

REVISÃO E AMPLIAÇÃO DE ÁRVORES DE DOMÍNIO COM BASE NA ANÁLISE DE CORPUS

Review and Expansion of Concept Systems through Corpus Analysis

DOI: 10.14393/LL63-v41S-2026-13

Amanda Letícia Valadares dos Santos*

Flávia de Oliveira Maia-Pires**

RESUMO: Independentemente do formato, árvores de domínio são maneiras de arquitetar informações para compreensão de conceitos conexos, para identificação do conceito de um termo e suas variações, sendo úteis ao treinamento de Inteligências Artificiais (IA), entre outros. Durante sua revisão e atualização, o terminólogo pode contar com o apoio de ferramentas tecnológicas. Assim, este artigo demonstra a Linguística de Corpus (LC) como abordagem metodológica importante na construção de árvores de domínio, que identifica candidatos a termos. Assim, descreve o tratamento dado ao corpus de estudo no *Sketch Engine*, que dispõe de recursos de LC como: *Wordlist*, *Keywords*, *Word Sketch* e *Concordance*. Essa estratégia resultou na identificação de 12 termos importantes que não constavam no diagrama inicial, o que permitiu a revisão e a ampliação da árvore de domínio para o aprofundamento conceitual e o mapeamento das correlações terminológicas.

PALAVRAS-CHAVE: Terminologia. Árvore de domínio. Linguística de Corpus. Sketch Engine. Lei Geral de Proteção de Dados Pessoais.

ABSTRACT: Regardless of format, concept systems constitute methodological frameworks for organizing information to understand interconnected concepts, to identify term definitions and their variations, and support Artificial Intelligence (AI) training, among other applications. When reviewing and updating these structures, terminologists can leverage technological tools for enhanced efficiency. This article demonstrates how Corpus Linguistics (CL) serves as a crucial methodological approach in concept systems construction for identifying potential term candidates. Accordingly, a study corpus was processed using Sketch Engine, which provides the following CL resources: Wordlist, Keywords, Sketch Word, and Concordance. This methodology resulted in the identification of 12 significant terms absent from the initial diagram, thereby enabling the revision and expansion of the concept systems to achieve greater conceptual depth and comprehensive mapping of terminological correlations.

KEYWORDS: Terminology. Concept System. Corpus Linguistics. Sketch Engine. Brazilian Data Protection Law.

* Mestranda do Programa de Pós-Graduação em Linguística do Instituto de Letras. Universidade de Brasília (UnB). ORCID: 0009-0001-0902-176. E-mail para contato: linguista.amandavaladares(AT)gmail.com.

** Professora Associada do Departamento de Linguística, Português e Línguas Clássicas (LIP) do Instituto de Letras. Universidade de Brasília (UnB). ORCID: 0000-0001-7167-5214. Email para contato: maiapires(AT)unb.br.

1 Introdução

A Terminologia é a disciplina da Linguística responsável pelo estudo dos termos, isto é, das unidades linguísticas utilizadas em discursos especializados, cujo significado geralmente não é conhecido por leigos. Devido a esse recorte investigativo, entende-se que a Terminologia é interdisciplinar e envolve processos de familiarização com áreas distintas das Ciências da Linguagem. Conforme afirmam Cruz, Dallapicula e Maia-Pires (2023, p. 328):

A Terminologia, como disciplina, ocupa-se dos conceitos, definições e denominações de uma área do saber, mas também está relacionada ao conjunto de termos atrelados a uma área específica, como a terminologia da Matemática, da Física, da Linguística, etc.

Devido a esse caráter interdisciplinar, os terminólogos recorrem a metodologias distintas para se aproximarem e compreenderem conceitualmente as terminologias dentro da estrutura de uma nova área do saber. Dessa forma, nos estudos terminológicos, é recorrente a elaboração de árvores de domínio, definidas como “a representação, em uma forma piramidal, dos conceitos-chave de um domínio e das relações que eles mantêm entre si”¹ (Zafio, 1985, p. 161, tradução nossa).

As árvores de domínio elucidam, para o pesquisador inicialmente leigo, a estrutura conceitual em que a comunicação especializada ocorre, apresentando uma representação hierárquica e visual dos conceitos de uma área, capaz de organizar os termos ao demonstrar as relações de subordinação e coordenação entre os conceitos a partir das categorias mais gerais para as mais específicas.

Em geral, não há um consenso sobre quais metodologias os terminólogos devem utilizar para criar árvores de domínio com eficiência, cobrindo a maior extensão conceitual possível do subléxico de especialidade dos seus estudos, dado que cada área possui suas especificidades. Entretanto, percebe-se que há metodologias semelhantes utilizadas pela maioria das pesquisas terminológicas. Após leituras de contextualização na área de especialidade em estudo, o pesquisador mapeia as relações entre as palavras-chave e os termos observados, baseando-se, por exemplo, em frequências, em conceitos conexos ou em

¹ *la représentation, sous une forme pyramidale, des notions clés d'un domaine et des relations qu'elles entretiennent entre elles.*

variantes terminológicas identificadas durante o processo de mapeamento dos dados de análise.

Nesse contexto, para alcançar tal compreensão conceitual, a pesquisa terminológica está avançando mediante a implementação de recursos tecnológicos. Neste trabalho, apresenta-se a Linguística de *Corpus* (LC) como metodologia capaz de auxiliar nesse processo, tornando-o mais preciso e eficiente. Trata-se de uma abordagem metodológica que utiliza grandes coleções de textos autênticos (*corpora*) para investigar fenômenos linguísticos (McEnery; Wilson, 2001).

Ferramentas tecnológicas dedicadas à LC disponíveis no mercado permitem o processamento dos textos para analisar dados linguísticos de modo sistemático. Assim, é possível obter dados quantitativos sobre quais termos são mais chave em textos da área de especialidade. Em adição a isso, pesquisadores têm à disposição recursos para identificar aspectos conceituais, gramaticais e discursivos em árvores de domínio, que podem revelar lacunas de conceitos importantes não considerados inicialmente.

Diante do exposto, este trabalho tem o objetivo geral de apresentar como a LC pode auxiliar terminólogos no processo de elaboração das árvores de domínio. Notadamente, como objetivo específico, busca-se demonstrar como a LC pode ser aplicada para revisar e ampliar as árvores de domínio inicialmente elaboradas por pesquisadores — de modo a verificar a pertinência dos termos escolhidos para compô-las, além de quais podem ser adicionados ao diagrama inicial.

