadas em experiências acumuladas através da solução bem-sucedida de problemas anteriores" (MONARD; BARANAUSKAS, 2003, p. 39). De acordo com Antunes, Pardo e Almeida (2017), e em consonância com o que se propõe neste artigo, técnicas de AM podem ser aplicadas aos estudos linguísticos com a finalidade de verificar regularidades linguísticas não observáveis apenas com análises manuais, aprofundando, dessa maneira, a descrição linguística disponível.

Considerando-se o exposto, as questões que mobilizam a realização desta investigação são as seguintes: (i) de que maneira o AM pode realmente contribuir com a avaliação e o aprimoramento da descrição linguística humana?; (ii) quais informações linguísticas não haviam sido destacadas no momento da descrição manual que, à luz dos algoritmos gerados, passaram a ser evidenciadas?; e (iii) quais atributos (propriedades) das construções analisadas destacam-se e atuam como elementos fundamentais para a classificação dos verbos locativos estudados?

As considerações dessas questões são tidas como diretrizes para o aprofundamento de descrições manuais a serem realizadas em trabalhos futuros, já que, com base no algoritmo gerado, será possível saber quais são as propriedades linguísticas em potencial que caracterizam as classes dos verbos locativos do espanhol. Dessa maneira, esta investigação colabora para o estado da arte com a difusão do AM como tarefa contributiva à descrição linguística humana.

[4] LGLE foi a base de dados utilizada neste trabalho devido à facilidade de acesso aos seus dados.

Além disso, viabiliza a identificação e classificação automática desses tipos de verbos em possíveis aplicações linguístico-computacionais.

Tendo em vista a proposta e os objetivos, este trabalho se organiza da seguinte maneira: na seção 2 serão apresentadas a descrição e tipologia dos verbos locativos do espanhol (LGLE), as quais foram realizadas de maneira manual. Em seguida, na seção 3, são discutidos alguns aspectos relevantes para a compreensão do AM e apresentada a ferramenta utilizada nesta investigação (Weka) para geração dos algoritmos. Na seção 4, serão analisados os algoritmos gerados e suas medidas de avaliação. Por fim, serão apresentadas as considerações e contribuições deste trabalho, assim como os vislumbres de investigações futuras.

2. DESCRIÇÃO LINGUÍSTICA: A BASE DE DADOS LGLE

O Léxico-Gramática dos verbos Locativos do Espanhol (LGLE) é uma base de dados dos verbos que selecionam um argumento interpretado como *lugar* (*entrar, salir, poner, vivir* etc.), proposta por Rodrigues (2019), com base em estudos anteriores sobre o fenômeno em questão nas línguas francesa (GUILLET; LECLÈRE, 1992) e portuguesa (BAPTISTA, 2013). Em sua primeira versão, o LGLE apresenta a classificação sintático-semântica de 318 operadores (verbos locativos), distribuídos em 10 classes distintas, de acordo com as suas propriedades estruturais (preposições, posição e número de argumentos), distribucionais (tipo de argumento: papéis semânticos, nome humano, nome não humano, nome plural, nome locativo) e transformacionais (alterações sintáticas que mantêm, de maneira geral, as informações semânticas das frases de base: apassivação com *ser*, apassivação com *estar* e operação de *fusão*).

Os dados do LGLE estão dispostos em tabelas binárias, nas quais as linhas apresentam as 318 entradas lexicais estudadas e as colunas, as 49 propriedades sintático-semânticas analisadas. Trata-se de um produto linguístico que tem como arcabouço teórico-metodológico o modelo do Léxico-Gramática (GROSS, 1975), em que, partindo-se do léxico nuclear de uma frase de base, a qual se constitui pelo operador e os seus argumentos necessários e obrigatórios, são estabelecidas as suas relações sintático-semânticas em codependência. Quando determinado operador apresenta a propriedade em questão, é assinalado o símbolo "+"; quando não possui a propriedade, aponta-se o símbolo "-". A Tabela 1 exemplifica algumas das propriedades descritas no LGLE.

Verbo[5]	Classe	N_0 Papel Semântico	N_0=Hum	Prep1=0	N_1 Papel Semântico	N_1=nHum	$Prep_2$=0	$Prep_2$=Loc	N_2=0	N_2=Nloc	[Pas_Ser]	[Pas_Estar]	Frase de base
residir	35LS	agent-gen	+	-	Locative _Place	-	+	-	+	-	-	-	*Ana Julia residió en Burgos.*
viajar	37LD	agent-gen	+	-	Locative _Source	-	-	+	-	+	-	-	*La periodista viajó de Nicaragua a Londres.*
expulsar	38LS	agent-gen	+	-	Patient	-	-	+	-	+	+	+	*El rey expulsó a los judíos de Francia.*
cruzar	38L1	agent-gen	+	+	Locative _Path	-	+	-	+	-	+	-	*Una familia cruzó la frontera.*

Tabela 1*:* Exemplo da tabela binária do LGLE baseado em Rodrigues (2019).

Ressalta-se que os verbos locativos são entendidos em Rodrigues (2019) como aqueles que, *grosso modo*, respondem adequadamente à pergunta "(*preposição*) + *onde?*" e em que o argumento locativo, selecionado pelo verbo, é um elemento imprescindível para a constituição e compreensão da frase de base[6].

Como já mencionado, as classes locativas, assim como as siglas utilizadas (35LS, 35LD, 37LD, 38L1, entre outras – em que *L* se refere a *Locativo*, *S* a construções *eStáticas* e D, *Dinâmicas*), baseiam-se em trabalhos anteriores sobre as construções verbais locativas do francês (GUILLET; LECLÈRE, 1992) e do português europeu (BAPTISTA, 2013). A utilização de uma mesma terminologia segue os pressupostos do modelo do Léxico-Gramática, facilitando o desenvolvimento de estudos comparados entre as línguas. As construções locativas possuem variadas disposições sintático-semânticas, como se apresentam na Tabela 2:

[5] Notações: *Verbo*: verbo analisado; *Classe*: classificação da construção verbal locativa; N_0, N_1, N_2: sujeito e complementos; *Papel Semântico*: papel semântico do argumento; *Hum*: argumento preenchido por nome humano (entendido, *grosso modo*, como elemento animado); *nHum*: argumento preenchido por nome não humano (entendido, *grosso modo*, como elemento inanimado); *Loc*: preposição locativa; *Nloc*: argumento preenchido por nome locativo; *PAS_SER*: a construção admite passiva com o verbo *ser*; *PAS_ESTAR*: a construção admite passiva com o verbo *estar*. A notação "=0" significa que o argumento não é preenchido.

[6] A frase de base é a unidade mínima de análise para o modelo do Léxico-Gramática e se constitui pelo operador, que no caso é o verbo locativo, e seus elementos necessários e essenciais para a constituição da frase. É a partir da frase de base que se estabelecem os testes sintático-semânticos para o preenchimento das propriedades estruturais, distribucionais e transformacional de cada construção.

Classe	Estrutura[7]	Verbo	Frase de base[8]	#
35LD	N_0 Vdin Loc_1 $Nloc_1$	*acceder*	*Los migrantes accedieron a costas españolas.*	66
35LS	N_0 Vstat Loc_1 $Nloc_1$	*habitar*	*El reptil habitó en la zona.*	15
37LD	N_0 Vdin Loc-s_1 $Nloc_1$ Loc-d_2 $Nloc_2$	*saltar*	*El hombre saltó desde la plataforma hasta la pista.*	46
38L1	N_0 V $Nloc_1$	*abandonar*	*El jugador abandonó el recinto.*	43
38L2	N_0 Nloc-v [prep-a] N_1 [V=poner en Nloc]	*empaquetar*	*El actor empaquetó sus cosas.*	7
38L3	$Nloc_0$ V [prep-a] N_1	*albergar*	*El pequeño hotel albergaba a un grupo.*	4
38LD	N_0 Vdin [prep-a] N_1 Loc-d_2 $Nloc_2$	*guardar*	*El estudiante guardó el libro en el estante.*	101
38LS	N_0 Vdin [prep-a] N_1 Loc-s_2 $Nloc_2$	*evacuar*	*La policía evacuó a los turistas del lugar.*	19
38LT	N_0 Vdin [prep-a] N_1 Loc-s_2 $Nloc_2$ Loc-d_3 $Nloc_3$	*canalizar*	*Isabel II canalizó el agua del Lozoya hasta Madrid.*	14
38R	N_0 Vstat [prep-a] N_1 Loc_2 N_2	*localizar*	*Eduardo localizó el poblado minero en el mapa.*	3
Total				**318**

Tabela 2: Classes de construções verbais locativas do LGLE (RODRIGUES, 2019, p. 121).

No LGLE, são três as classes locativas estativas: 35LS, 38L3 e 38R. Em 35LS estão as construções constituídas por um nome locativo preposicionado, conforme se verifica em (1). A classe 38L3, por sua vez, agrupa construções transitivas diretas, em que o locativo ocupa a posição argumental de sujeito, como em (2). Por fim, a classe 38R, exemplificada em (3), é marcada por casos residuais que, até o momento, não parecem enquadrar-se nas construções recenseadas e nas outras classes propostas, ratificando a classificação para o português europeu de Baptista (2013).

(1) *Manuela vivió en Barcelona.*[9]
(2) *Un templo cobija imágenes de estilo gótico.*
(3) *El Instituto se sitúa en Barcelona.*

As demais classes são marcadas por construções dinâmicas. Embora apresente semelhança sintática à 35LS, a classe 35LD agrupa verbos que denotam a movimentação do argumento que ocupa a posição de sujeito, como em (4):

(4) *Robert entró en la sinagoga.*

As classes 38LD e 38LS se constituem por três argumentos que ocupam as posições de sujeito, complemento direto e argumento preposicionado locativo. A diferença entre essas classes está no fato de que em 38LD o argumento (na posição de complemento direto) se desloca a um lugar de destino, (5), enquanto em 38LS o deslocamento se dá desde um local de origem, (6):

(5) *El presidente encerró al equipo en el vestuario.*
(6) *Isabel desalojó a Muñoz de la alcaldía.*

[7] Notações: N_0, N_1, N_2, N_3: sujeito e complementos; *Prep*: preposição; [*prep-a*]: possibilidade da preposição *a*, quando N_1 é um nome humano; *Nloc*: nome locativo; *Loc*: preposição locativa, *-d* de destino, *-s* de origem; *V*: verbo, *Vdin*: verbo locativo dinâmico; *Vstat*: verbo locativo estativo, *Nloc-v*: verbo denominal locativo.

[8] Em Rodrigues (2019), as frases de base foram extraídas da base de dados ADESSE (GARCÍA-MIGUEL *et al.*, 2003) e da *web*, através da ferramenta WebCorp (*The Web as a Corpus*). Além disso, houve um processo de validação dos dados realizado por 20 falantes de diferentes países de língua espanhola, havendo um predomínio de informantes de nacionalidade espanhola. WebCorp disponível em: <https://www.webcorp.org.uk/live/>. Acesso em: out. 2021.

[9] As frases desta seção foram retiradas de Rodrigues (2019).

As classes 37LD e 38LT agrupam verbos que selecionam tanto um lugar de origem, quanto um lugar de destino. Distinguem-se, no entanto, quanto ao número de argumentos e ao elemento deslocado: em 37LD é o argumento que ocupa a posição de sujeito que se desloca, (7); já em 38LT é o argumento na posição de complemento direto que é movimentado, (8):

(7) *Los jóvenes bajaron de lo alto de la torre hasta el lago.*
(8) *Thomas transportó a los pasajeros de la ciudad hasta el archipiélago.*

Por fim, mencionam-se as classes transitivas diretas: 38L1 e 38L2. Em 38L1, têm-se construções estáticas (9a) ou dinâmicas (9b) de origem, destino ou trajetória: o elemento que ocupa a posição de sujeito *está* ou *se desloca* de/a/por um determinado local:

(9) a. *Pedro habita una aldea.*
b. *El inmigrante escaló un edificio.*

A classe 38L2, por sua vez, constitui-se por verbos denominais locativos, ou seja, verbos que se constituem por um nome de lugar, como *enjaular* em (10a). Conforme se verifica em (10b) e em (10c), essas construções são marcadas pela operação transformacional de *fusão*, em que ocorre a união de frases que possibilita a modificação do número de argumentos de um verbo, a partir da combinação de dois verbos entre si ou de um verbo e um argumento, resultando no apagamento de um dos dois elementos (GROSS, 1981, p. 45):

(10) a. *El humano enjauló el animal.*
b. *El humano puso # el animal está en la jaula.*[10]
c. *El humano puso # el animal en la jaula.*

A descrição e classificação das construções verbais locativas de Rodrigues (2019) contribuem para os estudos descritivos linguísticos em si, proporcionando a reflexão sobre as particularidades do fenômeno em língua espanhola. Os dados dispostos de maneira formalizada em tabelas binárias facilitam a sua compreensão e comparação entre as línguas naturais, além de atuar como recurso linguístico-computacional, com variadas aplicações à área de PLN.