2 Pressupostos teóricos

2.1 A expansão significativa de “árvore de domínio”

A “árvore de domínio” pode ser definida como a organização e representação sistemática de uma determinada área do conhecimento por meio da interconexão dos termos que a integram, em associação com outros recursos visuais (geralmente dispostos em formato arbóreo). Essa estrutura remete a uma árvore, em que o termo mais geral se localiza no topo, a base (por exemplo, "dados"), e os conceitos mais específicos se ramificam abaixo (por exemplo, "dados pessoais", "dados sensíveis", "dados anonimizados").

Tal representação configura-se por meio da chamada árvore do domínio, um construto teórico que desenha a hierarquia temática de cada domínio do saber, pretendendo, com isso, representar o sistema lógico-cognitivo que particulariza os universos de conhecimento especializado. Por isso, a árvore funciona como uma espécie de mapa conceitual do domínio, auxiliando a reconhecer a vinculação terminológica, nessa medida, a pertinência dos termos a uma área (Krieger; Finatto, 2004, p. 56).

De maneira geral, uma ferramenta de representação gráfica refere-se à organização visualmente estruturada de informações de modo associativo ou hierárquico, por meio de diagramas, árvores ou mapas conceituais. Nesse sentido, é importante considerar que a denominação da ferramenta de representação gráfica utilizada em Terminologia para indicar a associação entre conceitos não é uniforme entre diferentes autores e tradições terminológicas.

Em particular, “árvore de domínio” constitui uma tradução do francês *“arbre de domaine”*, termo que inicialmente assume um significado ligeiramente distinto daquele utilizado por Wüster no alemão *“begriffssystem”* — frequentemente traduzido para o inglês como *“concept system”*, ou seja, “sistema de conceitos” (Temmerman, 2000).

O termo *“arbre de domaine”* começa a aparecer em textos de pesquisadores da França e do Canadá (Quebec), nas décadas de 1970 e 1980, em que fica clara a distinção do que esses autores pretendiam estabelecer como “árvore de domínio”, em comparação com *“begriffssystem”*. Este último termo é mais abrangente, referindo-se a esquemas lógicos de organização dos conceitos, em geral; o primeiro, por sua vez, é mais específico, referindo-se à disposição gráfica hierárquica do domínio e de seus subdomínios.

O terminólogo deve concentrar a atenção na estruturação de noções que sustentam objetos no mundo real (domínio nocional). Essa estruturação geralmente é um reflexo de atividades específicas de um domínio particular. De fato, toda unidade terminológica tem a função de nomear, ou seja, estabelecer a relação entre uma forma linguística (palavra, termo, etc.) e uma realidade (abstrata ou concreta) por meio de uma noção. Consequentemente, o terminólogo deve estabelecer uma lista estruturada de subdomínios, dentro dos quais ele pode classificar as atividades específicas do domínio com o qual está lidando. Esses subdomínios podem ser organizados na forma de um diagrama de árvore (árvore de domínio)

que descreve a estrutura hierárquica do domínio.² (Auger; Rousseau, 1984, p. 14, tradução nossa).

Apesar dessa delimitação inicial mais específica, com o avanço dos estudos terminológicos de língua francesa, o termo “*arbre de domaine*” passou a ser utilizado, em certo nível, como equivalente de “*begriffssystem*”. Essa atualização de significado alinha-se com a tendência internacional de normatização promovida pela International Organization for Standardization (ISO), que prioriza as nomenclaturas “*concept diagram*” e “*tree diagram*” (ISO 704:2022) e apenas “*concept diagram*” (ISO 1087:2019), o que enfatiza mais o caráter epistemológico do que o hierárquico.

Tomando outras línguas como referência comparativa, Temmerman (2000) refere-se a essa estrutura como “*conceptual system*”, seguindo a tradição austríaca. O pesquisador uruguaio Barité (2017a) segue a tradição francesa clássica, adotando o nome “*árbol de dominio*”. Em contrapartida, a autora espanhola Cabré (1999) denomina essa estrutura de “*sistema conceptual*” ou simplesmente de “*diagrama*”, acrescentando que essa estrutura não precisa ser necessariamente em formato arbóreo.

No âmbito da língua portuguesa, Faulstich denomina essa estrutura como “diagrama” e “árvore conceitual”, ao traduzir o Manual de Terminologia de Pavel e Nolet (2002). Já Barros (2004), embora cite a nomenclatura “árvore de domínio”, prefere adotar as formas específicas “árvore de conceitos”, “árvore de características” e “diagrama”. Por outro lado, outros pesquisadores em língua portuguesa adotam especificamente “árvore de domínio” — a exemplo de Krieger e Finatto (2004), bem como Arcos e Bevilacqua (2020).

Diante dessa diversidade terminológica, atualmente existem muitos modelos de árvores de domínio possíveis, todos fundamentados teoricamente e muitos (embora não sempre) com disposição hierárquica. Dessa forma, há versões mais ontológicas, outras taxionômicas e outras próximas de mapas mentais. A título de exemplo, é possível citar os

² El terminòleg ha de centrar l'atenció en l'estructuració de les nocions retendes a objectes del món real (camp nocional). Aquesta estructuració és sovint el reflex d'activitats concretes pròpies d'un camp determinat. En efecte, tota unitat terminològica té per funció la denominació, és a dir, l'establiment de la relació entre una forma lingüística (paraula, terme, etc.) i una realitat (abstracta o concreta) a través d'una noció. En conseqüència, el terminòleg haurà d'establir una llista estructurada de sub-camps C' dins els quals pugui classificar les activitats pròpies del camp que ha de tractar. Aquests subcamps podran organitzar-se en forma d'un diagrama arbori (arbre de camp) que esquematitzi l'estructura jeràrquica del camp.

trabalhos de Maia-Pires (2024), acerca da terminologia da legislação brasileira sobre migração, e de Santos (2023), sobre a terminologia utilizada na Lei Geral de Proteção de Dados Pessoais (LGPD).

Neste artigo, essas formas de representação gráfica serão abarcadas pelo escopo do termo “árvore de domínio”, o qual também será considerado sinônimo de “sistema conceitual” — reconhecendo-os como equivalentes originários de tradições terminológicas distintas.