Tendo em consideração a descrição linguística de Rodrigues (2019), a qual foi realizada de maneira manual, as próximas seções apresentarão informações relativas ao aprendizado de máquina (AM) e em qual medida essa área pode contribuir para o aprimoramento da descrição manual humana.

3. APRENDIZADO DE MÁQUINA: CONCEITOS BÁSICOS

O AM pode ser compreendido como o estudo de métodos computacionais que tem o objetivo de adquirir novos conhecimentos e habilidades, além de meios de organizar o conhecimento previamente existente (MITCHELL, 1997). Ademais, o estudo das técnicas de AM é capaz de fornecer um melhor entendimento do processo do raciocínio humano (MONARD *et al.*, 1997).

De acordo com Monard e Baranaukas (2003), existem diversos paradigmas de AM, a saber: (i) *estatístico*: criação de métodos de classificação com base em modelos estatísticos; (ii) *baseado em exemplos*: classificação de novos exemplos com base em exemplos similares conhecidos; (iii) *conexionista*: construções matemáticas inspiradas no modelo biológico do sistema nervoso, redes neurais; (iv) *genético*: predição baseada em uma população de informações para a geração do classificador; e (v) *simbólico*: construção de representações simbólicas a partir da análise de exemplos e contraexemplos do conceito. Esta pesquisa se apoia no paradigma simbólico, tendo em consideração a afirmação de Souza (2019, p. 77) de que se trata de um

[10] O símbolo # marca a independência entre frases.

paradigma mais utilizado "em tarefas de interpretação humana, sendo representado[s] por árvores de decisão ou por conjuntos de regras".

Os algoritmos baseados em regra estruturam-se sob o modelo lógico *se-então*: dado que se satisfaçam certas condições, a instância é rotulada com determinada classe. Esse tipo de algoritmo, de maneira geral, constrói classificadores em conjuntos de regras, em que *se* uma instância é classificada com determinada classe, *então* compreende-se que as condições anteriores a ela foram satisfeitas; caso isso não ocorra, o classificador aplicará a classe mais genérica do conjunto de dados. Nesse sentido, Kubat (2017) aponta que os conjuntos de regras, além de facilitarem a interpretação das informações, permitem observar a recursividade dos atributos. Assim, tem-se que um atributo já utilizado na determinação de uma classe pode ocorrer novamente em combinação com outro, resultando no apontamento de uma classe diferente.

Nesta pesquisa, a partir de um modo de aprendizado *supervisionado*[11], com o fornecimento dos dados do LGLE previamente classificados, ou seja, com os dados cujo rótulo da classe já é conhecido, foi utilizado o algoritmo de aprendizado JRip (baseado em regras) (COHEN, 1995), por meio do *software* Weka, *Waikato Environment for Knowledge Analysis* (WITTEN; FRANK, 2005). O Weka é uma ferramenta que agrupa algoritmos de diferentes abordagens de AM, por meio de uma interface de fácil manuseio, além de funcionalidades de preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização.

Para medir a acurácia do classificador gerado (neste caso, o algoritmo JRip), foram selecionadas as medidas de precisão, cobertura e medida-f, definidas no Quadro 1, comumente utilizadas em avaliações na área de PLN.

Medidas	**Definição**	**Fórmula**
Precisão	Quantidade de instâncias corretamente classificadas em relação à quantidade total de instâncias. Baseia-se na quantidade de casos *true positives* (TP) em razão da quantidade somada de casos verdadeiros e *false positives* (FP).	$Precisão = \frac{TP}{TP + FP}$
Cobertura	Quantidade de casos corretamente detectados em relação à quantidade que deveria ser detectada, ou seja, a quantidade de instâncias cobertas pelo classificador. Baseia-se na quantidade de casos TP em razão da quantidade somada de casos verdadeiros e *false negatives* (FN).	$Cobertura = \frac{TP}{TP + FN}$
Medida-F	Média ponderada das medidas anteriores.	$Medida - F = \frac{(Precisão.Cobertura).2}{(Precisão + Cobertura)}$

Quadro 1: Medidas de acurácia. Quadro baseado em Jurasky e Martin (2000).

Além disso, optou-se pela verificação da distribuição dos dados na matriz de confusão como outra medida de avaliação, em que é possível observar a dispersão dos dados e possíveis similaridades entre as classes propostas pelo LGLE. Conforme definem Monard e Baranaukas (2003, p. 47), "a matriz de confusão de uma hipótese *h* oferece uma medida efetiva do modelo de classificação ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos *T*". Os dados relativos à acurácia e à análise da matriz de confusão serão apresentados na próxima seção.

[11] De acordo com Monard e Baranaukas (2003, p. 40), à contraposição do aprendizado *supervisionado*, tem-se o aprendizado *não-supervisionado*, o qual analisa os exemplos fornecidos e tenta determinar se podem ser, de alguma maneira, agrupados.

4. AM E LGLE: ANÁLISE DO CLASSIFICADOR

Visando a uma melhor organização, esta seção divide-se em duas partes. Na primeira, verificam-se os procedimentos metodológicos e a geração do algoritmo JRip, com a utilização do *software* Weka. Na segunda, abordam-se a interpretação linguística humana, a partir da análise do algoritmo gerado, e os desdobramentos descritivos em pesquisas e ações futuras.

4.1 Algoritmo JRip

O algoritmo JRip, proposto por Cohen (1995), consiste em um aprendizado de regra proposicional a partir de exemplos já existentes, em que encontra um conjunto de regras com bom desempenho de acordo com o conjunto de exemplos e a medida de performance. O desempenho do algoritmo, ao final, é medido pelo ganho de informação obtido com a regra. Além disso, o JRip trabalha tratando todos os exemplos em particular julgamento dos dados de treinamento como uma classe, e encontra um conjunto de regras que cobre todos os membros da classe. Esse processo é executado da mesma forma para todas as classes em sequência, até que todas as classes tenham sido cobertas (RAJPUT *et al.*, 2011; SILVA, 2019).

Para facilitar a interpretação dos dados gerados no *software* Weka, realizaram-se substituições de anotação do LGLE: "+" tornou-se "sim" e "-", "não", facilitando a leitura e interpretação dos dados gerados. Além disso, viu-se a necessidade do acréscimo da notação NSA (*não se aplica*) aos casos em que a propriedade em questão não pode sequer ser aplicada à construção verbal analisada, tal como a propriedade transformacional de *apassivação* que não se aplica às construções sem complemento direto.

É importante mencionar ainda que, para a construção do classificador por meio do algoritmo JRip, contou-se com a técnica de *k-fold cross-validation*, a partir dos dados da base LGLE. Nesse sentido, de maneira automática, as instâncias foram aleatoriamente divididas em k partições com tamanho proporcional e aproximado, utilizando-se as instâncias para treinamento ($k - 1$) e testadas no conjunto restante (ou seja, 1). Esse procedimento é repetido k vezes, resultando em um classificador que apresenta uma taxa média de erro entre os testes (DEPREN *et al.*, 2005). Neste trabalho, foram utilizadas 10 partições, cada qual indicada na literatura como quantidade recorrente em trabalhos dessa natureza (MONARD; BARANAUSKAS, 2003). Como o LGLE não possui uma quantidade de dados extensa, com o recenseamento de apenas 318 construções verbais, a utilização da técnica de *cross-validation* se faz necessária. Diante do desbalanceamento dos dados (a classe 38R possui 3 instâncias, enquanto a classe 38LD possui 101, por exemplo), utilizar essa técnica mitiga a possibilidade de os resultados gerados pelos classificadores não representarem a realidade linguística do fenômeno descrito, especialmente para os classificadores do tipo *conjunto de regras*, já que se compreende a relevância dos atributos segundo a quantidade de ocorrência.

Desse modo, foi construído o classificador utilizando todos os atributos do conjunto: as 49 propriedades estruturais, distribucionais e transformacionais descritas no LGLE. Entretanto, ao criar o primeiro classificador, percebeu-se que alguns atributos não ocorriam nas regras. É importante notar que, enquanto na descrição linguística humana há maior propensão para a valorização da precisão dos dados, na computação privilegia-se a otimização dos sistemas, os quais devem conter um número reduzido de regras que possam apresentar alta precisão.

Assim, decidiu-se investigar quais atributos poderiam ser mais proeminentes na caracterização e identificação de cada classe do LGLE. Para tanto, utilizou-se o algoritmo *InfoGainAttributeEval* (QUINLAN, 1986), o qual constrói um *ranking* entre os atributos, categorizando aqueles que são mais relevantes (com relação à recorrência) na caracterização das instâncias. Esse resultado se dá a partir do cálculo da entropia, o qual mede a pureza de subconjuntos: quanto menor a entropia, maior a surpresa nos resultados e, consequentemente, maior o ganho de informação. Com base nos resultados fez-se um corte de 70% no *ranking*, selecionando apenas 20 atributos a serem utilizados, como é apresentado na Tabela 3.

Ranking	Desempenho	Atributo[12]	Ranking	Desempenho	Atributo
1	155.621	N_1 Papel Semântico	26	0.26045	N_3=Npl
2	144.543	N_2 Papel Semântico	27	0.26045	$Prep_3$=0
3	134.412	$Prep_2$=de	28	0.26045	$Prep_3$=a
4	12.316	$Prep_2$=a	29	0.26045	N_3=Nloc
5	123.094	$Prep_2$=en	30	0.26045	N_3=0
6	106.682	Pas_ser	31	0.26045	N_3=sem-X
7	102.314	N_2=sem-X	32	0.26045	N_3=Hum
8	102.062	Pas_estar	33	0.26045	N_3=nHum
9	102.002	$Prep_2$=por	34	0.26045	N_3 Papel Semântico
10	0.99654	N_1=Nloc	35	0.26045	$Prep_3$=por
11	0.9835	N_2=Hum	36	0.26045	$Prep_3$=en
12	0.9835	$Prep_2$=0	37	0.26045	$Prep_3$=de
13	0.9835	N_2=nHum	38	0.26045	$Prep_3$=Loc
14	0.9835	N_2=Nloc	39	0.15792	N_0=nHum
15	0.9835	N_2=Npl	40	0.1526	fusão
16	0.9835	N_2=0	41	0.11478	N_0=Hum
17	0.9835	$Prep_2$=Loc	42	0.10827	N_1=sem-X
18	0.97058	$Prep_1$=Loc	43	0.09744	N_0=Nloc
19	0.72251	$Prep_1$=0	44	0.02898	$Prep_1$=por
20	0.71293	N_1=nHum	45	0.02898	N_0=Npl
21	0.5337	$Prep_1$=de	46	0.02485	N_0=sem-X
22	0.43512	N_1=Hum	47	0.00522	N_1=Npl
23	0.40439	$Prep_1$=en	48	0	N_1=0
24	0.31796	N_0 Papel Semântico	49	0	N0=0
25	0.30227	$Prep_1$=a			

Tabela 3: Atributos proeminentes nas classes do LGLE.

De acordo com a Tabela 3, os atributos mais relevantes para a classificação das construções verbais locativas estão compreendidos entre as posições 1 e 20 do *ranking*. Isso se dá porque se trata de características recorrentes em várias classes do LGLE e, por isso, são tidas como importantes no processo de classificação. Conforme será discutido na próxima seção, há atributos que ocupam outras posições do *ranking* que são fundamentais para a caracterização de algumas classes, como a propriedade "fusão" que, embora ocupe a posição 40, é decisiva para identificar as construções da classe 38L2.