Do ponto de vista estrutural, Barité (2017a) explica que árvores de domínio são formadas, em suas formas mais hierárquicas, por um anel nuclear, também denominado de nódulo conceitual, e outro de termos afins. Vale ressaltar que, os elementos do anel de termos afins podem ser denominados como “ramos”, principalmente nas estruturas mais arbóreas e hierárquicas das árvores de domínio. Desse modo, o ideal é que cada candidato a termo possa ser alocado em um desses anéis. Ademais, conforme o diagrama se estende e se torna mais complexo, convém utilizar verbos de conexão entre os termos, bem como cores para diferenciar os termos do anel nuclear.

Nessa perspectiva, o anel nuclear deve ser composto por termos que indicam conceitos diretamente e exclusivamente associados à área de especialidade em questão. No caso deste estudo, como será explanado mais adiante, podem-se citar “titular”, “controlador”, “operador” e “dado pessoal” como exemplos. Já os termos do anel afim seriam aqueles que se relacionam de forma mais circunstancial com aquela área do saber. Para esse segundo caso, no contexto da Lei Geral de Proteção de Dados Pessoais (LGPD), é possível citar os termos “segurança da informação”, “termo de uso”, “terceiros”, “fornecedores” e “prestadores de serviço”.

Diante do exposto, convém esclarecer como as árvores de domínio são elaboradas, verificadas, revisadas, ampliadas e aprovadas. Em geral, seguindo recomendações da própria ISO 704 (2000), as árvores de domínio, vistas como *conceptual system* (sistema conceitual), constituem uma etapa prévia de aproximação do terminólogo com relação ao seu objeto de pesquisa — conforme defendido por Cabré (1999), Barros (2004), bem como Krieger e Finatto (2004).

2.2 *Elaboração, revisão e ampliação das árvores de domínio*

No âmbito da Teoria Geral da Terminologia (TGT), os sistemas conceituais desempenhavam um papel central no processo de identificação e definição de termos. Nesse sentido, os sistemas conceituais eram não apenas uma delimitação inicial, mas também um trabalho constantemente revisitado e ampliado — a cada novo possível termo identificado.

Anos depois, Cabré (1999) propõe a Teoria Comunicativa da Terminologia (TCT), que rompe consideravelmente com as diretrizes rígidas e, por vezes, artificiais da TGT. No novo paradigma proposto pela autora, os sistemas conceituais passam a ocupar uma posição mais norteadora e menos seletiva. Tal abordagem não implica, entretanto, a diminuição da importância de o pesquisador elaborar construções ontológicas para organizar e validar os termos.

...a árvore de domínio permite enquadrar cada termo em algum de seus ramos ou subáreas e, desse modo, garante tanto sua existência quanto seu pertencimento ao domínio [...] cada um dos termos deve se situar em algum lugar da estrutura básica que a árvore proporciona. Se isso não for possível, o candidato a termo deverá ser considerado um caso duvidoso ou inclusive ser excluído.³ (Barité, 2017a, p. 96, tradução nossa).

Nesse contexto, as ferramentas tecnológicas utilizadas nos estudos terminológicos, que incluem a abordagem da Linguística de *Corpus*, contribuem significativamente para a identificação e alocação de termos relevantes para a árvore de domínio da área estudada.

Segundo McEnery e Hardie (2012, p. 1), a LC, no âmbito das abordagens de estudo das línguas, trabalha com “um conjunto de textos legíveis por máquina que é considerado uma base apropriada para estudar um conjunto específico de questões de pesquisa”⁴ (tradução nossa). Acrescenta-se, ainda, o fato de a LC realizar seus estudos linguísticos com base em exemplos de uso da 'vida real', conforme afirmam McEnery e Wilson (2001, p. 1). Para tanto, são utilizados recursos informáticos, os quais contribuem para a identificação de relações

³ *el árbol de dominio permite encuadrar cada término en alguna de sus ramas o subáreas y de ese modo garantiza tanto la existencia del término como su pertenencia al dominio. [...] cada uno de los términos deberá situarse en algún lugar de la estructura básica que proporciona el árbol. Si esto no es posible, el término candidato deberá considerarse caso dudoso o inclusive ser excluído.*

⁴ *some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions.*

quantitativas dos termos mais recorrentes e das construções linguísticas coocorrentes, por meio de uma exploração empírica.

No âmbito da Terminologia, a LC é amplamente utilizada como metodologia para identificação de candidatos a termos. Com esse auxílio, é possível mitigar o que Aubert (2001) chama de "risco de ruído" e "risco de silêncio". Isto é, o risco de incluir termos não relacionados ao campo estudado e o risco de não se incluírem termos importantes para a análise, respectivamente.

Neste artigo, será explorada especificamente a possibilidade de expandir o diagrama inicialmente elaborado pelas autoras, resultante da leitura e análise da própria LGPD, além de alguns outros documentos relevantes para o tema. Durante o processo, após a definição das duas árvores de domínio iniciais para estruturação da área estudada, o *corpus* reunido foi submetido a uma ferramenta de LC, principalmente para verificar se houve lacunas terminológicas despercebidas na primeira etapa do estudo epistemológico.

Com isso, foi possível preencher melhor os anéis, ou ramos, do sistema conceitual inicialmente elaborado. Tal aferição baseou-se em três critérios principais: frequência total, frequência comparativa e coocorrência. Neste trabalho, a LC contribuiu para a construção de árvores de domínio da área, pois submeter o *corpus* de estudo a uma ferramenta como o *Sketch Engine* (que dispõe de recursos como *Wordlist*, *Keywords*, *Word Sketch* e *Concordance*) tornou a análise mais completa, com base em dados objetivos e revelando as relações semânticas que, por vezes, não são óbvias na análise manual.

Destaca-se que a comparação entre *corpora* especializados e de língua geral possibilita validar quais termos são realmente específicos do domínio, evitando o "risco de ruído" e o "risco de silêncio", já mencionados (Aubert, 2001). Além disso, os dados quantitativos fornecem critérios objetivos para determinar a importância relativa dos termos na hierarquia conceitual, enquanto a possibilidade de atualização dos *corpora* permite que as árvores evoluam dinamicamente. Essa abordagem híbrida envolve rigor teórico e precisão empírica, o que resulta em estruturas conceituais mais robustas e fidedignas à realidade terminológica da área estudada.

Assim sendo, com um viés interdisciplinar mais recente, as árvores de domínio, em razão de sua estrutura frequentemente ontológica ou taxinômica, também podem ser

compreendidas como uma forma de arquitetar informações. Nesse sentido, contribuem para o treinamento de Inteligências Artificiais (IA), ao constituir bases simplificadas e computacionalmente interpretáveis de conhecimento (Knight, 2017).

2.3 Árvores de Domínio e Inteligência Artificial (IA)

Segundo Lara (2010), a Ciência da Informação (CI) e a Terminologia se interseccionam na perspectiva compartilhada de que unidades linguísticas representam conceitos especializados, os quais se conectam de forma estruturada, a partir da lógica que organiza e constrói as áreas do conhecimento. Isso porque as unidades terminológicas podem funcionar tanto como um recurso independente e autônomo quanto como um componente integrado a um sistema, conforme explica Freitas (2022, p. 59). Sendo assim, ambas as disciplinas científicas buscam, de alguma forma, organizar termos, embora a CI não use essa nomenclatura específica.

Barité (2017b) explica que, em geral, há duas formas principais de organização do conhecimento, compartilhadas entre a CI e a Terminologia: alfabética e sistemática. As árvores de domínio enquadram-se no segundo tipo, caracterizado pela seleção, ordenação e correlação ontológica. Ainda segundo o autor, visto que termos não existem isoladamente, mas sim em concatenação mútua para criar contextos, devem-se evitar, ao máximo, os chamados “termos órfãos”.

Uma outra área de destaque, que utiliza muitos preceitos da CI como base, é a Tecnologia da Informação — principalmente quando aplicada à organização do conhecimento para interpretação por tecnologias baseadas em informática. Para os fins deste estudo, serão abordadas em mais detalhes as Inteligências Artificiais Generativas de cunho textual. Em suma, essas tecnologias funcionam a partir da ampla análise e posterior associação de dados, para gerar determinado resultado (nesse caso, textos que se assemelhem à produção humana).

Como ferramenta de apoio, esses sistemas inteligentes permitem atualização contínua das estruturas terminológicas, complementando eficientemente o trabalho tradicional de terminólogos e cientistas da informação na organização sistemática do conhecimento especializado.

Com a hegemonia do aprendizado de máquina e do conhecimento obtido apenas a partir dos dados, as ontologias — desde sempre associadas ao PLN simbólico de influência racionalista — cada vez mais são utilizadas como uma maneira de representar o conhecimento específico de um domínio ou uma maneira de organizar a informação de uma área, e podem integrar sistemas para viabilizar o raciocínio lógico ou alimentar, com suas classes e termos, a anotação de corpus. (Freitas, 2022, p. 62).

A matéria prima de todo esse processo é um conjunto bem estruturado de dados, na forma de bases de dados, construídas de forma bruta ou refinada. A forma refinada reduz o tempo de treinamento das máquinas, pois já apresenta algum insumo para a interpretação das associações entre palavras. A área de Processamento de Linguagem Natural (PLN) utiliza desde a classificação gramatical até a estruturação sistemática para criar bases de dados. Esta segunda abordagem envolve a criação de ontologias.

Essa interdependência é importante porque máquinas não possuem o conhecimento de mundo necessário para distinguir categorias semânticas. Essas relações partitivas precisam ser informadas às IAs, de forma logicamente estruturada. É nesse contexto que as árvores de domínio figuram como uma ferramenta importante de organização do conhecimento — principalmente as definimos de forma mais ampla, abarcando diferentes construções associativas, conforme descrito na metodologia do estudo aqui apresentado.

3 Metodologia

Considerando a importância das árvores de domínio como ferramenta de organização do conhecimento e sua aplicabilidade em contextos tecnológicos contemporâneos, torna-se fundamental demonstrar sua utilização prática em pesquisas terminológicas atuais. Para comprovar a necessidade de revisão e atualização dessas estruturas organizacionais em uma pesquisa terminológica, na primeira etapa da pesquisa da qual se apresenta aqui um recorte, foram elaboradas árvores associadas a uma pesquisa em *corpus* com textos associados à Lei Geral de Proteção de Dados Pessoais (LGPD) — isto é, à Lei Federal nº 13.709 (BRASIL, 2018).

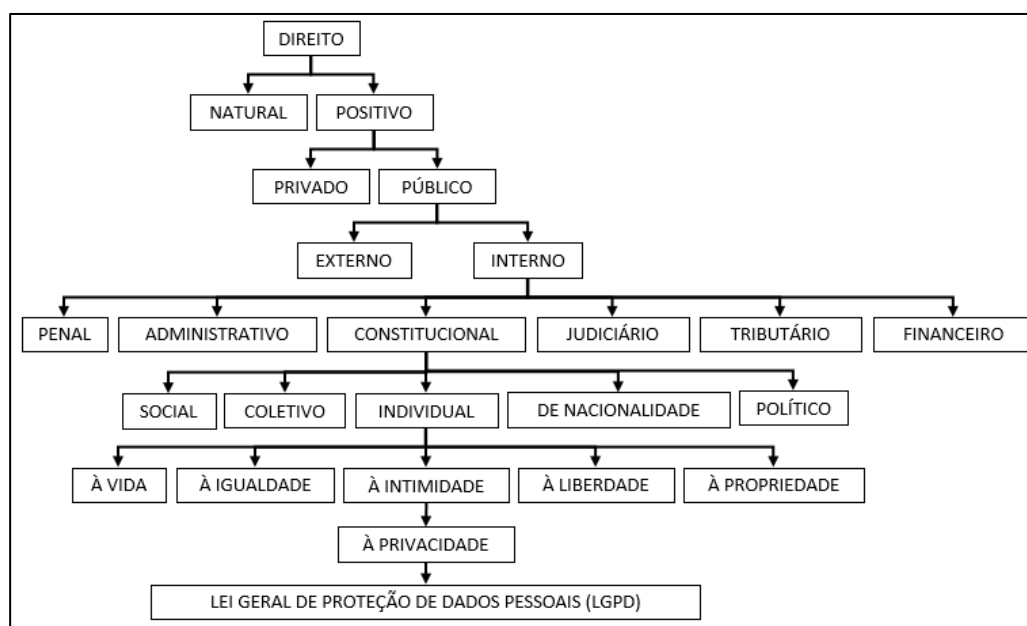
Esta Lei, que rege a privacidade dos brasileiros, garante-lhes o direito de decidir o que fazer com suas informações, com quem compartilhá-las, com qual finalidade e por quanto tempo. Sua relevância é ainda maior atualmente, em um contexto global no qual empresas de tecnologia compram e vendem informações pessoais de seus clientes para treinar

Inteligências Artificiais e para enviar anúncios direcionados a perfis de consumo — conectando-se diretamente às discussões anteriores sobre IA e processamento de dados.