Como visto, o JRip é um algoritmo baseado em regras. Assim, o algoritmo testa a primeira regra e, caso ela não satisfaça a identificação e classificação dos dados, passa a testar as demais até que todas sejam analisadas. Caso nenhuma seja verdadeira, uma última regra servirá de *default* para a descrição dos dados, como demonstrado no Quadro 2.

[12] Notações: *N_0, N_1, N_2, N_3*: sujeito e complementos; *Papel Semântico*: papel semântico do argumento; *Hum*: argumento preenchido por nome humano (entendido, *grosso modo*, como elemento animado); *nHum*: argumento preenchido por nome não humano (entendido, *grosso modo*, como elemento inanimado); *Npl:* argumento preenchido por nome plural; *Nloc*: argumento preenchido por nome locativo*; Loc*: preposição locativa; *Prep*: preposição; PAS_*SER*: passiva com o verbo *ser*; sem-X: marca de traço semântico específico; *PAS_ESTAR*: passiva com o verbo *estar*. A notação "=0" significa que o argumento não é preenchido.

JRIP – SEM SELEÇÃO DE ATRIBUTO (100% de acurácia – 318 instâncias corretamente classificadas)	JRIP – COM SELEÇÃO DE ATRIBUTO (96,54% de acurácia – 307 instâncias corretamente classificadas)
1. **Se** N_2 Papel Semântico = *Locative_Place*, **então** a classe é *38R* (3.0/0.0)	1. **Se** N_2 Papel Semântico = *Locative_Place*, **então** a classe é *38R* (3.0/0.0)
2. **Senão** N_0 Papel Semântico = *Locative_Place*, **então** a classe é *38L3* (4.0/0.0)	2. **Senão** Pas_ser = *não* **e** N_1=nHum = *sim*, **então** a classe é *38L3* (4.0/0.0)
3. **Senão** fusão = *sim*, **então** a classe é *38L2* (7.0/0.0)	3. **Senão** $Prep_2$=0 = *NSA* **e** N1 Papel Semântico = *patient*, **então** a classe é *38L2* (7.0/0.0)
4. **Senão** $Prep_3$= 0 = *não*, **então** a classe é *38LT* (14.0/0.0)	4. **Senão** N_2 Papel Semântico = *Locative_Source* **e** Pas_estar = *não*, **então** a classe é *38LT* (13.0/4.0)
5. **Senão** N_1 Papel Semântico = *Locative_Place* **e** Prep1= 0 = *não*, **então** a classe é *35LS* (15.0/0.0)	5. **Senão** N_1 Papel Semântico = *Locative_Place* **e** $Prep_1$=0 = *não*, **então** a classe é *35LS* (15.0/0.0)
6. **Senão** N_2 Papel Semântico = *Locative_Source*, **então** a classe é *38LS* (19.0/0.0)	6. **Senão** N_2 Papel Semântico = *Locative_Source*, **então** a classe é *38LS* (20.0/5.0)
7. **Senão** $Prep_2$=0 = *NSA* **e** $Prep_1$=0 *sim*, **então** a classe é *38L1* (43.0/0.0)	7. **Senão** $Prep_2$=0 = *NSA* **e** $Prep_1$=0 = *sim,* **então** a classe é *38L1* (43.0/0.0)
8. **Senão** N_1 Papel Semântico = Locative_Source **e** $Prep_2$=0 = não, então a classe é 37LD (46.0/0.0)	8. **Senão** N_1 Papel Semântico = *Locative_Source* **e** $Prep_2$=0 = *não*, **então** a classe é *37LD* (46.0/0.0)
9. **Senão** $Prep_1$=Loc = *sim*, **então** a classe é *35LD* (66.0/0.0)	9. **Senão** $Prep_1$=Loc = *sim*, **então** a classe é *35LD* (66.0/0.0)
10. **Caso contrário**, a classe é *38LD* (101.0/0.0)	10. **Caso contrário**, a classe é *38LD* (101.0/0.0)

Quadro 2: Regras geradas pelo algoritmo JRip sem e com a seleção de atributos.

No Quadro 2, apresenta-se o algoritmo *sem* e *com* a seleção de atributos. Como se verifica, a primeira regra do classificador *sem* a seleção de atributo prevê que, se N_2 Papel Semântico tiver o valor *Locative_Place*, a classe do verbo locativo será 38R, cobrindo todos os 3 casos recenseados. Como ainda há classes que não foram contempladas, o classificador testa a segunda regra, que indica que, se o atributo N_0 Papel Semântico tiver o valor *Locative_Place*, então a classe é 38L3, cobrindo todas as 4 instâncias. O classificador continua testando todas as nove regras e, caso nenhuma seja verdadeira, o verbo pertence à classe 38LD, funcionando como a regra *default*. Como resultado, obteve-se um classificador que identifica corretamente todas as 318 instâncias, o que significa 100% de acurácia.

O classificador *com* seleção de atributos foi proposto a fim de analisar se seria possível manter uma alta acurácia em detrimento da diminuição do conjunto de atributos. Assim, foram escolhidos apenas os atributos mais relevantes do conjunto inicial (cf. Tabela 3), obtendo-se um classificador ainda com 10 regras e com quase 97% de acurácia, errando 11 instâncias. É pertinente considerar que, para manter uma boa acurácia, os classificadores (*com* e *sem* seleção de atributos) propõem regras compostas por mais de um atributo, como é o caso, por exemplo, da regra 2 no classificador *com* a seleção, que avalia os atributos *Pas_ser = não* **e** N_1*=nHum = sim*.

O Weka disponibiliza dois mecanismos de avaliação dos classificadores propostos: matriz de confusão e medidas de avaliação (precisão, cobertura e medida-f). A Tabela 4 apresenta os dados referentes ao classificador *com* a seleção de atributos, já que *sem* a seleção obteve-se uma acurácia de 100%. Na próxima seção, serão discutidos e interpretados os dados gerados pelo algoritmo JRip.