O processo inicial de criação das árvores de domínio exigiu uma dupla compreensão: i) em qual ramo do Direito a LGPD se insere? ii) Como a própria LGPD se organiza internamente? Assim sendo, tais necessidades explicativas culminaram na elaboração de duas árvores, partindo de textos de referência diferentes. Os critérios para a escolha de como dispor cada termo no diagrama foram as relações de hiponímia, hiperonímia e associação, conforme contextos de uso nesses textos. Após a construção de ambas as árvores de domínio, estabeleceu-se uma relação comparativa entre as duas estruturas, a partir da LC, considerando-se dois aspectos: a) tema abordado e b) disposição visual adotada (hierarquicamente ou sob a forma de mapa conceitual).

Isto é, a primeira árvore referiu-se à alocação da LGPD em um dos subdomínios do Direito, o que exigiu a consulta a duas referências que estruturassem a referida área. Para tanto, observaram-se as áreas do Direito definidas pela base de dados do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), em associação com a doutrina jurídica de Montoro (2016). Ao correlacionar ambas as referências, foi possível elaborar a seguinte árvore de domínio, com a temática de subdomínios do Direito e com disposição hierárquica.

Figura 1 – Árvore de domínio do Direito, contendo os domínios em que se insere a LGPD



Fonte: elaboração própria, no PowerPoint.

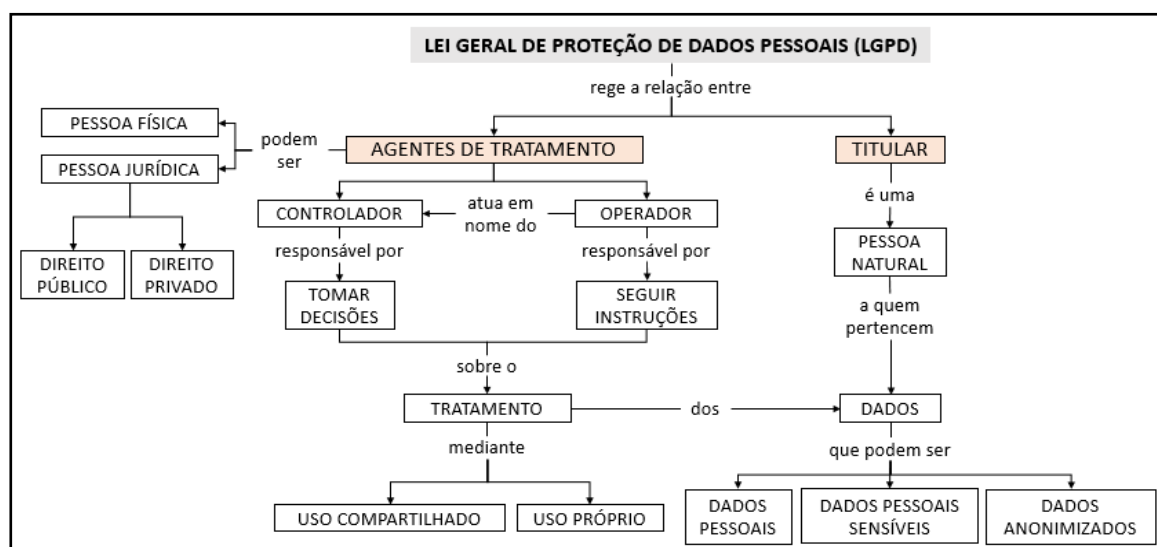
Por outro lado, a segunda árvore de domínio abarcou a temática da organização interna da LGPD com a disposição associativa de mapa conceitual. Os principais dados para este segundo diagrama foram extraídos da própria legislação em estudo, que foi acrescentada a um *corpus*, elaborado para a análise terminológica da referida área do conhecimento.

O *corpus* inclui nove textos:

- (1) LGPD, com redação final de 2019;
- (2) Guia de Elaboração de Termo de Uso e Política de Privacidade para Serviços Públicos, na Versão 1.3;
- (3) Norma ABNT/NBR ISO 27701;
- (4) Política de Privacidade do BRB;
- (5) Política de Privacidade da Caixa;
- (6) Política de Privacidade do Banco do Brasil;
- (7) Política de Privacidade do Itaú;
- (8) Política de Privacidade do Bradesco;
- (9) Política de Privacidade do Santander.

Tais textos constituem o *corpus* de referência da LGPD, selecionado para as próximas etapas do estudo terminológico. Por conseguinte, como resultado da leitura (humana) inicial desses nove textos, foi elaborada a segunda árvore de domínio desta pesquisa:

Figura 2 – Árvore de domínio das relações internas à LGPD



Fonte: elaboração própria no PowerPoint.

Uma vez definidas as duas árvores de domínio, avançou-se para a etapa de análise do *corpus*, com uma ferramenta informática especializada, o *Sketch Engine*, programa online de análise de *corpora*, amplamente utilizado por possuir variados recursos automáticos de LC. O recurso *Wordlist* oferece uma lista das palavras mais frequentes em todo o *corpus* inserido na ferramenta. Já o *Keywords* apresenta candidatos a termos a partir de cálculos comparativos com outro *corpus* de língua geral, que o próprio *Sketch Engine* possui em seu banco de dados.

A etiquetagem automática do corpus pelo *Sketch Engine* gera *tags* de classes gramaticais (*part-of-speech*), que se correlacionam quantitativamente e estatisticamente. O *Word Sketch* apresenta as relações de coocorrência de determinada unidade linguística com outras categorias sintáticas, baseando-se nessa anotação automática do corpus. Por fim, o *Concordance* apresenta, ao centro da tela, uma unidade linguística escolhida com seus contextos de ocorrência nos textos, à esquerda e à direita — procedimento que recebe a denominação de *Keyword in Context* (KWIC).

Esses quatro recursos foram consultados para identificar possíveis candidatos a termos que se encaixariam nas árvores de domínio inicialmente elaboradas a partir da leitura dos textos, visando expandir seus anéis ou ramos e tornar essas árvores o mais completas possível. A pertinência dos termos identificados pelo *Sketch Engine* se baseou em frequência, no contexto de uso (hiponímia, hiperonímia e associação) e no conhecimento da área, adquirido pelas pesquisadoras na etapa anterior do estudo, com previsão de verificação futura por especialista na área do Direito.

Para construção visual das árvores, após as análises, tanto da leitura contextualizante quanto dos resultados obtidos pelo *Sketch Engine*, as pesquisadoras utilizaram o programa PowerPoint. Apesar de haver ferramentas específicas para elaboração de ontologias no mercado, optou-se pelo referido programa, devido à sua acessibilidade em sistemas *Microsoft*.

4 Resultados

Após a elaboração inicial que resultou na construção das duas árvores deste estudo, o *corpus* da LGPD foi submetido à ferramenta *Sketch Engine*. Em seguida, efetuou-se uma busca pela categoria “*noun*” (substantivo) utilizando recurso *Wordlist*⁵. Essa categoria gramatical foi escolhida para investigação mais detalhada, pois a maioria dos termos são sintagmas nominais — seja em forma simples de substantivo, seja em forma complexa com um substantivo acompanhado por especificadores (Krieger, 2006). Paralelamente, consultou-se a lista de *Keywords*⁶ da ferramenta, que destaca termos significativos para o domínio estudado. A Figura 3 ilustra os 15 primeiros substantivos e *keywords* (simples e complexos) identificados pelo *Sketch Engine*, evidenciando termos importantes que estavam ausentes nas árvores de domínio iniciais.

Figura 3 – Substantivos mais frequentes e *keywords* do corpus da LGPD.

Text types 1 (9) ***		KEYWORDS		KEYWORDS	
noun (2,185 items 26,574)		Text types 1 (9) ***		Text types 1 (9) ***	
Noun	Frequency ? ↓	SINGLE-WORDS ✓	MULTI-WORD	SINGLE-WORDS ✓	MULTI-WORD TERMS ✓
1 dado	810 ***	reference corpus: Portuguese Web 20		reference corpus: Portuguese Web 2020 (ptTenTen20) (Items: 8,84)	
2 dp	734 ***	Lemma	Frequency	Term	Frequency
3 informação	733 ***	1 lec	521 ***	1 diretriz para implementação	162 ***
4 iso	529 ***	2 hyperlink	174 ***	2 titular de dp	125 ***
5 tratamento	527 ***	3 dp	734 ***	3 informação estabelecida	100 ***
6 lec	521 ***	4 abnt	487 ***	4 tratamento de dp	87 ***
7 abnt	487 ***	5 nbr	379 ***	5 diretriz adicional	68 ***
8 organização	420 ***	6 sgpi	45 ***	6 tratamento de dados pessoais	91 ***
9 nbr	379 ***	7 iso	529 ***	7 tratamento de dados	123 ***
10 serviço	379 ***	8 anonimização	35 ***	8 operador de dp	48 ***
11 lei	374 ***	9 subcontratado	44 ***	9 direito reservado	93 ***
12 segurança	315 ***	10 anpd	31 ***	10 Termo de uso	53 ***
13 titular	303 ***	11 lgpd	69 ***	11 segurança da informação	124 ***
14 direito	297 ***	12 convir	274 ***	12 controlador de dp	40 ***
15 controle	286 ***	13 privacy	32 ***	13 dado pessoal	394 ***
		14 unibanco	39 ***	14 implementação do controle	33 ***
		15 conglomerar	20 ***	15 Dados pessoal	82 ***

Fonte: extraída de análise no programa Sketch Engine.

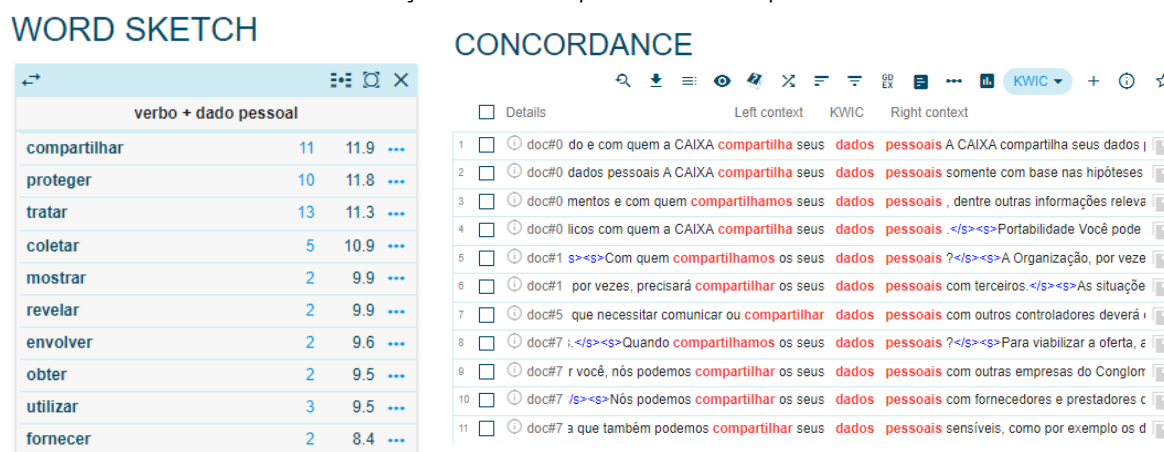
⁵ O recurso Wordlist possibilita a listagem por classes gramaticais, pois a ferramenta Sketch Engine realiza uma etiquetagem automática do corpus. É importante ressaltar, porém, que esse processo automático pode conter erros, sendo necessária atenção e checagem dos dados gerados, por parte dos pesquisadores.

⁶ Para extração da relação de Keywords no Sketch Engine, foi utilizado como corpus de referência o Portuguese Web 2020 (ptTenTen20).

A análise comparativa entre esses resultados e as árvores de domínio iniciais revelou lacunas terminológicas significativas. A partir dessa listagem, é possível notar que “dp”, “informação”, “organização”, “serviço”, “lei”, “segurança”, “direito”, “controle”, “anonimização”, “subcontratado”, “ANPD”, “termo de uso” e “segurança da informação”, todos exemplos de candidatos a termo, estão ausentes nos diagramas iniciais, selecionados com base nas informações fornecidas pelas análise “manual” das pesquisadoras, de acordo com a pertinência conceitual dos conceitos envolvidos da LGPD identificados no *corpus* de base via leitura. Cabe observar que, pelos termos estarem diretamente relacionados à LGPD, e não ao Direito como um todo, apenas a segunda árvore de domínio sofreu alterações nessa etapa.