Matriz de confusão											Medidas de avaliação		
35LD	**38L1**	**38LD**	**38L3**	**38LS**	**37LD**	**38LT**	**35LS**	**38L2**	**38R**	**Classe**	**Precisão**	**Cobertura**	**Medida-F**
66	0	0	0	0	0	0	0	0	0	**35LD**	1	1	1
0	43	0	0	0	0	0	0	0	0	**38L1**	1	1	1
0	0	100	1	0	0	0	0	0	0	**38LD**	0,99	0,99	0,99
0	0	0	4	0	0	0	0	0	0	**38L3**	1	0,88	0,89
0	0	0	0	15	0	4	0	0	0	**38LS**	0,78	0,76	0,75
0	0	0	0	0	46	0	0	0	0	**37LD**	1	1	1
0	0	0	0	5	0	9	0	0	0	**38LT**	0,64	0,66	0,65
0	0	0	0	0	0	0	15	0	0	**35LS**	1	1	1
0	0	1	0	0	0	0	0	6	0	**38L2**	1	0,92	0,92
0	0	0	0	0	0	0	0	0	3	**38R**	1	1	1

Tabela 4: Acurácia do algoritmo JRip *com* a seleção de atributos.

4.2. Interpretação linguística do algoritmo e das medidas de avaliação

Acredita-se, neste trabalho, que os dados gerados pelo algoritmo JRip são capazes de contribuir para o refinamento da descrição linguística humana. Assim, na matriz de confusão, verificam-se os desvios precisos na classificação das instâncias, que, consequentemente, aparecem nas medidas de avaliação com os seguintes pares: 38L3/38LD; 38L2/38LD e 38LT/38LS.

Como se verifica nas regras geradas *sem* a seleção de atributos (e com 100% de acurácia), as classes 38L3, 38L2 e 38LT possuem atributos particulares e exclusivos que as caracterizam, a saber: (i) a classe 38L3 é marcada pelo papel semântico *Locative_Place* na posição de sujeito da oração (N_0); (ii) a classe 38L2 se constitui por verbos denominais locativos, apresentando o polo positivo para a propriedade transformacional de *fusão*; e (iii) a classe 38LT se caracteriza por apresentar valência quatro, em que o verbo seleciona argumentos nas posições de sujeito, complemento direto e dois argumentos preposicionados, com os papéis semânticos de *Locative_Source* e *Locative_Destination*, respectivamente.

Sendo assim, os atributos apontados (*N_0=Locative_Place=Sim; Fusão=Sim; $Prep_3$=0=Não*), embora sejam fundamentais para caracterizar cada uma dessas classes, não são produtivos se comparados aos atributos recorrentes nas demais construções locativas anotadas no LGLE. Portanto, ao considerar as regras geradas *com* a seleção de atributos, essas propriedades, por ocuparem no *ranking* posições pouco relevantes para a máquina, são desconsideradas na geração de regras. Isso impacta, consequentemente, na avaliação da cobertura das classes envolvidas (38L3, 38L2, 38LT), além das relações que se estabelecem com outras classes da base de dados (38LD e 38LS). Para além da interpretação das regras geradas e da compreensão do uso de atributos mais ou menos relevantes, os desvios destacados na matriz de confusão permitem revisar e refletir sobre a descrição linguística humana e as relações estabelecidas entre estas classes 38L3/38LD, 38L2/38LD e 38LT/38LS.

Além da ausência do atributo *N_0=Locative_Place=Sim*, que caracteriza os verbos da classe 38L3, concomitantemente, o fenômeno que parece haver classificado uma construção verbal de maneira indevida na classe 38L3 – e não na 38LD – foi a apassivação com *ser* (*pas_ser*). Os verbos da classe 38LD selecionam um argumento na posição de complemento direto que se desloca a um lugar de destino e apresentam o preenchimento "sim" para o atributo *pas_ser*. No entanto, revisitando a descrição humana, verificou-se um único caso em que a *pas_ser* apresentou a anotação "não": o verbo *pasear*, como exemplificado em (11):

(11) *La arqueóloga paseó a un grupo de profesores por Palma.*

Conforme consta no *Diccionario de la Lengua Española* (REAL ACADEMIA ESPAÑOLA, 2014), o verbo *pasear* admite um uso transitivo sinônimo de *hacer pasear*, tais como: *pasear a un niño* e *pasear a un caballo.* Sendo assim, na frase (11), constata-se uma construção em que o elemento que ocupa a posição de sujeito (*La arqueóloga*) "faz passear" o

elemento que ocupa a posição de complemento direto[13] (*a un grupo de profesores*) em um determinado lugar (*por Palma*). No entanto, essa construção se destaca por não admitir a passiva com "ser", como as demais construções anotadas na classe 38LD, o que demonstra ter um comportamento particular na língua, com restrições de preenchimento na posição de complemento direto, que precisam ser analisadas em investigações futuras[14].

Da relação 38L2/38LD não foi possível compreender um aspecto específico do desvio de classificação nas regras geradas, além da ausência do atributo *fusão* como elemento fundamental para a caracterização dos verbos da classe 38L2. Não obstante, tal desvio possibilitou um olhar mais atento à descrição das duas classes e a elaboração de duas reflexões relevantes: (i) a necessidade de especificar o papel semântico do elemento que ocupa a posição de N_1; e (ii) a análise de construções denominais que possibilitam a explicitação de um argumento preposicionado locativo externo. Sobre o primeiro aspecto, no LGLE, além dos papéis semânticos locativos dos argumentos que ocupam a posição sintática N_1 (*Locative_Source, Locative_Path, Locative_Destination* e *Locative_Place*), têm-se construções em que N_1 é ocupado pelo elemento que é (des)locado na construção locativa. Nesses casos, o papel semântico anotado no LGLE é o de Paciente [*Patient*]:

(12) a. *El padre confinó a Elisabeth en el sótano del edificio.* [*Patient*]
b. *El actor empaquetó sus cosas*. [*Patient*]

Tanto em construções 38LD (12a), quanto em construções 38L2 (12b), o elemento que ocupa a posição de N_1 tem papel semântico *Patient*, o que aproxima os verbos das duas classes. No LGLE, a anotação é limitante, no sentido de apresentar apenas o papel *Patient*, tanto para os casos em que o argumento que ocupa essa posição é um nome humano, quanto para as situações em que se tem um nome não humano. No ViPEr (BAPTISTA, 2013), uma base de dados verbais do português europeu, propõe-se a anotação *Patient* para referir-se a nomes humanos e, dentre outros papéis, *Object_Generic* para referir-se a nomes não humanos (SANTOS, 2014), que ocupam a posição N_1. A realização dessa distinção pode, eventualmente, gerar dados relevantes de análise em trabalhos futuros.