Posteriormente, procedeu-se à análise contextual de ocorrência de um dos termos considerado central no domínio LGPD, a saber: “dado pessoal” (conforme apresentado na Figura 4). Para isso, utilizou-se o recurso *Word Sketch*, que gerou automaticamente informações gramaticais do termo, identificando padrões colocacionais e relações sintático-semânticas. A análise revelou que “dado pessoal” estabelece relação com diversos verbos, sendo “compartilhar” o mais frequente (11 ocorrências), que se associa ao termo “uso compartilhado”, presente na árvore da Figura 2, o que confirma a coerência terminológica do mapeamento inicial.

Figura 4 – Verbos mais frequentes que têm “dado pessoal” como complemento e colocações de “compartilhar dado pessoal”.

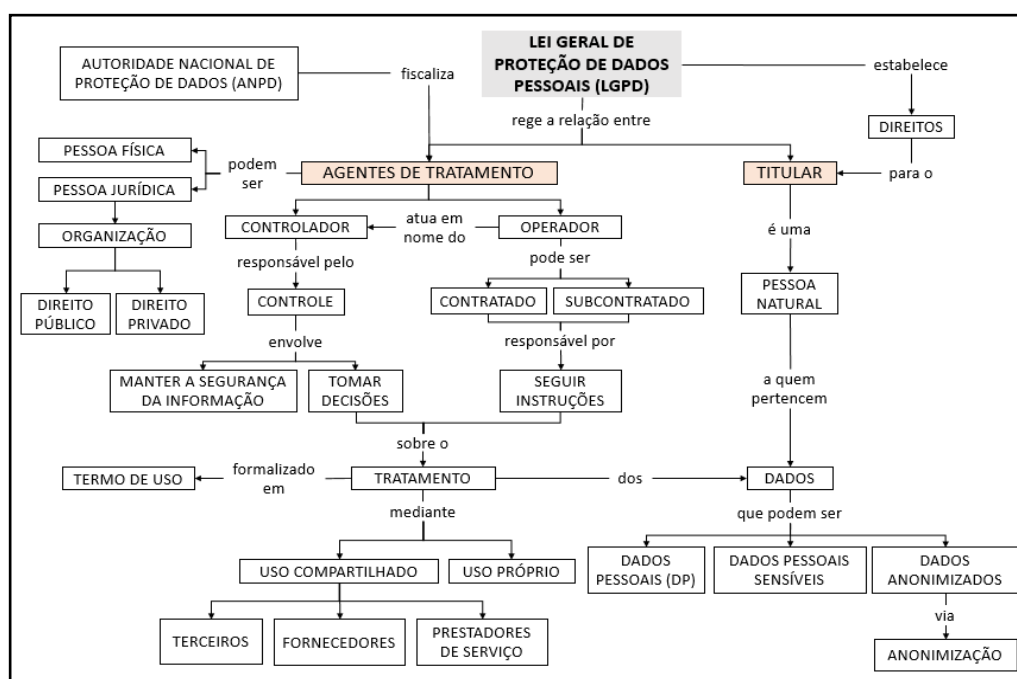


Fonte: programa Sketch Engine.

Uma verificação detalhada dessas ocorrências de “dado pessoal” associado a “compartilhar”, por meio do recurso *Concordance*, revelou a presença de “entes de compartilhamento”, como “terceiros”, “outros controladores”, “outras empresas do conglomerado”, “fornecedores” e “prestadores de serviço” — os quais também convém considerar incluir na árvore de domínio elaborada. Assim, a segunda árvore de domínio pôde ser consideravelmente ampliada a partir dos dados fornecidos, aumentando seus ramos com sucesso, além de tornar seu conteúdo mais completo, preciso e pertinente à área de Privacidade e Proteção de Dados Pessoais (vide Figura 5).

Novamente, nesta etapa, os termos identificados relacionaram-se especificamente à LGPD, resultando em alterações apenas na segunda árvore de domínio. Isso revela que, para enriquecer a primeira árvore de domínio, que discute como a referida lei se encaixa nos subdomínios do Direito, seria necessário um segundo *corpus*, mais abrangente, que englobasse os ramos específicos da área.

Figura 5 – Árvore de domínio revisada e atualizada da LGPD.



Fonte: elaboração própria no PowerPoint.

Aventou-se a necessidade de se elaborar um segundo *corpus*, mas essa etapa foi cancelada. Visto que o foco desta pesquisa consiste no entendimento da LGPD como marco

inaugural de uma nova área do conhecimento — a privacidade e proteção de dados pessoais no contexto digital —, optou-se por concentrar os esforços na revisão e ampliação exclusivamente da segunda árvore de domínio. Destaca-se que essa decisão metodológica não comprometeu a qualidade do trabalho, uma vez que obras de referência consolidadas fundamentaram a elaboração do primeiro diagrama, além da previsão de revisão por especialista da área em etapa posterior da pesquisa.

A seguir, destacamos as contribuições da metodologia empregada para revisão e ampliação da árvore de domínio, visando o aprofundamento conceitual e o mapeamento das correlações terminológicas, em adição às suas implicações para trabalhos futuros na área de proteção de dados pessoais.

5 Considerações finais

Diante do exposto, comprova-se que a Linguística de *Corpus* pode ser uma abordagem metodológica importante na construção de árvores de domínio durante a pesquisa terminológica. No *Sketch Engine*, ressaltam-se os recursos *Wordlist*, *Keywords*, *Word Sketch* e *Concordance* como auxiliares, especialmente, na identificação de possíveis candidatos a termos, relevantes para o diagrama de representação conceitual de uma área.

Os resultados obtidos mostram que a metodologia pode contribuir para a elaboração de árvores de domínio mais completas e condizentes com o uso real dos termos na comunicação especializada, desde que o *corpus* de estudo também seja bem elaborado e, de fato, representativo da área objeto de análise. Assim, o terminólogo poderá dispor de mais ferramentas além de sua leitura dos textos para construir a esquematização conceitual de como os conceitos-chave se relacionam.

Além dos benefícios terminológicos de norteamto para a seleção e compreensão dos candidatos a termos, essa abordagem para revisão e atualização das árvores de domínio pode oferecer outras vantagens. No contexto das IAs, por exemplo, essas configurações cada vez mais densas de arquitetura da informação auxiliam no treinamento direcionado e completo de tecnologias relacionadas à linguagem em áreas de especialidade.