O segundo aspecto elencado refere-se a outro fenômeno que evidencia a relação estabelecida entre os verbos das classes 38L2 e 38LD: a existência de construções de base nominal que selecionam um argumento preposicionado locativo, como se vê em (13) com o verbo *almacenar*:

(13) *La asociación almacenó los papeles en sus hangares.*

Como se nota em (13), o verbo *almacenar* tem como base o nome *almacén*, mas sua construção seleciona outro argumento locativo (*en sus hangares*). Rodrigues (2019, p. 133) menciona esse comportamento verbal em sua pesquisa e afirma que, nesses casos, "a construção verbal se afastou do sentido locativo do nome da qual deriva, possibilitando a seleção de outros argumentos locativos, introduzidos, sobretudo, pela preposição *en*". Sendo assim, a autora opta por classificar tais construções na classe 38LD, devido à sua valência, distanciando-as do comportamento dos verbos da classe 38L2. Esse fenômeno explicita a relação sintático-semântica entre verbos da classe 38L2 e 38LD, o que pode justificar, em determinada perspectiva, a necessidade de estudos específicos sobre os verbos denominais locativos da língua espanhola.

Por fim, a valência verbal também justifica os desvios encontrados nas regras do algoritmo JRip entre as classes 38LS e 38LT. Ambas apresentam *Locative_Source* como o papel

[13] A tradição gramatical de língua espanhola prescreve o uso da preposição *a* para introduzir um objeto direto que designa uma ou várias pessoas ou coisa personificada (RAMMÉ, RODRÍGUEZ, 2020).

[14] Em uma análise preliminar de dados extraídos do *Corpus del Español del Siglo XXI* (CORPES), a construção *N_0 pasear a N_1* parece possuir algumas restrições de preenchimento, tais como a seleção de um nome humano na posição de complemento direto, sobretudo *niño*, e um nome de animal, principalmente *perro*. Além dessas restrições, verifica-se seu uso em locuções com os verbos *sacar* e *salir* (*a pasear*). CORPES, disponível em: <https://apps2.rae.es/CORPES2/>. Acesso em: out. 2021.

semântico do argumento N_2 e têm dados muito dispersos quanto ao preenchimento da passiva com *estar* (*pas_estar*). Como a seleção dos atributos mais relevantes ignorou a informação sobre a valência quatro da classe 38LT, essa classe se "fundiu" à classe 38LS. Essa união motivou a retomada de tais classificações e evidenciou a necessidade de uma revisão sobre a anotação da propriedade transformacional de apassivação com *estar* para os verbos de movimento, além de indicar um estudo mais refinado sobre os verbos locativos que selecionam tanto um argumento de origem, quanto um argumento de destino, como se observa na frase (14):

(14) *Lemke desvió el avión [de su plan de vuelo previsto] [hacia Nueva York].*

Segundo Rodrigues (2020, p. 473), cerca de 20 construções locativas apresentaram um comportamento semelhante à frase (14), em que se nota a omissão de um argumento locativo (de *origem* e/ou de *destino*), assemelhando-se e, em certa medida, mesclando-se ao comportamento de verbos das classes 38LD e 38LS. Compreende-se, assim, como necessário o estudo mais aprofundado sobre a possibilidade de omissão de argumentos locativos dessas três classes (38LD, 38LS e 38LT) para criar critérios mais formais para a classificação verbal.

Conclui-se que computacionalmente o conjunto de regras gerado *com* a seleção de atributos não foi produtivo, já que, além dos desvios descritos, foi mantido o número de regras (dez) para o processamento da máquina. Por outro lado, a ação de *forçar* o algoritmo com a seleção dos atributos mais relevantes pôde direcionar o olhar do linguista a determinadas particularidades do comportamento sintático-semântico dos verbos locativos da língua espanhola, que, em momento anterior, não tiveram a atenção e, consequentemente, descrição adequada. Sendo assim, esse *exercício* de geração e interpretação do algoritmo apresentou caminhos possíveis para ampliar a descrição linguística humana do fenômeno em questão.

CONSIDERAÇÕES FINAIS

Esta pesquisa teve como principal objetivo demonstrar que a relação entre a *descrição humana* e o *aprendizado de máquina* não consiste somente na importância da descrição como *insumo* para a máquina, mas, principalmente, em como é possível utilizar algoritmos (e suas medidas de avaliação) para validar e aperfeiçoar a descrição de diferentes fenômenos das línguas naturais. Como sabido, descrições formais linguísticas são elementos basilares para a construção de ferramentas linguístico-computacionais. No entanto, este trabalho buscou demonstrar como os recursos computacionais também podem contribuir com as descrições linguísticas, principalmente as realizadas de maneira manual. Para tanto, analisou-se o algoritmo simbólico (JRip), com vistas a avaliar e refinar a proposta de tipologia dos verbos locativos em língua espanhola da base de dados LGLE (RODRIGUES, 2019).