Este trabalho mostrou que a aplicação prática dessa metodologia é eficaz. Após as análises conduzidas com o apoio da LC, no âmbito do estudo da LGPD, os seguintes doze

termos finais foram selecionados para alocação na árvore de domínio revisada e atualizada: “dp”, “organização”, “direito”, “anonimização”, “controle”, “ANPD”, “subcontratado”, “termo de uso”, “segurança da informação”, “terceiros”, “fornecedores” e “prestadores de serviço”. Os procedimentos aqui descritos demonstram o potencial dessa integração metodológica para estudos terminológicos em diferentes domínios do conhecimento.

Referências

ARCOS, M.; BEVILACQUA, C. R. UFE eventivas na área da conservação e restauração de bens culturais móveis em suporte papel: identificação e análise. **Debate Terminológico**, n. 18, p. 4-18, 2020. Disponível em: <https://seer.ufrgs.br/index.php/riterm/article/view/98700>. Acesso em: 07 jun. 2025.

AUBERT, F. H. **Introdução à Metodologia da Pesquisa Terminológica Bilingüe**. São Paulo: Humanitas Publicações (FFLCH/USP), 2001.

AUGER, P.; ROUSSEAU, L. J. **Metodologia de la recerca terminològica**. Tradução de Maria Teresa Cabré. Catalunya: Departament de Cultura, 1984. Disponível em: <http://hdl.handle.net/20.500.12368/1042>. Acesso em: 22 de junho de 2025.

BARITÉ, M. Los árboles de dominio. In: CATALÁ, S. A.; BARITÉ, M. (org.). **Teoría y praxis en terminología**. 1. ed. Montevideo: Ediciones Universitarias, Unidad de Comunicación de la Universidad de la República, 2017a, v. 1, p. 91-102.

BARITÉ, M.; RAUCH, M. El árbol de dominio como herramienta en organización del conocimiento: aplicación al área de educación superior. In: SIMÕES, Maria da Graça; BORGES, Maria Manuel (ed.). **Tendências Atuais e Perspetivas Futuras em Organização do Conhecimento**: atas do III Congresso ISKO Espanha e Portugal – XIII Congresso ISKO Espanha, Universidade de Coimbra, 23 e 24 de novembro de 2017. Coimbra, Centro de Estudos Interdisciplinares do Século XX – CEIS20, 2017b.

BARROS, L. A. **Curso Básico de Terminologia**. São Paulo: EdUSP, 2004.

BRASIL. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, ano 155, n. 157, p. 59-64, 15 ago. 2018.

CABRÉ, M. T. **Terminology**: theory, methods, and applications. Amsterdã: John Benjamins Pub. Co., 1999.

CRUZ, C. L. S.; DALLAPICULA, A. L. N.; MAIA-PIRES, F. O. Ficha terminológica informatizada: meio para manutenção do banco de dados do Glossário Terminológico da COVID-19. In: VICENTE, Renata Barbosa; SANTOS, Mônica Maria Soares. (org.). **Ficha terminológica**

informatizada: meio para manutenção do banco de dados do Glossário Terminológico da COVID-19. 1. ed. São Paulo: Estige Editorial, 2023, v. IV, p. 326-343.

FREITAS, C. **Linguística Computacional**. 1. ed. São Paulo: Editora Parábola, 2022.

INTERNATIONAL STANDARD ORGANIZATION (ISO). **ISO 1087**: Terminology work and terminology science — Vocabulary. Genebra: ISO, 2019.

INTERNATIONAL STANDARD ORGANIZATION (ISO). **ISO 704**: Terminology work: principles and methods. Genebra: ISO, 2022.

KNIGHT, M. Taxonomy vs Ontology: Machine Learning Breakthroughs. **Dataversity**, Los Angeles, 17 de out. de 2017. Disponível em: <https://www.dataversity.net/taxonomy-vs-ontology-machine-learning-breakthroughs/>. Acesso em: 07 jun. 2024.

KRIEGER, M. G. Do ensino da terminologia para tradutores: diretrizes básicas. **Cadernos de Tradução**, [S. l.], v. 1, n. 17, p. 189–206, 2006. DOI: <https://doi.org/10.5007/%25x>. Disponível em: <https://periodicos.ufsc.br/index.php/traducao/article/view/6862>. Acesso em: 07 jun. 2024.

KRIEGER, M. G.; FINATTO, M. J. B. **Introdução à Terminologia**: teoria e prática. São Paulo: Contexto, 2004.

LARA, M. L. G. Terminologia e Ciência da Informação In: **Os Estudos Lexicais em Diferentes Perspectivas**, volume II. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2010, v.2, p. 8-16.

MAIA-PIRES, F. O. Estudo do fluxo migratório no Brasil: a redução da desigualdade por meio dos termos. **Palimpsesto** - Revista do Programa de Pós-Graduação em Letras da UERJ, Rio de Janeiro, v. 23, n. 46, p. 273–293, 2024. DOI: <https://doi.org/10.12957/palimpsesto.2024.78817>

MCENERY, T.; HARDIE, A. **Corpus linguistics**: method, theory and practice. Cambridge: Cambridge University Press, 2012.

MCENERY, T.; WILSON, A. **Corpus Linguistics**: an introduction. 2nd Edition, Edinburgh: Edinburgh University Press, 2001.

MONTORO, A. F. **Introdução à ciência do direito**. 33. ed. São Paulo, SP: Revista dos Tribunais, 2016.

PAVEL, S.; NOLET, D. **Manual de terminologia**. Tradução de Enilde Faulstich. Canadá: Direção de Terminologia e Normalização Departamento de Tradução do Governo Canadense, 2002.

SANTOS, A. L. V. **Consentimento Informado**: Acessibilidade Terminológica e Fraseológica da Lei Geral de Proteção de Dados Pessoais, aplicada às Relações Bancárias. Monografia (Bacharelado em Letras) – Instituto de Letras, Universidade de Brasília, Brasília/DF, 2023. Disponível em: https://bdm.unb.br/bitstream/10483/37189/1/2023_AmandaLeticiaValadaresSantos_tcc.pdf. Acesso em: 07 jun. 2025.

TEMMERMAN, R. **Towards New Ways of Terminology Description**: The Sociocognitive-Approach. Amsterdã: John Benjamins Pub. Co., 2000.

ZAFIO, M. N. L'arbre de domaine en terminologie. **Meta**: Journal des traducteurs, Montreal, v. 20, n. 2, p. 161-168, jun. 1985. Disponível em: <https://www.erudit.org/fr/revues/meta/1985-v30-n2-meta308/004635ar.pdf>. Acesso em: 07 jun. 2024.

Recebido em: 30.06.2025

Aprovado em: 10.05.2026