Os dados demonstraram que a anotação humana das 49 propriedades (atributos) estruturais, distribucionais e transformacionais do LGLE apresenta primorosa acurácia (100%), quando considerados todos os seus atributos no AM. Para *forçar* os dados, em um segundo momento, optou-se por verificar o resultado do algoritmo baseado apenas nos atributos mais relevantes da tarefa, o que gerou uma avaliação de aproximadamente 97% de acurácia e possibilitou o apontamento de reflexões linguísticas a serem tratadas em investigações futuras sobre os seguintes fenômenos:

a) análise e verificação da produtividade na língua de construções verbais locativas que possuem atributos específicos e bem delimitados que as caracterizam, como o ocorrido com os verbos das classes 38L2, 38L3 e 38LT;
b) estudos de construções semelhantes à exemplificada com o verbo *pasear*, as quais apresentam restrições de seleção para o preenchimento de um determinado argumento que impactam outros atributos;
c) revisão da propriedade transformacional *pas_estar* (passiva com *estar*) nas construções de movimento, considerando as particularidades do fenômeno da apassivação em língua espanhola (ARAÚJO JÚNIOR, 2014);

d) ampliação da anotação semântica dos dados do LGLE, sobretudo do argumento que ocupa a posição N_1;
e) análise (sincrônica e/ou diacrônica) de construções denominais locativas (*enjaular, enlatar, embotellar*) que parecem ter se desvencilhado – se não totalmente, ao menos parcialmente – dos nomes dos quais derivam (*anidar/nido; almacenar/almacén*), gerando uma valência verbal distinta e, consequentemente, uma construção nova; e
f) revisão dos verbos que transitam entre a valência quatro (38LT) e a valência três (38LD, 38LS), por possibilitarem a omissão do argumento N_2 (de *origem*) e/ou N_3 (de *destino*).

Retomando as questões que mobilizaram a realização desta investigação, pode-se afirmar que o AM contribui tanto com a validação, quanto com o aprimoramento da descrição linguística humana, ao mostrar, a partir das medidas de precisão, quais instabilidades descritivas e classificatórias encontram-se na base de dados. Mais especificamente, destacam-se as relações estabelecidas entre as classes 38L2-38LD e 38LT-38LS. A tarefa de AM indicou ainda quais atributos das construções analisadas destacam-se e atuam como elementos fundamentais para a classificação dos verbos locativos estudados (cf. Tabela 3), além da geração do algoritmo em si, que contribui com tarefas para a área de PLN.

REFERÊNCIAS

ANTUNES, R. A. M. R.; PARDO, T. A. P. S; ALMEIDA, G. M. B. *Formação de gentílicos a partir de topônimos: descrição linguística e aprendizado automático*. *In*: STIL 2017 - XI BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY AND COLLOCATED EVENTS - Proceedings of the Conference. Uberlândia, 2017.

ARAÚJO JÚNIOR, B. J. As formas passivas. *In*: FANJUL, A. P.; GONZÁLEZ, N. M. (orgs.). *Espanhol e português brasileiro:* estudos comparados. São Paulo: Parábola Editorial, 2014.

BAPTISTA, J. ViPEr: uma base de dados de construções léxico-sintáticas de verbos do Português Europeu. *In:* Textos Selecionados, XXVIII ENCONTRO NACIONAL DA ASSOCIAÇÃO PORTUGUESA DE LINGUÍSTICA. Coimbra: APL, 2013. p. 111-129.

COHEN, W. W. *Fast effective rule induction*. *In:* MACHINE LEARNING PROCEEDINGS. Morgan Kaufmann, 1995.

DEPREN, O.; TOPALLAR, M.; ANARIM, E.; CILI, M. K. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. *Expert systems with Applications*, v. 29, n. 4, 2005.

FINATTO, M. J.; LOPES, L.; CIULLA, Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida. *Domínios de Lingu@Gem*, v. 9, 2015.

GARCÍA-MIGUEL, J. M.; COSTAS, L.; MARTÍNEZ, S. Diátesis verbales y esquemas construccionales: Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. *In*: VI CONGRESO INTERNACIONAL DE LINGÜÍSTICA HISPÁNICA. Leipzig, 2003.

GROSS, M. *Méthodes en syntaxe*. Paris: Hermann, 1975.

GROSS, M. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 1981.

GUILLET, A.; LECLÈRE, C. *La structure des phrases simples en français:* constructions transitives locatives. Genebra: Librairie Droz S.A, 1992.

JURASKY, D.; MARTIN, J.H. *Speech and Language Processing:* An introduction to natural language Processing. Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey, 2000.

KUBAT, M. *An introduction to machine learning*. 2ª ed. CoralGables/EUA: Springer International Publishing, 2017.

MITCHELL, T. M. *Machine Learning*. Nova York: McGraw-Hill, 1997.

MONARD, M. C.; BATISTA, G.; KAWAMOTO, S; PUGLIESI, J. B. *Uma introdução ao aprendizado simbólico de máquina por exemplos*. São Carlos: ICMSC-USP, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *In*: *Sistemas Inteligentes:* Fundamentos e Aplicações, v. 1, 2003.

QUINLAN, J. R. Induction of decision trees. *Mach Learn*, v. 1, p. 81-106, 1986.

RAJPUT, A.; AHARWAL, R. P.; DUBEY, M.; SAXENA, S. P.; RAGHUVANSHI, M. J48 and JRIP rules for e-governance data. *International Journal of Computer Science and Security* (IJCSS), v. 5, 2011.

RAMMÉ, V.; RODRIGUEZ, D. G. V. O uso da preposição ‘a’ como objeto direto (OD) e objeto indireto (OI): uma análise contrastiva entre o espanhol e o português brasileiro. **Caletroscópio**, v. 8, 2020.
REAL ACADEMIA ESPAÑOLA: *Diccionario de la lengua española*, 23.ª ed., [versión 23.5 en línea]. https://dle.rae.es. 2014. Acesso em: out. 2021.
RODRIGUES, R. *Contribuições para um léxico-gramática das construções locativas do espanhol*. 2019. Tese (Doutorado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2019.
RODRIGUES, R. Tipología con fines pedagógicos de los verbos locativos del español. *Domínios de Lingu@gem*, v. 14, 2020.
SANTOS, R. P. T. *Automatic Semantic Role Labeling for European Portuguese.* Dissertação (Mestrado em Ciências da Linguagem) – Universidade do Algarve, Faro, 2014.
SILVA, A. V. V. *Classificação baseada em regras para estudo da produtividade do algodão no estado do Mato Grosso*. Dissertação (Mestrado em Matemática, Estatística e Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2019.
SOUZA, J. W. C. *Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento*. Tese (Doutorado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2019.
VIEIRA, R.; STRUBE, V. L. *Linguística Computacional:* princípios e aplicações. *In*: IX ESCOLA DE INFORMÁTICA DA SBC-SUL. Porto Alegre - RS: UFRGS, 2001.

Recebido: 15/9/2021
Aceito: 13/10/2022
Publicado: 14/10/2022